Linking employer-employee data.

By
Niels Westergaard-Nielsen,
CLS, IZA and CCP, Aarhus School of Business, NWN@els.dk

Introduction

Linking employer-employee data opens up for analyses of the demand side of the labour market. For many years we have within empirical labour economics been able to analyse the supply side using register and survey data. These studies have produced large amounts of knowledge about labour supply, wage structure, impact of education, gender wage gaps, immigration and integration, unemployment and employment that now constitutes most of the knowledge in empirical labour economics and sociology, but they all stop short as to the factors related to firms. In most empirical analyses the firm is considered to be a black box based on theories of what happens inside firms. In the rare cases with empirical evidence, it is observations of a single firm.

Linked employer-employee data makes it possible to study differences across firms and the reactions to various policy interventions. The linked data makes it possible to answer policy-relevant questions on where jobs are created and where they are destroyed. It also makes it possible to answer questions on how different policy interventions affect firms and their demand for labour. This has a European policy dimension, a national dimension and a management dimension.

The European dimension. Linked employer-employee data is important information for policy makers in a Europe, where regulations of firms and employment protection of individuals are so different across countries and where there is an obvious shortcoming in Europe compared to the US. It is often believed that the “lacking behind” arises because to a large extent regulations hinder the best allocation of people and their time on jobs in Europe. Thus, many economies have difficulties in employing young people, in general mobility is low and employers are forced to give workers the same wage irrespective of their individual productivity. However, Europe differs widely in respect to different dimensions of this problem. Some countries are less constrained while others are more constrained. It is also recognized that different markets find different ways of optimizing irrespective of regulations. Before policy makers start homogenizing these laws and regulations, a more thorough knowledge of the impact on firms and employment is necessary.

The local labour market dimension. Studies of the impact of labour protection, new wage setting schemes, whether unemployment or early retirement has a firm dimension or not. The latter is important, if you want to address one of these issues politically, since it indicates to what degree policy should be directed towards firms and to what degree it should be directed towards persons.
The management dimension. Using linked employer-employee information gives a new and unique way of analyzing the impact of different management tools on productivity and profit. Good data gives a much better picture of the efficiency of management tools that can be used to increase the overall performance of firms and the well being of their employees for the general benefit of society. This is a relatively new direction but with large perspectives.

Other dimensions. There is a huge potential in linking information on where people have worked with medical information on their current health. This would give new insight in work-related hazards.

The paper contains a discussion of data requirements for different levels of employer-employee analyses. Second, I will survey some of the results that we have obtained using Danish EE-data. Third, I will look at different ways of accessing data. Finally, I will summarize the main conclusions of this paper.

Data requirements

Studying employer-employee relations demands special types of data and different sample sizes compared to the traditional supply-side studies.

Sample size. Where traditional supply-side studies can use relatively small samples of individuals, most EE-studies require several people from each employer, which means a larger sample. Depending on the sample size, this would allow you to estimate wage trajectories (wage functions) for each firm. In principle you can also investigate how large a proportion of the total wage is due to observed individual effects and how much is due to firm effects depending on the number of people observed in each firm. Analyses of the composition of total rewards consisting of observed wage elements and observed job satisfaction are also possible with this data set-up. Models including unobserved heterogeneity on either side need more complicated data as will be explained.

If you want to analyse creation and destruction of jobs and labour turnover, you will need access to data, where you can follow workers employed in the same firms over at least two periods in order to determine how many stay and how many move to other firms. If you want to estimate models including a person and a firm effect, because your model presumes unobserved heterogeneity on both sides, the data has to follow workers over a longer period and at different employers (see e.g. Abowd and Kramarz, 1999). Again, you can use samples, but the sampling frame has to be rather complicated, as it should follow individuals from employer to employer. Furthermore, the data should also contain information on all the other workers in these firms. For this type of study you need to observe some individuals at more than one employer. If turnover is relatively large and you want to follow people over a long period, the sample becomes very large and it will eventually (at least in a small country) require the full set of individuals and firms since workers leaving and entering the sampled firms will be more and more likely to have come from and leaving for a firm outside the original sample. Only access to the full set of observations will guarantee that we can track people from employer to employer and only this type of data will also allow for a full analysis of mobility between firms.

Data-types for EE-analyses. EE-analyses may in principle be based on the same data types as information on individuals, i.e. survey data and administrative records formed as panels or
not. However, there are a number of specific problems with surveys of businesses. Businesses often consider the time and effort as a large cost with the result that attrition (respondents do not respond any longer) may be big. The attrition problem is of course much bigger for panels than for one-wave surveys. One has to bear in mind that all the time firms are asked to fill out surveys and give one or the other sort of information. Adding another may have devastating effects for the possibilities of getting responses to other survey questions. Data are clearly best acquired as part of an official reporting system, tax filing or similar procedures.

Furthermore, there is a specific identity problem with EE-data because firms (unlike individuals) merge, split, are sold or reshaped under a new owner structure. Often the official register number of a plant will change, when one of these changes occurs. However, some of the changes do not constitute real changes and should be reclassified. One method is to use the number of employees, who continue in the plant. If this number is bigger than a threshold of 50-60% of all employees, a correction is made, so that the plant is said to continue.

Linking employers and employees with common identifiers presents possibilities of linking with other data as long as the same identifier is used. One possibility here is to link with special survey information on individuals (labour force surveys, EHCP etc.) and special surveys on employers.

This obviously presents new possibilities for improving the efficiency of employment policies. Another important field is ways of quantifying some of the concepts within management theory in order to see which practises are most efficient with respect to wage systems, promotions and hiring policies.

Most studies consider firms to be homogenous at least within industries. Working with the new data makes it possible to show that this assumption is wrong and that firms are actually rather idiosyncratic even within industries, Bingley and Westergaard-Nielsen, 2000 and 2001.

So far, a few countries have succeeded in creating linked employer-employee data. The Nordic countries have been able to create excellent linked EE-data sets because of their extensive registers of persons. These data cover the whole population and longitudinal in plant and worker dimensions. France has good linked data based on official surveys and registers. Portugal has extensive data from their official survey of firms, which can be linked to data on workers. Many other countries are working at creating these data (see Abowd and Kramarz, 1998, for an overview). In this presentation, I will give a brief introduction to the linking of employer/employee data in Denmark, which is typical for the data available in Finland and Norway as well. Second, I will present some results of our analyses so far. And finally, I will say something about how the important issue on data access can be solved.

**Register-based data – the Scandinavian way**

Statistics Denmark has linked the whole population with employers in the so-called IDA database. The attachment between employers and employees is established through tax records. The November attachment is used by Statistics Denmark as a fix point for the EE-relationship. It is longitudinal and covers the years after 1980. So far we have data until 1998. For research purposes at the Centre for Labour Market and Social Research, Statistics
Denmark has made a selection of variables on individuals and plants. The population data consists of data on wages, number of hours worked, education, training on the job, total experience, tenure, family background (information on children, spouse, residence, wealth), unemployment, sickness benefits, occupation plant and firm identifiers and industry. Economic information (business statistics) has been added on all firms with more than 20 employees. The business statistics come from an annual official survey of business and cover the years 1992-1998. The firm-level data has information on total revenue, profits, value added, investments, exports, loans, physical capital, equity, location and type of ownership. There are longitudinal data about all 6000 firms with more than 20 employees but information also exists on an increasing number of the smaller firms. One of the real advantages of the register-based system is that it allows for supplementing with data from surveys. In this way more qualitative data on employees and employers can be merged with the longitudinal data. So far, we have supplemented with two business surveys. One where we have asked firms about their use of performance-based wages, on-the-job-training, new management systems and changes in their policies. 1600 firms were surveyed and the survey information was merged with register information on employees and firms. We have done the same with another survey on competencies of firms. For details on the variables in these data sets, see documentation on www.cls.dk.

Research on EE-data in Denmark

In order to give the reader an impression of the potential in this type of data, I will use some space explaining what we have got out of the Danish EE-data. First, a number of studies on job creation, destruction and worker turnover where the main dimensions have been the macro perspective and the relation between turnover and business cycles.

The main results of this research can be described as follows: There is substantial worker mobility between firms. Thus, it is found that about 30% of all employees in Danish firms leave their workplace each year. Turnover is somewhat higher among young employees and is reduced with age. Turnover does not differ across most industries (Werwatz et al., 1999). Turnover is found to differ over the business cycle. Thus, it is found that total turnover measured by separations in the private sector differs from about 27% in the beginning of an upswing to about 32% at the end of the upswing.

Similarly, it is found that hires move from about 28% to 34% of total employment during an upswing. Figure 1 gives an intuitive illustration of how hires in an upswing are leading the separations and the opposite in a recession. In the beginning of an upswing, when there are enough vacant workers, hires can happen without raiding the already employed. But gradually, the new hires have to come from the already employed and the number of separations increase. It is also easy to see that recessions are marked by reductions of hires more than increases in separations. Worker dismissals are less correlated with business cycles indicating that firms are reluctant to fire.
Figure 1. Hires and separations in the private sector over the business cycles (own calculations).

Separations are further studied in Frederiksen and Westergård-Nielsen, 2001, where we follow those who leave their jobs. It is found that 2/3 of those leaving their workplace find a new job within the same year and that the remaining 1/3 are not employed the year after. Of the non-employed, 2/3 are unemployed, others are leaving the labour force for a period, fewer are beginning educations, and much fewer leave because of retirement (3% of all separated and 9% of all non-employed). While young people have a higher probability of leaving their job, they are less likely to become unemployed compared to the middle-aged because they are better to find a new job. Furthermore, it has been found that both higher growth and lower unemployment benefits have the impact that more people find other employment and less become unemployed. The data requirements for this sort of analysis are data, where you can follow persons over time and between jobs.

Turning to the management dimension of turnover, it is found that mobility is closely related to the wage and the entire personnel policy of the firm (Bingley and Westergård-Nielsen, 2000); it was found that increasing wage level reduces turnover. However, this conclusion is not straightforward to reach, since there is a serious issue of direction of causation. Statistical methods of disentangling this have to be employed (IV-methods) and for that purpose data on other plants in the same firm was applied. Another line of study has looked at management pay (Eriksson, 1999). It has been found that the level of management pay can be explained by the so-called tournament theory, which resembles the competition at golf and tennis tournaments. Ortega and Werwatz, 1999, find that the wage level is also related to the span of control. Earlier studies on these topics have all been based on studies of single firms, which means that it is impossible to generalize the results. The problem is that single firms are selected in a non-random way. Our results are based on many plants and firms, so results are robust to the sample size. Finally, it is found that the wage differential between firms at the same level and industry is relatively large (Bingley and Westergård-Nielsen, 1998). The wage dispersion for white-collar workers in the US and Denmark is found to be smaller in Denmark though wage dispersion is of similar magnitude in the two countries (Bronars et al. 1999). The data necessary for these studies is detailed information on individuals on the plant level and to some extent the firm relationship as well.
The third group of studies exploits the information on business statistics. Here we have linked information on profit at the firm level with turnover and wage policy. It has been shown that hires are indeed costly for firms in the sense that productivity is lower with more hires (Bingley and Westergård-Nielsen, 2001) and furthermore: higher wage levels and lower wage growth are both associated with higher profit; a workforce having less tenure, other things being equal, is more profitable. In a different context, we have estimated production functions with the attempt to investigate the returns to different educational compositions at the firm level (Frederiksen and Westergaard-Nielsen, 2002). Plant- and firm-level aggregates including production data (for the latter study) are the necessary data for this type of studies. An essential requirement is the link between plants and firms.

A fourth group of studies has used EE-data linked to other data sets. Thus, we have linked the EE-data to a comprehensive register of pension contributions, which started in 1964. With this extra information, we are able to investigate if employees stay shorter or longer during the period 1980-1998. It appears that women tend to increase their tenure, while men shorten their tenure. The result is that both men and women in the 1990s have an average tenure (ongoing) of about 4.7 years. Another interesting result of this study is that the return to tenure has increased, probably as a result of a decentralisation of the wage setting that occurred in the early 1990s in Denmark. We are also currently working on a matched data set that links graduates from our Business School with the EE-data in order to analyse career patterns. The necessary data is the link to persons and plants and firms.

A fifth group of studies uses the information from surveys together with EE-data in order to investigate which firms are most eager using new performance-related wage systems and using new management tools (Eriksson, 2001).

A common result of all our work is that the industry classification used in many investigations is rather meaningless because there are larger differences across firms within one industry than across industries.

I have mentioned these studies in order to demonstrate some direct uses of different types of EE-data. Other researchers working with other national data sets have used them for other purposes depending on the specific characteristics of their data sets. However, I believe that the different applications mentioned present fix points on a chart, where there are many unknown places. The number of applications is of course legio.

The condition for working with this sort of data is that researchers can work with data in a reasonable way. Below, I have tried to describe different ways of accessing data.

**Data access**

Data access has been widely discussed in relation to data on individuals. Confidentiality has always been a major concern in this research. However, data providers have become more relaxed over time with respect to how much information the researchers are allowed to work with and to access from their own computer. First, it has been realized that also researchers have a clear interest in not making disclosures that can identify individuals (they have all their human capital at stake), and second, there are a number of well-known data sets you can actually download or buy very cheap. (PSID(USA), BHPS(UK) and SOEP(FRG)). And third, experience has shown that the use of individual data for research purposes does not present a
big risk. As far as I remember, the only cases of breach of security have occurred when somebody within the administrative system has revealed private information from the registers.

These arguments have been taken up again in relation to discussing accessibility to EE-data. Confidentiality is even taken to be a more serious matter when the data are about firms compared to individuals. The main problem seems to be that individual firms can be identified on a few variables. The belief is that information on location (municipality), industry and size will identify the firm. This is of course in principle the case with large firms, but not with the huge number of small firms. However for the majority of large firms, there is already very thorough information in the public domain. First of all, most stock exchanges demand that extensive information on publicly traded firms is given to investors and to the public. Furthermore, authorities often demand that information on economic performance and owner structure is publicly available. Second, various commercial business registers bring extensive information on the firms, their owners and managers. Third, some countries demand that option plans for managers are public (UK). Some of these points also apply to smaller firms.

A different viewpoint is that we have experienced that it is not easy accidentally to identify a firm, because there are many variables in the data, which are not easily projected on the general knowledge of a firm. This is probably a different matter if one deliberately wants to find a firm in the data, though I am proud to say we have never tried it. Furthermore, data used by researchers are usually lagged 3-4 years and compared to the public sources, I think that most of the information is put dated with respect to sensibility.

Various statistical offices have introduced different methods of accessing confidential data. The first requirement everywhere is that researchers are only allowed to work on data where identifiers have been scrambled or substituted by other identifiers allowing that observations on firms and individuals are consistent over time. Some hand out un-identified data directly to researchers and allow researchers to work on their own PC’s and in their own work environment. The condition is that the principal researcher becomes responsible for any breach of security. Responsibility here means that the researchers can be prosecuted if the rules of disclosure are broken. This is the case in Norway. The advantage is that researchers can work with data in their own environment and this facilitates clearly cooperation with other researchers and comparative research.

A different approach is the creation of safe houses where researchers can come and use un-identified data. Here they have to use the computers and software made available by the statistical office or the owner of the data. This is the method used in USA, the Netherlands and Denmark. The virtue of this arrangement is that data do not have to leave the premises of the owner, but may be analysed from safe houses at different locations. Data can either be in the safe house or remain in the statistical office and accessible through terminals. The inconvenience for researchers is that they cannot work in the usual environment and cooperation is made more difficult. A consequence of this is that research assistants are normally sent to work in the safe house. This means less efficient management of RA’s, time waste, logistically problems and fewer possibilities for senior researchers to have hands on, because other obligations often prevent using much time on transportation between work place and the safe house. The advantage for the statistical agencies owning the data is that they can have physical control with the admission to the safe house and at the same time have
full control of the use of computers and data. The latter is done by controlling printed output and preventing downloads.

A third approach is in between these two methods and could be named virtual safe houses. Here remote access to computers in the statistical office is provided over the Internet. The security here has different elements: access is only possible via specific computers identified by IP-numbers and special terminal server software that can lock the local drives so that no data can be stored here. Nor can data be downloaded to other devices. Each user has a personal password and to each research group belongs a device that gives a check number that is only valid in 1 minute. A similar check number is generated in the server computer. The responsibility of controlling the physical access and use of the data is here given to a person in the local research institute hierarchy who may be prosecuted in case of a breach of security. The owner of the data maintains administration of passwords and usernames. The latter method makes it possible for researchers to work in their usual environment with all the mentioned advantages with respect to management of RA’s, hands-on, co-operation and the building of a data environment. The data never leave the owners premises. With terminal software (Windows terminal server software for example) it is possible submit jobs online, see the output on the screen etc completely like sitting next to the server. Printing of output is possible by means of a special delaying procedure that makes it possible to log and monitor all output. In practice, print is emailed to the user after a screening procedure. Of course, lists of data are not possible to print; neither can single observations be printed.

In Denmark we started out with the safe house method with an office in Aarhus in order to give access outside the main statistical office in Copenhagen. That worked well with the limitations and problems mentioned. However, the experiences were good and Statistics Denmark introduced in beginning of 2002 the virtual safe house method. Only the larger data users were at first considered to be suitable for this access. See Andersen, 2002 for details on the arrangements.

Another related issue has been the data files that can be accessed in this way. Data from the criminal register, from the register of children removed from their homes and some data from the business statistics are currently only accessible from the safe house. There may apply other restrictions to the size of the data matrices put on this system of access. A variant of such restrictions is a situation where data are somewhat scrambled so that you can never get back to the underlying records. Methods have been developed so that averages and moments in the real distributions are maintained after the scrambling procedure. Such procedures are described in Abowd and Woodcock, 2001. The problem is, however, that even the most perfect scrambling of the real data at some point may be used for a purpose, where the scrambling is not appropriate. And this is just what most research is about: using new methods on old or new problems.

A fourth method is where researchers submit their programs to be executed by the data owner. This prevents that researchers can use new methods and prevents the investigative behaviour that is so important when working scientifically with empirical matters. I don’t think that this brings good research out in any country.

For researchers, the work environment is essential. Let me list some points of importance: It is essential that others are working on the same type of data; first, data problems are more easily solved; second, essential knowledge on data and estimation problems can be shared;
errors in data are more easily discovered and solved. In this context it is essential that replication studies can easily be done. New findings should also quickly be incorporated in the research of others.

Whatever method of access is chosen, it is important that researchers can use the superior quality of data, where it exist. For once, Europe has something better than what can be found on the other side of the Atlantic, see Hamermesh, 2002. This situation should be exploited to attract good researchers to work with issues that are important in the European context.

**Conclusion**

The potentials of using linked employer-employee data are large because employer-employee data can be used to do empirical research of what happens inside a firm. This has implications for our understanding of what makes firms expand and reduce employment and the relationship of labour input and output. Results from comparative European research could have a big potential as foundation of new EU employment policies. It has also a huge and so far almost unused potential in selecting best practices in management and remuneration systems. For once, Europe has an advantage compared to the USA.

Within medicine most citizens and policy makers understand the necessity of using all sorts of data. Within the social sciences it has been more difficult and it is even sometimes claimed that it is too dangerous to let researchers use the linked registers. For some reasons this argument has been used more rigorously about EE-data. Of course, all sorts of research using sensitive data require a security policy. The security policy should consist of a combination of admission control (screening of persons who can be granted access) and some limitations in the total matrix of information that the researchers can work with at the same time. But it should be realized that all arrangements require a certain element of trust to the researcher, since no security system can be constructed that fully prevents intended disclosures. Similarly, the researchers should also have some trust in the work of statisticians who create the data. There are examples that policemen and administrative people with access to individual records have been found guilty in disclosures, but there are to my best knowledge no examples of researchers making intended disclosures.

In my opinion, the real danger with register based data is that they are not used to improve the living conditions of the population.

**References**


Bingley, P. and N. Westergaard-Nielsen, Tenure and the returns to firm specific human capital, CLS, 2002

Bronars, Bingley, Famulari and Westergaard-Nielsen, Employer Wage Differentials in the US and Denmark In: Haltiwanger, Lane, Spletzer and Troske (eds.): *The Creation and Analysis of Employer-Employee Matched Data*, North Holland, 1999


Eriksson, 2001


Frederiksen, A. and N. Westergaard-Nielsen, Production and Education Heterogeneity: A Panel Data Study, CLS, 2002


Ortega and Werwatz, 1999