The Use of Register Data in Economic Analysis

By Niels Westergård-Nielsen, Aarhus School of Business, Denmark

1. Introduction

In recent years an increasing number of empirical studies on micro economic behaviour have been based on panel data sets containing information on individuals followed over a number of years. Most often, these data sets have been collected for purely analytic reasons. At the same time in many countries specific computer registers have for years been constructed for administrative purposes. In some countries these registers are merged with each other and have replaced the censuses and most surveys. The cost of surveying the whole population justifies this tendency, but other issues as data reliability may count as well.

Though various type of longitudinal data sets could easily be constructed from existing administrative registers in these countries, the Statistical Bureaux seem in general to be very reluctant to let researchers use these rich sources of information. The reason seems to be that the use of registers has become very sensible to critique from the public calling about (mostly unjustified) fears of "A brave new world". The point made by economists that we are mainly interested in covariance matrices and definitely not in identifying the behaviour of individuals has, however, had a difficult time being heard.

In Denmark, *Danmarks Statistik* (The Danish Statistical Bureau) has for a number of years been working on utilizing the registers as much as possible for general statistical purposes. The last "real" census was held in 1970. Since then, "censuses" have been carried out using the registers. This change has been helped by the fact that the country is thoroughly organized in the sense that one can hardly do anything, getting a wage income, becoming married, getting sick, having children, moving, receiving any transfer pay or ultimately – die without being registered by the social security number (cpr.).

As a by-product, a research team in Aarhus has been able to work on a sample of a longitudinal data set based on several administrative registers made for statistical purposes. This has given us some knowledge about the pros and cons using administrative data in economic analysis. In the following I will discuss some of our experience working with longitudinal data and more specifically I will summarize some of the most recent results using the data.

2. Longitudinal data

The use of data on individual behaviour is relatively new in labour economics. Historically, it started with analyzing survey data coming from surveys most often designed and performed by other social scientists.

2.1 Longitudinal versus cross-section data

In genuine longitudinal data the same people are followed over a prolonged period with recurrent observations. This may happen either in a panel that is surveyed again and again or in the form of combining administrative data. Generally, these data are superior to cross section data. Observing the same people a number of times makes it possible to control for unobserved heterogeneity in the population. Thus, the fact that a person persistently has a higher income than other seemingly identical individuals tells us that this person has better abilities or other unobserved capabilities. If this fact is not taken into account, inefficient and biased estimates may be expected.

A classical example is the effect of education in a wage regression. Neglecting unobserved differences causes upward biased coefficients to education, because ability is believed to be positively related to education as well as wage. Using panel data, where identical persons are followed over a number of years, makes it possible to estimate models using the differentials between years, thus differencing out any constant (or even random) effect from unobservable variables. (See *Hausman/Taylor*, 1981 and for an application on LDB, see *Pedersen/Schmidt-Sørensen/Smith/Westergård-Nielsen*, 1988).

Consequently, there has in recent years been a clear upswing in the construction of panel data sets in USA and Europe. In USA the well known PSID and NLS panel data have existed for years. In Holland, Germany and UK various panel data sets based on surveys of 1,000 – 12,000 people are now covering 2 – 4 years. At great length these data sets are based on surveys. In Denmark, a panel data set based on administrative data is now covering 11 years.

2.2 Register data versus surveys

These two types of panel data have their pros and cons. Surveying people has to deal with the problems of a declining response rate, attrition and an effect from having the same people to answer the same battery of questions time after time. These points may clearly limit the number of interviews. One advantage is that the researchers may provide the questions that fit best into their analytical framework.

In the administrative data set repetition is no problem since most of the registers by statute are updated regularly.

Looking at the available data sets, the cross section data may come either from surveys conducted for a specific purpose or the data may come from surveys conducted regularly by the respective statistical offices. Often such data sets include some retrospective questions that help to identify transitions and to reveal information on the background variables. In other cases – the EC labour force surveys for example – people are followed three to four periods after a rotating scheme.

Using retrospective questions raises a reliability problem. Can people remember what happened six months ago, can they remember their wage rate etc. A recent paper by *Torelli/Trivellato* (1989) investigates the consistency in the answers to a EC labour force survey for Italy confronting the retrospective answers at the survey time with the answers at previous surveys. This shows a clear bias towards "improving the history". A related question is how reliable are other types of information, also given by the respondents, for example about wage rates. A paper by *Greenberg/Halsley* (1983) raises serious doubts on this point. Another paper by *Bound/Brown/Duncan/Rodgers* (1989) goes a step further. They investigate the reliability of answers to the NLS by confronting the answers of the employees and the employers with the "true" answers in the files of the firms. Their results show also that people tend to "improve" their record if it is just a few months old. All these studies indicate the existence of serious measurement problems with surveys.

The most serious problem with administrative data seems to be that data is collected for administrative purposes and does not always reveal the most relevant information for analytical purposes. The measurement problem then shifts to a problem of data registration.

In general, register data gives answers on all the items that are necessary for the administrative purposes of the data. But this is not always identical to those questions you would pose for analytical purposes. An example is the amount of work measured in hours. Often the register data is weak on this point. The survey on the other hand may have a partly reliable answer about the normal working hours. At the same time one may assume that it has relatively unreliable data on wage income. A way out of the dilemma is to supplement the register data with data from surveys. This can be done either directly, where data on the same persons is merged from different registers. However, this is seldom possible. The alternative is to use statistical merging, which could be done by estimating an hour function in a data set where this information exists and then use the predicted number of hours in the register data. A necessary requirement is obviously that some of the same variables are present in both data sets.

2.3 The Confidentiality issue

One of the major obstacles to forming data sets from register data or letting social researchers use the data has been the confidentiality question.

There are several ways to secure confidentiality. The first is that the statistical bureau only hands out data, where the personal identifiers have been scrambled in such a way that the data as such can be updated. The second is that variables are censored with respect to details, so that persons cannot be identified by running cross tabulations on several detailed variables.

The fear has mostly been that somebody from the public should get the impression that it was possible to get information on individuals out of the data with an increasing public suspicion towards public registers as the effect.

3. The Danish Longitudinal Data-set

The complete data set consists of a random sample of approximately 240,000 adults (approximately 5 % of the Danish adult population between 15 and 74) with information taken from a number of Danish administrative registers[1]. A common registration number (cpr) is used to link the different registers. After forming the sample, this identifier is encoded. The data set contains information on a large number of demographic, educational, and labour market variables as well as information on income and wealth for the years 1976–85. However, some of the variables have not been available right from the beginning, but have been added in subsequent years as an effect of general improvements of the registers. Thus, the data set only contains detailed (weekly) information on unemployment from the beginning of 1979.

The total data set consists of a little more than 70 variables (in average) for about 240,000 individuals for 10 years (in 1989). That is 168 million data which is stored in RDB-files. This file is obviously too big for most statistical packages and computers. As a consequence, a much smaller data set of about 11,000 individuals is constructed in SAS. The sub-sample is formed around a group of approximate 6,000 persons that are married to another participant in the sample in the years 1979–1981. The rest of the sub-sample is randomly chosen among those who were not married in these years. By selecting this sub-sample, the household concept can partly be utilized. The appendix contains a variable list and summary statistics concerning the SAS-files.

3.1 Quality of data

The quality of the variables in LDB is in general very high but may vary according to the procedure for collecting data. First, key data as sex and age is used and checked in numerous administrative routines. Second, all tax information is checked by employers and employees who have conflicting interests in misreporting. Third, other variables as industry and education come from different registers which are used for a number of statistical purposes, involving double checking.

[1] A detailed description of the data set is given by *Westergård-Nielsen* (1984).

3.2 Manipulation of data

Some of the original variables are, however, not directly applicable in economic analysis. Therefore more relevant variables have been formed from the register information. That is the case for experience, labour supply, wage rates and spells of unemployment and employment among others. In the following I will give a brief description of how we have constructed some of these variables.

3.2.1 Experience

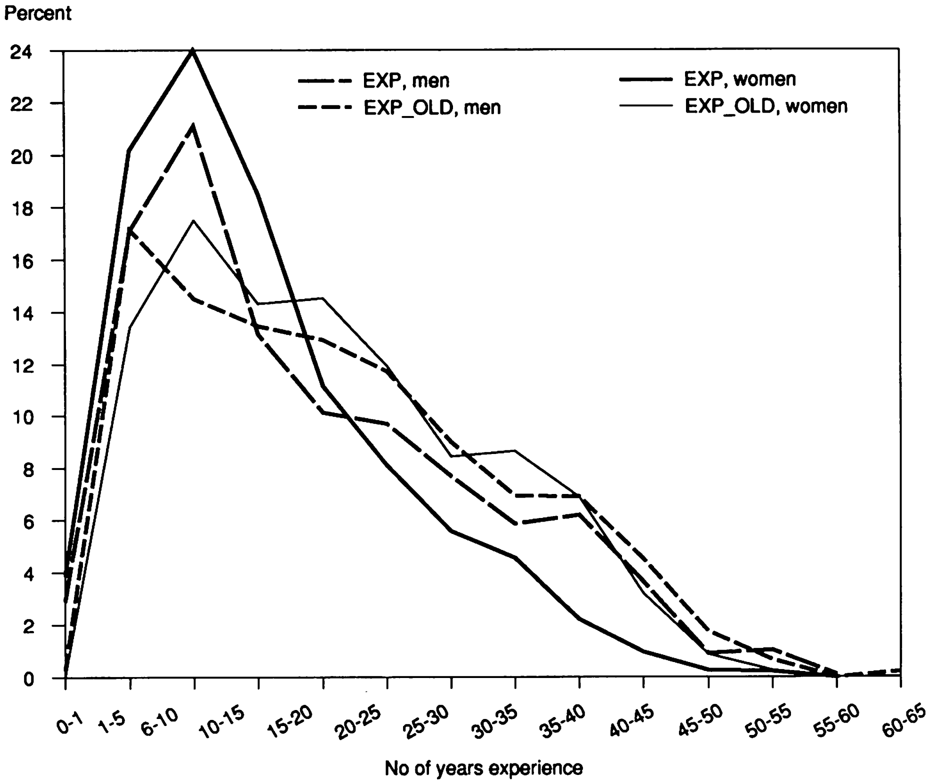
Most data sets do not contain any direct information on labour market experience. Instead a constructed variable based on the number of years in school and age is used, see for example *Chowdhury/Nickell* (1985). The presumption is that people follow the same employment/educational pattern over the life-cycle. This may be a sensible thing to do for men, but much less sensible for women who often have had periods with no or less participation.

LDB has information on actual working experience as wage earner from the supplementary pension scheme (ATP), which started in 1964. The hourly contribution to the supplementary pension scheme is paid by the employer as well as by the employee. The amount is fixed by law and reflects on a ladder scale the amount of work for the individual. Since the employee and the employer have conflicting interests in misreporting the true amount of hours, the data is expected to reflect reality. The difference between the true and the calculated measure in Figure 1 makes it clear that using the calculated experience measure (CAL_EXP) increases the experience for women, inducing errors. The remaining problem is that the ATP only reflects the hours on a ladder scale. The conjecture is that this induces an error in the measurement of the true number of hours for part time workers.

Furthermore, the information from ATP on the number of hours worked in a year is used to calculate the hourly wage rate from the annual wage income in the tax registers. The resulting wage is subsequently filtered to catch out-layers. Out-layers may come about because the number of hours is not fully represented by ATP. This may either be due to the ladder scaling of ATP or errors in the ATP accounts for ATP. In these few cases hours worked is calculated from the number of days unemployed. After this alternative calculation a little less than 7 % of all calculated wages are stopped by the filters[2]. Another type of problem is related to those who have been unemployed all year or have just entered the labour market. For those, one can either use their wages from the preceding year

[2] The sources to the errors are: 1. Normal hours (i.e. 40 hours per week taking account of the calendar variations, changes in vacation etc) may vary for groups and individuals. 2. ATP is measured on a ladder scale with few steps. Upward bias arises if people work less than indicated by the steps. 3. ATP does not include overtime which means that overtime gives an upward bias.

Figure 1:
Calculated experience (CAL_EXP) compared with the ATP-based experience measure (EXP)

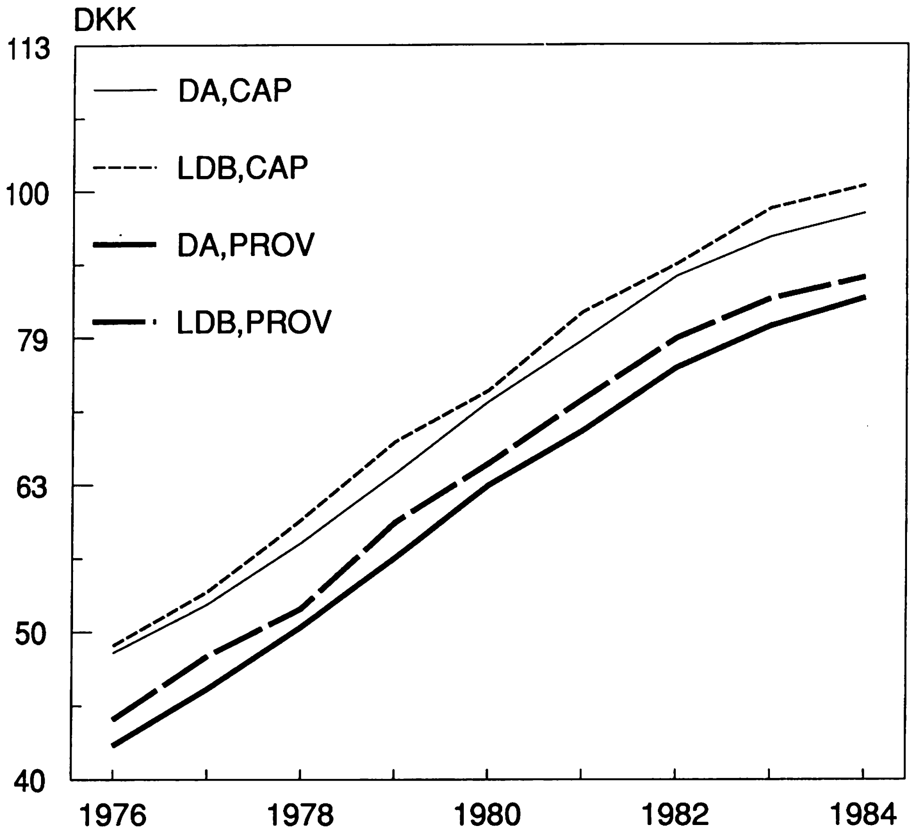


(if they exist) corrected for general increases, or one can use estimated wages depending on the purpose.

The final question in this connection is naturally, how well do we calculate hourly wages? Using the only other Danish source, i.e. wage rates collected by the Danish Employers Association from its members as a bench-mark produces the graph in Figure 2. Similar graphs can be made for other groups.

The graph seems to justify the use of the calculated hourly wage. One finding is that the calculated wage rate for men is between 3 and 7 % higher than the wage rate from the Employers Association only covering those employed as members. This is probably due to over time or the other factors mentioned above. For unskilled women the differences are more ambiguous probably due to problems measuring the correct number of hours for part time workers because of the use of a ladder scale as the relation between the true number of hours and the ATP.

Figure 2:
Comparing the calculated wage rate from LDB with wage rates from the Employers Association (DA) for skilled male workers in manufacturing for the Capital (CAP) and for the Provinces (PROV)



In any case, the quality of the calculated wage has to be compared with what could have been obtained from a survey where you use less reliable data for the wage rate. In that context, I think it is worthwhile using the reliable wage income from the tax register and work on improving the hours function following the lines in section 2.2.

3.2.2 Spells

The unemployment register has on a weekly basis information on the fraction of the week for which unemployment benefits are paid. For analytical purposes it is, however, necessary to convert these fractions of weeks into genuine spells of unemployment.

Ultra short spells lasting only a few hours have been excluded when constructing spells. Thus, by definition a week with more than 50 % unemployed hours, i.e. more than 20 hours of unemployment, is classified as a week of unemployment, while a week with less is employment. Since the smallest time unit is a week, all spells are defined to start at the beginning of a week. Similarly, all spells are defined to end at the end of a week.

4. Recent Results from LDB

In this section I will give a few examples of some recent uses of the data.

4.1 Types of spells

Weeks with a fraction of unemployment of 0 are under most conditions taken for employment. However, it may also cover the fact that the person has not been available for a job because of temporary or permanent withdrawal from the labour market, sickness leave, or vacation paid by a former employer. The first is checked by drawing on other information in the data set on labour market status in the current, the previous, and the following year. If the person is in the labour force in the surrounding periods he is as a main rule characterized as employed in a period with a fraction of unemployment of 0.

Furthermore, it seems obvious that short unemployment spells which are due to the vacation system should be re-classified so that the classification reflects if a person has a job or not. This is accomplished by simulating the vacation rules for all persons in the sample and subsequently searching for spells that fit the simulated results[3].

The next step is to identify the unemployment spells that end in re-employment with the same employer. Again, certain assumptions have to be made. Since there will always be an element of judgement, it has been chosen to apply a relatively restrictive strategy, which rather identifies too few spells as temporary, than the opposite.

As a consequence only unemployment spells directly following an employment spell can be considered as temporary lay-off. This is accomplished by letting the algorithm examine the unemployment spell together with the 3 previous and the 3 subsequent months. If only one employer is registered as having paid contributions to the supplementary pension scheme in this period, the spell will be characterized as temporary lay-off. There are, however, a few caveats in the procedure.

First, the employers are identified by an employer registration number, that is made for administrative and not for analytical purposes. Thus, changes in the legal ownership of the firm will lead to a change in registration number. As a consequence, we will be mislead to classify too many on-going spells as

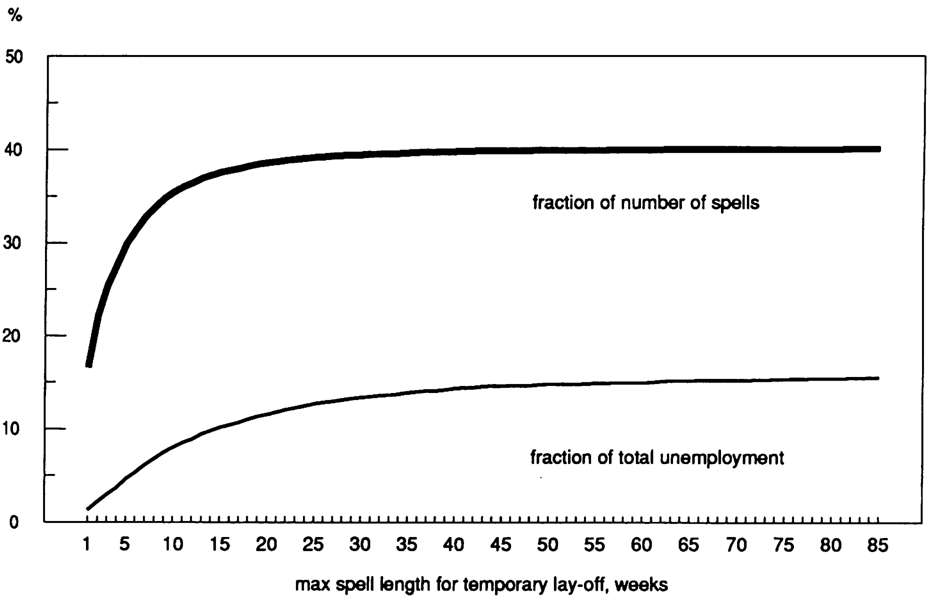
[3] For further details, see *Jensen/Westergård-Nielsen* (1988).

permanent lay-offs[4]. Secondly, unemployment spells that are not terminated at the end of the observation period (i.e. 1984) or are initiated before the beginning (i.e. 4th quarter of 1978) will not be classified as temporary lay-offs. Thirdly, unemployment spells that end with recall, but where the person has had a supplementary job with another employer within 3 months before and after the unemployment spell is not classified as temporary lay-off.

The classification algorithm therefore leads to a downward bias in the estimate of the number of temporary lay-offs.

The result of this effort is that 40 % of all spells of unemployment end with re-employment with the previous employer. Because temporary lay-offs are more prevalent for the shorter periods, they amount to 16 % of all unemployment in Denmark. These numbers both depend on the choice of a limit on the duration of unemployment after which we will not classify re-employment as temporary lay-off. Figure 3 shows that as long as the cut-off duration is above 20 weeks, the temporary lay-offs are still a substantial part of all spells and of all unemployment.

Figure 3:
The frequency of temporary lay-off depending on the length of the unemployment spell before re-employment

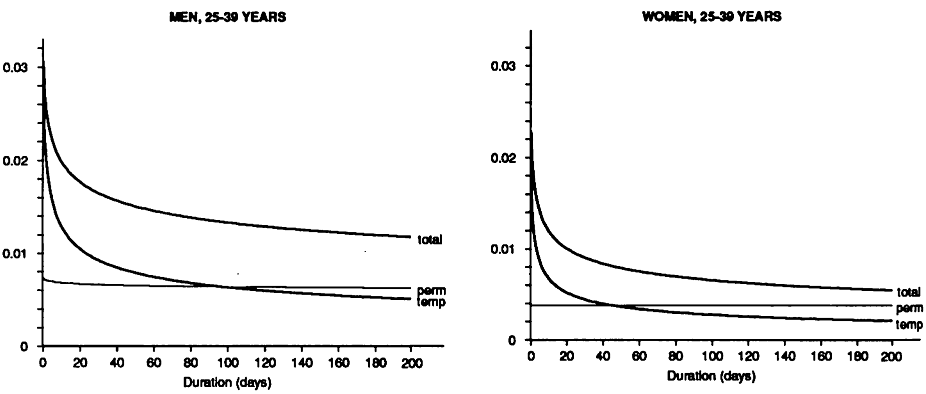


[4] An alternative approach would be to use other sources to identify firms that despite change of registration number are identical. *Hansen* (1985) has done that for a completely different sub-sample of Danish firms, where he considers a firm as identical across numbers, if it in two consecutive years employs as least 40 % of the same employees.

These data have been used to estimate a competing risk model for the risk of being re-employed or being employed by a new employer (see *Jensen/ Westergård-Nielsen*, 1989). The estimations show that the decreasing hazard function (a Weibull function) for leaving unemployment found in many studies of the duration of unemployment, is due to a decreasing hazard function for becoming re-employed with the previous employer and a constant hazard function for becoming employed with a new employer.

Our results indicate that neglecting the different types of lay-offs may lead to a serious bias in the results and conclusions. Generally speaking, all the negative duration dependence in the total hazard rate for re-employment (neglecting temporary lay-offs) seems to be due to the negative duration dependence in the temporary lay-off hazard rate.

Figure 4:
Hazard functions for men and women age 25 to 39 for the two competing risks of getting a new job and returning to their previous employer



4.2 Wage functions

Another line of study has been estimations of wage functions where different viewpoints have been used. In a study of the gender differences, *Smith/ Westergård-Nielsen* (1988) estimate a human capital model specified as a fixed effect model that takes into account unobserved factors such as individual skills, the tendency to have an extra job and others. Despite the observation from macro statistics that women have had the highest observed increases in wage rates, the models show that this increase is mainly due to an improvement in their background characteristics and that men still receive a higher return to their characteristics. The main difference between genders appears to be that female workers do not in general get any return to their experience, contrary to men. The estimates also show negative effects on the wage rate of previous spells of unemployment. The main conclusion therefore seems to be that the wage gap in Denmark between genders is found to be increasing during the period 1976–1984.

In another recent paper[5], it is investigated if the wage differential between the private and public sectors has increased in recent years. Here a Hausman & Taylor method with instrumental variables for schooling is used to get an unbiased coefficient to schooling so that any likely correlation between schooling and the unobserved effects from ability and motivation is avoided.

The return to education is found to be 5 % to men in the private sector and 6 % for men in the public sector. For women in the private sector the return is not significantly different from 0 while it is about 2 % for women in the public sector. The overall finding is that there has been an increasing wage gap between the private and public sectors. It is also shown that the wage rate in the public sector depends much more on experience (or seniority) related growth than in the private sector.

5. Future extensions of LDB

It is the hope that it will be possible to extend LDB with not only more years but also with more variables. The value of the database is clearly increasing with the period it covers. In the near future, LDB will be expanded with data on labour market educations for skilled and semi skilled workers. The purpose is to see, if this activity has any short or long term effect on the risk of becoming unemployed and the wage rate. Similarly, data on cohabitant, non-married couples will be added in order to improve labour supply estimations.

An extension in the future will be to link individuals with information about firms, so that hypotheses from efficiency wage theory about the relationship between the wage policy of the firm and its profitability can be tested. At the moment the Danmarks Statistik is working on establishing this link.

6. Summary and Conclusions

I hope to have demonstrated that register data are highly reliable, and in general of a very good quality compared to most survey-based data sets.

Until recently economist have not worried too much about the reliability and consistency issues with survey-based panel data simply because they have only rarely been collecting the data themselves, and probably also because alternatives

[5] *Pedersen/Schmidt-Sørensen/Smith/Westergård-Nielsen* (1988).

were scarce, if existing at all. With register data the situation is different. The difficult part here seems more to be to get access to the data and to interpret the data correctly.

The major shortcoming is that data are collected for administrative purposes, which in some cases differ from the analytical purposes in an economic approach. However, it has been demonstrated that it is possible in many cases to transform the original data into variables of more relevance in economic analysis with little or no loss in efficiency.

In the long run I have the vain hope that the statistical bureaux will let researchers use the statistics on individuals in the same way as they currently are letting researchers use macro statistics.

References

Bound, J./Brown, C./Duncan, G./Rodgers, W. (1989): Measurement Error in Cross-sectional and Longitudinal Labor Market Surveys. Results from Two Validation Studies. In: *Hartog, J./Ridder, G./Theeuwes, J.* (ed.): Panel data and labour market studies, forthcoming North-Holland.

Chowdhury, G./Nickell, S. (1985): Hourly Earnings in the United States: Another Look at Unionization, Schooling, Sickness and Unemployment Using PSID Data. Journal of Labor Economics, *3*, No. 1, 1985.

Hausman, J.A./Taylor, W.E. (1981): Panel Data and Unobservable Individual Effects. Econometrica *49*, no. 6, 1377–1398.

Jensen, P./Westergård-Nielsen, N. (1988): Ledighed, midlertidig hjemsendelse og ferie. Studies in Labour Market Dynamics, Working Paper 88–4 (in Danish).

Jensen, P./Westergård-Nielsen, N. (1989): Temporary Layoffs. In: *Hartog, J./Ridder, G./Theeuwes, J.* (ed), Panel data and labour market studies, forthcoming North-Holland.

Pedersen, P.J./Schmidt-Sørensen, J.B./Smith, N./Westergård-Nielsen, N. (1988): Wage differentials between the public and private sectors. Studies in Labour Market Dynamics, Working Paper 1988–5. The Aarhus School of Economics and Business Administration.

Pedersen, P.J./Westergård-Nielsen, N. (1987): Multiple spells of unemployment – the Danish experience. In: *Pedersen, P.J./Lund, R.* (ed.), Unemployment: Theory, policy and structure. de Gruyter.

Smith, N./Westergård-Nielsen, N. (1988): Wage differentials due to gender. Journal of Population Economics *1*, 115–130.

Torelli, N./Trivellato, U. (1989): Youth unemployment duration from the Italian labour force survey: Accuracy issues and modelling attempts. European Economic Review *33*.

Westergård-Nielsen, N. (1984): A Danish longitudinal data base. In: *Neumann, G./Westergård-Nielsen, N.* (ed.), Studies in labour market dynamics. Springer Verlag.

List of variables for working sample including 1976–1985

VARIABLE	TYPE	NAME	NO OF OBS	MIN	MAX	MEAN
Registering no	integer	LNR	99860	2	252207	106507.9
Year	integer	AAR	99860	76	85	80.75046
Sex	marker	KQN	96005	1	2	1.499182
Woman	marker	KVINDE	96005	0	1	0.499182
Place of residence	marker	SMKNMA RK	95450	1	2	1.659927
Province/Capital	marker	PROV	95450	0	1	0.659927
Sickness marker	marker	SYGEMAR K	95450	0	1	0.065846
Year of birth	classes	FQDAARG	96005	1	41	8.267684
Age	integer	ALDER	96005	13	78	42.13806
No of children, 0-2 years	integer	BARN1	50115	0	3	0.092407
No of children, 3-6 years	integer	BARN2	50115	0	3	0.143889
No of children, 7-14 years	integer	BARN3	50115	0	5	0.318507
No of children, 0-17 years	integer	BARN4	50115	0	7	0.675646
Children 0-6 years,marker	marker	BQRN06	70271	0	1	0.157917
Children 0-17 years,marker	marker	BQRN17	87381	0	1	0.342935
Married in sample.	marker	GIFTIGR	70073	0	1	0.55064
Reg.no of spouse	integer	LNR2	57242	0	247483	69793.61
Taxed with spouse	marker	SAMSK	95450	1	2	1.58571
Gross income of spouse	integer	BRUTTO2	77058	0	220000	42965.29
Occupation of spouse	classes	ARBST2	77058	0	9	2.30538
Wage earner income	integer	AINDK	99240	0	1000000	68785.69
Income as self employed	integer	BINDK	95450	-200000	1000000	13655.18
Gross income	integer	BRUTTO	99240	0	1000000	83582.77
Taxable income	integer	SKPLINDK	99240	-200000	1000000	66463.71
Wage income	integer	LQN	99240	0	1000000	56657.07
Surplus, own firm	integer	OVSKVK	99240	-200000	1000000	8116.495
Wealth	integer	FORMUE	95450	-1000000	1000000	96008.06
Unemploym.benefit	integer	ARBLHU	99240	0	150000	3651.189
Other transfer income	integer	OFFYD	95450	0	110000	5316.029
Surplus own house	marker	OVSKEJD	70073	0	1	0.299488
Deficit own house	marker	UNSKEJD	70073	0	1	0.097299
Housing assessment	marker	EJDVARDI	60039	0	1	0.288213
Positive wealth	integer	AKTIV	70073	0	1000000	146514.3
Negative wealth	integer	PASSIV	70073	0	1000000	86672.04
Ongoing education	classes	IGUDD	35070	101	3024	2709.630
Graduation year	int	EAFGAAR	31801	0	99	78.51061
School education	classes	ALMUDD	77026	0	3030	434.6009
Vocational training	classes	ERHVUDD	76991	307	3030	1984.359
Industry	classes	ERHVGR	95450	0	99	23.65903
Occupation	classes	ARBST	95450	0	9	3.827721
Industry, extended cat.	classes	BRANCHE	60161	0	99	44.03105
Occupation, extended cat.	classes	STILGR	60161	-1	550	443.9100
Supplem.pension points	marker	ATP	95449	0	30	4.705392
UI-fund,Union	classes	AKASGR	68075	1	8	6.252736
Insurance category	classes	FORSK	68075	0	3	1.937804
Full time insurance	marker	HELFORS	68075	0	1	0.581065
Part time insurance	marker	DELFORS	68075	0	1	0.084304
Non insured,public help	marker	HENVIST	68075	0	1	0.026001
Non insured	marker	EJ_FORS	68075	0	1	0.30863
Degree of unemployment (weekly)	integer	LGR	47227	0	1000	109.3249
Censored wage rate	marker	TRUNK	99860	0	1	0.004887
Wage rate	real	WAGE	54681	19.14894	649.5633	68.45334
Supply of hours	real	UDBUD	95172	0	5520	959.5067
Received UI-benefit per hour	real	UCOMPH	15921	0	2454.373	40.83702
Received UI-ben/wage	real	UCOMPGR	12890	0	45.92737	0.810623
Potent. UI pay per hour	real	UCH	54681	15.73885	50.25	40.06246
Potential replacement ratio	real	REPLAC	54681	0.08222	0.9	0.716856
Schooling,no of years	integer	S	82076	7	18	10.13688
Experience as wage earner	real	ERF	82916	-10.0180	59.66667	17.71878