Longitudinal Administrative Registers in Economic Research – A Norwegian Experience[*]

Knut Røed

1. Introduction

Administrative registers have begun to make their way into microeconometric labour market research on a fairly large scale, particularly in smaller countries with developed and mature registers. They are used to examine a wide range of topics; such as labour supply, retirement behaviour, educational choice, unemployment and job search behaviour, absenteeism, health problems, and the intergenerational transfer of socioeconomic status. Typical datasets involve millions of observations across time and space, with a lot of explanatory variables describing the subjects/individuals under study. However, many applications so far have barely scratched the surface of the information-content in the data. This 'utilisation deficit' has to some extent been caused by a number of more or less trivial obstacles, such as inadequate compatibility between different register providers, legal restrictions on the usage of sensitive information, administrative inability or unwillingness to supply the required data and user knowledge to the research community, and lack of computational power to handle (not to say analyse) the data. But progress has also been hampered by a shortage of (readily available) appropriate statistical techniques. A typical register data 'problem' is that it is difficult to see the wood for the trees; the researcher becomes victim of a kind of 'excess information syndrome'. Consequently, many register studies are based on small (random) samples of entrants into a single easily defined state, such as unemployment. As a result, the associated econometric work has typically been similar in kind to that of survey-based analyses. More recently, however, advances in statistical theory, computational techniques and power, together with the steady accumulation of experience and knowledge about the data generating and recording processes, have opened up new avenues for promising research.

[*] This is a shortened, revised and updated version of Røed and Raaum (2003), prepared for the conference 'Longitudinal Social Surveys in an International Perspective' in Montreal January 25-27 2006. Correspondence to knut.roed@frisch.uio.no,

During the past 10-15 years, researchers and affiliates at the Ragnar Frisch Centre for Economic Research in Oslo have invested heavily in the advancement of register based microeconometric research, with emphasis on individual labour market behaviour. This investment has taken the form of a comprehensive labour market *event history database*, realized by merging available Norwegian administrative registers. In its present form, 'The Frisch Centre Database' covers the whole Norwegian population and contains information about labour market status from 1992 through 2003. For each individual (and to some extent at each point in time) the records contain information on demographic factors (age, gender, country of birth), schooling and educational attainment, family situation, income, previous income, and work experience. In this paper, we want to spell out some of the opportunities for scientific progress that, according to our experience, lie hidden in large-scale register data, and share with the broader research community some ideas for future research. The paper is *not* a survey, i.e., it does not attempt to provide a representative overview of register-based labour market research. The strategy of the paper is to use the Frisch Centre Database as a sort of 'case' to illustrate a number of 'lessons' and ideas believed to be of general interest.

2. The longitudinal design of register data – time and state aggregation

Administrative registers are typically designed to serve the needs of public institutions. And since no administrative body has any valid reason for tracking the labour market career of individuals in full detail, such individual histories are not recorded in any single register. Consistent event histories can nevertheless be constructed on the basis of different administrative information sources. In practice, this is a giant jigsaw puzzle, where the different bits and pieces do not always fit together. While some labour market activities (such as insured unemployment spells, disability spells, long term sickness) can be traced virtually day-by-day, other activities (such as ordinary jobs) must often be identified on the basis of *annual* tax records. Educational activities are typically recorded semester by semester.

Two major issues in the construction of event histories are those of *time aggregation* and *state aggregation*. In the construction of the Frisch Centre Database, we have chosen the *calendar month* as the basic measurement unit for time, reflecting that this is the highest frequency that can be observed with a reasonable degree of accuracy. The database has a point-in-time structure, in the sense that labour market states are updated by the end of each calendar month. This has the implication that all state-durations (spells) are *interval-censored*, i.e., events are not exactly timed – the researcher only knows in which month an event occurred. More importantly, some spells are also *left-truncated*, i.e., they are not observed at all if they start and stop between two observation time-posts.

Our experience suggests that little information-content is lost as a result of interval-censoring; see Gaure *et al* (2005). However, it is extremely important to take the precise nature of data-generation, including the interval-censoring and left-truncation problems, into account in the design of the statistical analysis. If interval-censored data are used *as if* they were continuous in, e.g., hazard rate modelling, serious bias problems arise.

The issue of *state aggregation* is closely tied to the analytical context, and at this point the registers typically offer a large scope for discretion. In some cases, two or three 'main' states are sufficient, e.g., employment, unemployment and out of the labour force. In other cases, a more detailed state space is required. For example, in analyses of unemployment duration, it may be important to treat different types of unemployment (insured unemployment, uninsured unemployment, programme participation, etc.) as distinct states.

Administrative registers offer a wide range of information about the individuals and their families, such as age, gender, schooling and educational attainment, income history, country of birth and citizenship, place of residence, number of children, diagnosis (in case of long term sickness or disability), work preferences (in case of unemployment), etc. The Norwegian registers also identify biological parents, and they include school/university identifiers for students and firm identifiers for workers. Since registers cover the whole population, there is no sampling error involved. The number of observations is rarely a concern. And there is no attrition bias to worry about.

The huge set of characteristics and outcomes is of course a great asset from the researcher's point of view, both because many of these variables may be of interest in their own right, but also because they capture much of the heterogeneity in individuals' preferences and opportunities, and hence make it easier to establish chains of causality. The quality of some of these variables is questionable, however. As a rule of thumb, we have found that information that is essential for its administrative purpose tends to be of high quality, while more secondary information may be unreliable simply because there is less need to check and update the records. For example, the tax register records annual income, as well as starting and stopping dates for all jobs. The former of these pieces of information is highly reliable, since it directly determines each individual's tax liabilities. The second is less reliable, since errors typically have no administrative consequences.

3. Uncovering truly independent variation in explanatory variables

A fundamental problem facing virtually all microeconometric applications is that of *separating causal effects from the influence of unobserved characteristics of the* *individuals or of the environment in which they operate.* The driving force behind this problem is that variables for which causal effects are to be identified can rarely be considered exogenous or independent of unobserved individual characteristics. For example, the correlation between education and employment performance (e.g., income) not only reflect that education *causes* performance, but also that (unobserved) attributes that affect performance (such as ability) have affected the choice of educational attainment. Similar arguments apply to the relationship between unemployment benefits and unemployment duration, between various forms of treatments and outcomes, between retirement incentives and retirement behaviour, between wages and labour supply, between social status and health, and so forth. Consequently, the important causal parameters characterising these relationships cannot be established beyond doubt. Although we frame this problem here in a reduced-form-setting, it also applies to theoretically founded choice models. In attempts to estimate deep structural parameters on the basis of observed behaviour, unobserved heterogeneity will be 'thrown' into the parts of the structural model that provide the 'best fit' to the data. Consider for example the issue of recovering preferences over consumption (income) and leisure on the basis of cross sections of observed wages and hours of work. If there are unobserved job attributes that are correlated to the wage level (e.g., that jobs paying a high wage also tend to more interesting and exciting), the estimated structural parameters may be biased (e.g., we may over-estimate the wage responsiveness of labour supply).

In this section, we explain how register data can provide some virtually purely data-based solutions (i.e., solutions that are not dependent on either questionable distributional assumptions or on non-testable theoretical restrictions) to the problem of unobserved heterogeneity. The solutions are based on the idea that there indeed exist *sources of truly independent* variation in the explanatory variables of interest. This kind of variation appears in at least three forms: First, public bureaucracies are in need of relatively simple and verifiable rules and regulations, and can often not avoid generating individual economic incentives that contain elements of random assignment. Second, exogenous discretionary policy changes take place from time to time, either because of unforeseen events or by pure accident. And third, the elapse of *time* itself may provide a useful source for the separation of causality from unobserved heterogeneity, since the *calendar times* at which certain events occur may often be exogenous. We now look at these different sources of independent variation in turn.

Public regulations often rely on practices that, at least to some extent, appear *arbitrary from a behavioural viewpoint*. Typical examples are the use of *calendar year* income as the basis for computation of taxes and transfers and the application of various thresholds in tax and benefit systems. These kinds of practices may provide the researcher with event histories that are 'equal' up to a certain point at which the arbitrariness of the rules suddenly endow different individuals with different incentive structures. This phenomenon is of course not uniquely linked to register data. But the arbitrary component in economic incentives may often be small, relative to other sources of variation that are suspected to be correlated to unobserved heterogeneity, except, perhaps, for a very tiny group of people. Register data often make it possible to identify these kinds of small groups and isolate the random component in the economic incentives they face. The basic idea can be illustrated by an attempt to identify the elasticity of the escape rate from unemployment with respect to the unemployment insurance (UI) replacement ratio, presented by Røed and Zhang (2002; 2003). This is a well-known topic with a long history of empirical research; yet the issue is far from settled. The identification problem in this case arises from the fact that the size of UI benefits to a large extent is determined by the individual's own past behaviour, hence it cannot reasonably be assumed independent of unobserved characteristics. Fortunately (from the researcher's point of view), bureaucratic considerations are responsible for introducing some elements of random-assignment-like variation as well. The main source of 'random' variation in the benefits allocated to Norwegian job seekers is that the Public Employment Service (PES) uses income from the previous calendar year as the basis for calculating benefits. In some cases, this implies that two persons with exactly the same background (in terms of job length and associated income) may receive different benefit levels simply because their unemployment spells began in different months. This procedure has of course no behavioural justification (it is motivated by verifiability considerations), and it produces a variation in benefits that is similar to the way the tax level depends on the extent to which a given income is spread out across tax years. In the case of unemployment, individuals cannot time their spells optimally as persons who have quit their last job voluntarily do not receive benefits during the first part of their spell. By relying on this (and some other minor sources of independent variation in benefits) *only*, Røed and Zhang (2002; 2003) show that the Frisch Centre Database not only makes it possible to identify the average benefit elasticity relatively precisely, but that the information content in the data also suffices for identifying the way disincentive effects interact with business cycles, spell duration and the economic resources of the household.

The second source of independent variation in explanatory variables is that of exogenous shifts in public policies or practices. Various kinds of 'natural experiments' have indeed become popular tools for identification of causal effects. Such experiments arise, for example, when policy reforms are implemented such that they affect some individuals, but not others (and such that the affected individuals are not systematically different from the unaffected individuals), or when reforms are introduced in a stepwise procedure (e.g. state by state, or municipality by municipality), such that the timing of the reform varies in a 'random' manner. The 'natural experiment approach' is of course also open to research based on survey data. But systematic 'maintenance' of registers ensures that no such natural experiences are 'lost'. Many experiments occur 'by accident', and are therefore not announced in advance; hence no data collection activities can be set up in time. In this sense, the registers operate as a sort of permanent surveillance camera, which 'recordings' become valuable when something of interest happens. In fact, the registers can play a crucial role in merely establishing that a natural-experiment-like event has ever occurred.

The third and final source of independent variation in explanatory variables that we discuss in this section is that of *calendar time itself*. The reason why register data can take advantage of calendar time for this purpose is that they are built upon two dimensions of time, *process time* and *calendar time*. While process time (e.g. the timing of labour market decisions within a life history, the duration of time spent in a particular state, etc) is obviously endogenous to each individual (and thus affected by unobserved heterogeneity), calendar time may in many cases be considered exogenous and hence play the role as an instrumental variable. The presence of multiple cohorts ensures that process time and calendar time are not perfectly correlated. As an illustrative example, consider the well-known issue of identifying the true pattern of duration dependence within a hazard rate framework, which has preoccupied the unemployment duration literature. The problem is that as spell durations of a cohort of unemployed persons proceed, a selection process takes place (the best job seekers leave unemployment first) that may result in declining escape rates for the cohort as a whole, even though individual escape rates remain constant over spell duration or even rise. It has proved surprisingly difficult to identify the true underlying individual *duration dependence* based on samples of spell durations, without relying on unjustified assumptions, such as *mixed proportional hazard rates* (MPH) or distributional assumptions about unobserved heterogeneity and/or the shape of structural duration dependence. Since these assumptions typically have no behavioural justification, the resulting estimates of spell duration effects in, e.g., the escape rate from unemployment, are questionable. With access to register data for multiple cohorts of unemployed this problem can be solved (Brinch, 2000; Røed and Zhang, 2002). The key to the solution is the additional dimension of time ensured by multiple cohorts. 'Lagged' calendar time can then play the important role as instrument (together with other lagged time-varying explanatory variables). The only exclusion restriction required is that labour market conditions experienced *earlier* in a given unemployment spell do not have a direct causal effect on the present hazard rate, given the present state of the labour market.

The same basic idea can be used for evaluation of treatment effects; see Abbring and van den Berg (2003) for a theoretical discussion of this approach, Gaure *et al.*

(2005) for a Monte Carlo based evaluation, and Røed and Raaum (2006) for a recent application. If one looks at the entry into labour market programmes, there is a calendar-time variation that is clearly related to administrative procedures or policy changes, and hence is not fully driven by the composition of the unemployed. A person who has not (yet) been enrolled into a programme despite having been unemployed through a period of particularly active enrolment, will on average have unobserved characteristics that are unfavourable to enrolment. Moreover, within a competing risks framework (with transitions to labour market programmes and/or to employment) it is possible to identify the correlation structure in the unobserved factors that affect the different transitions.

4. Disentanglement of micro- and macro phenomena

Since the micro units covered by (complete) administrative data typically add up to the macro numbers, register data provide the often 'missing link' between the micro- and the macro level of the economy. Hence, access to registers also entails freedom with respect to the *level of aggregation*; from no aggregation at all (i.e. analysing the decisions of the individuals), via aggregation at the levels of, e.g., households, firms, birth-cohorts or municipalities, and up to a complete economy-wide aggregation. This makes it possible to *decompose macroeconomic- or aggregate patterns into their appropriate micro phenomena*. For example, we can decompose changes in some economy-wide or local aggregate into changes that result from changes in *the composition of subjects who currently constitute the aggregate* on the one hand, and changes in the economic environment faced by these subjects as a group on the other.

To make our ideas more transparent, assume that we wish to measure changes in the tightness of the national labour market (from the job seekers' point of view), i.e., the extent to which job prospects are improving or deteriorating for a given (representative) job seeker. The first thing we would look at is probably the rate of unemployment, or (after second thoughts, perhaps) the rate at which job seekers obtain jobs, if such a statistic is readily available. Time series of these statistics can, however, yield a misleading picture of how job prospects change. The rate of unemployment may give a wrong impression for two reasons. The first is that the number of unemployed people continues to rise (fall) as long as inflow exceeds (falls short of) outflow, even if the individuals' employment prospects has improved (deteriorated) due to a higher (lower) outflow and lower (higher) inflow into unemployment. Consider a turning point in the business cycle defined as the point in time when the development of job prospects of job seekers change from deterioration to improvement (or vice versa). The corresponding turning point in the stock of unemployed individuals (the unemployment rate) will then typically be delayed as it takes some time before the flows into and out of unemployment level out. The second problem with the unemployment rate as a labour market tightness indicator is that the exact time it takes to level out these flows depends on the composition of the unemployment pool (in terms of e.g. average 'employability'), and this composition may vary systematically over the business cycle.

Even the observed outflow rate is affected by the composition of the unemployment pool; hence its development over time does not always mirror changes in labour market tightness. What we really want to look at is how job prospects develop over time, controlling for the composition of job seekers. By contrast, what the aggregate outflow rate can tell us is how average job prospects develop for a constantly changing group of people. With access to register data, these differences can be sorted out. Gaure and Røed (2003) have estimated a grouped hazard rate model for transitions from unemployment to employment in Norway during the 1989-2002 period without – apart from some proportionality assumptions - parametric restrictions on either spell duration, calendar time, observed- or unobserved heterogeneity. In practice, this is accomplished by means of several hundred dummy variables, e.g., a separate dummy variable for each calendar month that is represented in the dataset. The parameters associated with these calendar months measure the influences of time itself. And after appropriate seasonal- and trading-day adjustments, these parameters together constitute a business cycle (or labour market tightness) indicator. The freedom with respect to aggregation level implies of course that similar decomposition exercises can be made within regions, particular industries or with respect to particular demographic groups. Carlsen *et al.* (2006) estimate regional hazard-based business cycle indicators for 90 travel-to-work areas in Norway, and use them to estimate the impact of labour market tightness in dynamic panel models explaining local wage formation and regional migration flows during the 1990's. In both cases, the hazard-based tightness indicator outperforms the rate of unemployment, and basically renders the latter variable superfluous in both the wage and in the migration model.

One can also use 'creative aggregation' in order to identify 'contextual variables', e.g., in the form of attributes of institutional units that are represented in the dataset, yet cannot be directly observed. For example, having an economic model of individual wage determination, one may find that there are *firm specific wage components* that are not explained by the composition of the workforce within the firms. These 'residual' effects convey information about firms' wage policies, even though the firms may not be represented at all in the dataset (apart from the firm identifier observed for each individual worker). In a similar fashion, one may extract information about firms' technology (through the educational composition of the workforce), their retirement programmes (through observed retirement behaviour in the firm), their working environment (through the observed sick-leave behaviour) and so forth.

5. Work-place events and consequences for the employees

Since firms are represented in the data through their employees (by means of their firm identifier), the data implicitly keep track of firm (or work-place) events. In particular, it is possible to identify 'shocks' that lead to large changes in the number or composition of employees, e.g., in the form of reorganisation, downsizing, or closure. There has been a growing interest in the short- and long-term consequences of being hit by such shocks; for future employment prospects, earnings, social security dependency, and health. On the basis of administrative registers the employees can be traced both before and (for a long time) after such events happened. Various approaches have been used to account for the obvious selection problems involved in analysing the causal effects of being hit by, e.g., a downsizing or a closure (both employees being laid off in a downsizing and employees remaining in a firm at the time of a closure may be strongly selected). Huttunen *et al*. (2005) take the stock of employment in the firm some time *before* the event in question happened as a starting point for their analysis, and distinguish between 'early leavers' (those that quit the firm in the period prior to a closure) and 'exit-layoffs' (those who lost their job when a plant closed down). They follow these individuals (and control groups) for up to eight years after the closure, and assess the impact of the event on their probability of being employed and on their earnings (if they are still employed). Rege *et al.* (2005) circumvent the selection problem by looking at the 'effects' of working in a firm that (with the wisdom of hindsight) is going to downsize substantially or close down during the coming years (i.e., they do not look at the effect of actually being laid off). They also utilise the opportunities represented in administrative register data by looking at impacts on the probability of becoming disabled as well as on mortality. Røed and Fevang (2006) attempt to model the selection process directly within the framework of a competing risks hazard rate model for a group of Norwegian nurses who, during the course of the eight-year data period experience a number of organisational changes. They evaluate the impact of these changes on hazard rates to new jobs, to sickness absence, to more lasting social-security dependence (rehabilitation, disability), as well as to direct withdrawals from the labour force (without social security benefits).

An important question arising in the evaluation of firm-specific events of the type discussed here is how to define a firm. Administrative registers may not provide much choice here; firm identifiers may refer to big corporations or single plants. In the former case, it may be difficult to identify the events of interest from observed data (or more creative techniques must be used, such as looking for larger groups of individuals being registered at local employment offices, with a common firm-identifier). The researcher may also be confused by 'spurious' events, created by what we may call 'paper-reorganisations'. For example when two firms are merged it may appear from the data that two closures have occurred, even though little (or nothing) has happened to the employees. Consistency checks (with, e.g., the unemployment register) are therefore always called for.

6. Intergenerational issues

The extent to which socio-economic outcomes depend on family background is an issue of great interest to social scientists as well as policy makers. ‘Equality of opportunity’ is a principle within the core of welfare policies, hence knowledge about what kinds of institutions that provide the most equal opportunities for new-born citizens and the best insurance against being born into less privileged environments is of great value. Administrative register-data contain family identifiers that facilitate analysis of the intergenerational transmission of economic wealth, of socio-economic position and of health. The data are also sufficiently rich to go beyond the standard measures of parent-offspring correlations in, e.g., income and education. Jäntti *et al.* (2006) provide a recent example, in which intergenerational mobility is evaluated by means of a number of statistics, including income-position transition-matrices. This paper also tries to combine administrative register data from the Nordic countries with survey data from the United States and the United Kingdom to provide an appropriate foundation for international comparisons. The basic idea is that Nordic register data are so ‘rich’ that they can be used to mimic the survey data generated in other countries (i.e., one can draw cohorts that match surveys in other countries, and construct income measures that are comparable to those used in these studies).

Administrative registers also provide opportunities for identifying the ‘genetic part’ of the intergenerational transmission mechanisms. The basis for such studies is intergenerational panel data containing parents and siblings, where the siblings can be divided into twins, non-twins, and non-biological siblings (adopted children). For example, one may look at the differences in intergenerational correlations between parents and biological children and parents and (early) adopted children. Combining administrative data with specific surveys can also be highly informative. Björklund et al (2005) merge the Swedish Twin Registry with register data to, first, obtain information on earnings, and, second, establish a wide range of siblings. Hence, the authors utilize resemblance in earnings between mono- and dizygotic twins (of which some are ‘reared apart’) as well as between full non-twin siblings, half siblings and adopted children.

7. The challenges of non-parametric inference

It is a well-known fact that micro-econometric identification of causal effects often rests on functional form restrictions, some of them apparently innocent simplifications, others of vital importance for identification of the effects of interest. Even when purely data-based identification is proved to exist, strong functional form restrictions are typically needed; hence it is sometimes hard to tell whether the results are driven by the data (as assumed) or by the restrictions. Register data provide the opportunity for 'non-parametric' approaches, in the sense that functional form assumption can be replaced by a huge number of dummy-variables. We realise, of course, that no empirical approach is completely non-parametric. Even when the degrees of freedom are virtually unlimited, parametric restrictions may be imposed directly on the basis of economic theory, or (more ad hoc) in order to ensure that the results are economically 'interpretable'. And since the degrees of freedom never are unlimited, one also has to trade off flexibility/richness in the model against precision in parameter estimates. Hence, the issue of non-parametric modelling is really a matter of degree, more than of kind. Our own experience suggests that when the register data are allowed to speak for themselves, 'everything affects everything' and the model typically becomes complex and un-interpretable. This reflects, of course, the enormous degree of individual heterogeneity. It is probably no exaggeration that, unless an explanatory variable is deliberately created by a random generator, the issue of whether it will be attributed a coefficient significantly different from zero or not in a statistical analysis of human behaviour, is *only a matter of sample size*. With sufficiently large dataset all coefficients are significant; if they are causally irrelevant they will inevitably pick up some correlation to unobserved (relevant) factors. Consequently, the linking back to structural parameters of interest quickly becomes an insurmountable task. This illustrates the difficulties associated with making a direct confrontation between economic theory and data. Economic theories are partial and simple, and not even designed to represent the 'whole truth' (in the sense that a model explains data fully). The real world is extremely complicated, and, if unaccounted for, the complexities will push their way into the parameters of simple models in ways that render their interpretation intractable. This problem not only makes it difficult to test and quantify economic theories, it also makes the usage of micro-simulation models for policy-evaluation purposes questionable. Large datasets may to some extent solve this problem. In relation to partial economic theories, their most important role is to take care of the enormous heterogeneity that exists in individuals' behaviour and resources, and make sure that this heterogeneity does not disrupt the researcher's attempt to identify and assess a well-defined structural parameter or causal mechanism.

Non-parametric modelling raises huge computational problems. For example, in duration analysis one would typically like to represent process time by one dummy for each possible duration time (e.g., for each month or each week), and one dummy for each possible calendar time. Similarly, one would like to represent variables such as age and educational attainment by a large number of dummy variables, rather than through 'arbitrarily' selected linear, log-linear or quadratic functions. More critically, one would perhaps also like to represent unobserved heterogeneity through a number of unobserved dummy variables, along the lines suggested by Heckman and Singer (1984), and estimate their size and probabilities together with the other parameters of interest. In addition to all this, one would perhaps like to ease proportionality assumptions by allowing interaction terms between, e.g., spell duration, calendar time and explanatory variables. The result is probably an extremely complicated likelihood function. This raises the issue of computational costs. Computation time tends grow in the square of the number of parameters to be estimated, implying that going from, say 20 parameters within a restrictive functional form setting to 1,000 parameters in a non-parametric dummy setting multiplies computation time with a factor of 2,500. And those who have worked with the modelling of non-parametric unobserved heterogeneity know that it is a large computational task to locate the global maximum of the likelihood function even in the former case, and even with relatively few observations. Apparently simple models often require days of computation time. Fortunately, this problem can be substantially alleviated by what we call *implicit dummy variables* (see Gaure and Røed, 2003, for details). When a standard estimation procedure is called upon to recover parameters associated with, say, 100 mutually exclusive dummy variables, it computes the inner product of all the dummies and parameters a large number of times. Every time it does 99 multiplications with zero, and 1 multiplication with 1, and it then does 99 additions with 0. Doing that kind of calculations perhaps billions of times is of course a tremendous waste of resources. The implicit dummy approach amounts to replacing all the 100 mutually exclusive dummies with one single variable, which takes 100 different values, and then to estimate separate parameters for each of these alternative values without doing all the superfluous zero-calculations. Now, many researchers do not pay much attention to computation time, since modern computers do the job so quickly in any case. Whether estimation takes a few seconds or 10 minutes may not be a big issue. However, working on register data, the issue may be a few hours versus several months. Much more important, efficient computational tools allow the researcher to formulate and estimate models that would be completely infeasible (by several orders of magnitude) with standard methods. And at this point, our experience suggests that *the supply creates its own demand*, in the sense that new computational tools stimulate the development of ideas for future research.

8. Benefits versus costs

In the previous sections, we have explored some of the opportunities offered by national administrative data. Compared to surveys, a number of advantages can be identified. First, administrative registers cover all permanent residents; hence researchers do not have to worry about representativity or attrition bias arising from selective response behaviour. This is particularly important in longitudinal surveys, where the apparent trend of increased unwillingness to answer questions raised by researchers often makes survey data virtually useless.

Second, there are no reporting or recollection errors. Third, there is a virtually continuous panel dimension represented in the data. Fourth, we can identify individuals who share a common environment and those who are closely related to each other, for social or biological reasons. Fifth, the large number of observations makes it possible to ease functional form restrictions and still obtain sharp estimates. Sixth, the researcher is not limited by data in studying important, but small, dimensions or groups. And finally, the aggregation level can be determined freely by the researcher.

There are of course also disadvantages associated with administrative data. First, since the register information is collected for administrative purposes, the definitions of states and variables are frequently different from what researchers prefer. Second, there may be administratively generated measurement errors or inaccuracies, particularly with respect to the timing of events and with respect to variables that are not important from the administrative point of view. Third, the data-generating process may not have been constant over time; collection procedures as well as variable-definitions may have changed, and the exact nature of such changes is not always recoverable. Fourth, due to their sensitive nature, access may be restricted, hampering free exchange of data and thereby possibly also the scope for verifiability of research results. In practice, requests by journals to make data publicly available often conflicts with the confidentiality requirements set by data providers or data inspection authorities.

Research communities considering embarking upon register data projects also face costs, and these costs must of course be weighted against potential benefits. Now, from a social point of view, there are no costs at all associated with data collection, since these costs are already paid for by other users (and hence sunk). But there are real costs associated with acquiring the required knowledge about the data. Substantial resources are also needed to develop the required tools for handling them and adapt them for research purposes. However, these costs are, too a large extent, investment costs. Once the administrative registers are linked and adapted in a 'researchable' fashion, the marginal cost associated with using them for a new project is virtually zero. Hence, there are huge economics of scale involved. The transformed registers should therefore ideally be considered public goods where access should not be regulated by the willingness to pay (a price above marginal costs).

In practice, register data providers are likely to charge the research community for any additional expenses that have to be incurred when the data are made available for research purposes, and they may also take the opportunity to cover their already sunk costs. Our experience suggests that that these expenses are of minor importance, and that the price paid for the 'raw' registers is small. In many cases, the register providers consider it to be in their own interest that the registers are used by researchers, because they then get feedback on the register quality and often also get a more direct communication line to research activities that are relevant for their own administrative purpose. Data collectors are commonly public bodies closely associated with policymaking administrations. When research provides a better understanding of the (heterogeneity of) individual behaviour that generates the records, the foundation for policymaking is improved.

14. Concluding remarks

Microeconometrics is the art of extracting chains of causality from observed behaviour. One fundamental problem that inevitably has to be overcome in one way or another, is that outcomes of interest not only depend on the variable for which a causal parameter is to be identified, but also on unobserved individual characteristics that might be correlated to the variable in question. In this paper, we have argued that the emergence of large scale register data that keep track of all citizens' performance in the labour market and related activities over time, contain the seeds for substantial scientific progress within the field of labour market econometrics. But despite large efforts the past 10-15 years, we also consider register-based microeconometrics to be in its infancy. Efficient utilisation of register data requires development of new methodological approaches, and new statistical- and computational techniques. The investment in administrative data for research purposes is likely to improve our understanding of choice behaviour in the labour market.

The development of a new large-scale labour market event history database can be seen as a public good *investment project*. The investment dimension is not primarily related to the procurement costs (which tend to be low), but the costs associated with acquirement and maintenance of knowledge about the underlying data-generating processes, the adaptation (and documentation) of the data into some 'researchable' format, the construction of storage- and access facilities, the development of statistical tools and optimisation programmes (standard software is often of little help when complicated likelihood functions are to be maximised with the aid of millions of observations). A lot of problems are bound to emerge during this investment period, such as inconsistencies between different registers, errors in variables, breaks, computational limitations etc. The complicated process of transforming the well of information embedded in the data into familiar 'states' and 'variables' is likely to add new errors. Once the investment is undertaken, though, huge research opportunities emerge, in which the cost of providing data for a new research topic or a new approach may often be virtually zero. Expanding the time (period) dimension of the data incurs, in sharp contrast to longitudinal surveys, basically no costs.

Particularly promising research opportunities will arise as registers from more countries become available and as the design of the various registers converge towards some common standards. It will then be possible to combine the huge information content embedded in the longitudinal and cross-sectional variation in individual outcomes within each country with institutional differences between different countries. For the time being, however, there is a lot do be done in order to exploit the information already embedded in country-specific data.

Data from administrative registers do of course not offer a solution to *all* the intriguing problems of interpreting correlations and identifying causal relationships in empirical labour economics. A lot of interesting information can never be recovered from registers, such as individual attitudes and reactions to hypothetical situations. Hence, interview-data still have important roles to play. But registers can relieve interviewers from asking a number of questions for which registers provide (superior) answers, and hence make it possible to focus the interviews entirely on the type of questions that really need to be answered directly by the respondents. Hence, access to register data may improve the scope for cost-effective provision of good interview data as well.

Even with the best possible data, there is a lot that can never be observed, such as counterfactual behaviour and intrinsically unobservable individual characteristics. No matter the data, there will always be plenty of scope for sophisticated methods to account for these fundamental problems. Good data do not solve all problems, but definitely help a lot. It is frequently argued that more effort should be devoted to the provision of better data, rather than developing techniques that contain the damage caused by bad data. This general argument is reinforced by the increased access to administrative data, simply because the potential of these data is enormous, raising the (marginal) return to data providing activities.

References

Abbring, J. H. and Van den Berg, G. J. (2003) 'The Nonparametric Identification of Treatment Effects in Duration Models'. *Econometrica*, Vol. 71, 1491-1517.

Björklund, A., Jäntti, M. and Solon, G. (2005), 'Influences of Nature and Nurture on Earnings Variation: Preliminary Results from a Study of Various Sibling Types in Sweden', in Bowles, Gintis and Osborne (eds) Unequal Chances: Family Background and Economic Success, Russel Sage, New York.

Bound, J. and Solon, G. (1999) 'Double Trouble: On the Value of Twins-based Estimation of the Return to Schooling', *Economics of Education Review,* 18 (1999) 169-182.

Brinch, C. (2000). 'Identification of structural duration dependence and unobserved heterogeneity with time-varying covariates'. Memorandum No. 20/2000, Department of Economics, University of Oslo.

Carlsen, F., Johansen, K. and Røed, K. (2006) 'Wage Formation, Migration, and Local Labour Market Tightness'. *Oxford Bulletin of Economics & Statistics,* Vol. 68 (2006), No. 4, 423-444.

Gaure, S. and Røed, K. (2003) 'How tight is the labour market? A micro-based macro indicator'. Memorandum No. 9/2003, Department of Economics, University of Oslo

Gaure, S., Røed, K, and Zhang, T. (2005) 'Time and causality: A monte carlo assessment of the timing-of-events approach. Memorandum No. 10/2005, Department of Economics, University of Oslo.

Heckman, J. and Singer, B. (1984). 'A method for minimizing the impact of distributional assumptions in econometric models for duration data'. *Econometrica*, Vol. 52, 271-320.

Huttunen, K., Møen, J., and Salvanes, K. G. (2005) How Destructive is Creative Destruction? The Costs of Worker Displacement. NHH Working paper.

Jäntti, M., Bratsberg, B., Røed, K., Raaum, O., Naylor, R., Österbacka, E., Björklund, A. and Eriksson, T. (2006) American exceptionalism in a new light: a comparison of intergenerational earnings mobility in the Nordic countries, the United Kingdom and the United States. Memorandum No. 34/2005, Department of Economics, University of Oslo.

Rege, M., Telle, K. And Votruba, M. (2005) 'The Effect of Plant Downsizing on Disability Pension Utilization'. Discussion Paper 435, Statistics Norway.

Røed, K. and Fevang, E. (2006) 'Organisational Change, Absenteeism and Welfare Dependency' *Journal of Human Resources*, forthcoming.

Røed, K. and Raaum, O. (2003) Administrative Registers – Unexplored Reservoirs of Scientific Knowledge? *Economic Journal*, Vol. 113. (2003), F258-F281.

Røed, K. and Raaum, O. (2006) 'Do labour market programmes speed up the return to work? *Oxford Bulletin of Economics & Statistics*, forthcoming.

Røed, K. and Zhang, T. (2002) 'The duration and outcome of unemployment spells - the role of economic incentives'. Memorandum No. 6/2002, Department of Economics, University of Oslo

Røed, K. and Zhang, T. (2003) 'Does unemployment compensation affect unemployment Duration? Economic Journal, Vol. 113 (2003).