United States™ Census Bureau

NORC at the UNIVERSITY of CHICAGO

Abstracts for the Conference on

Utilizing Administrative Data:
Technical, Statistical & Research Issues

October 27th and 28th, 2011
Washington D.C.

Contents

Day I, Session I: Experiences with Linked Data

Day I, Session II: Theory and Methods

Day I, Session III: Building Integrated Data Sets

Day II, Session I: Register Based Population Censes

Day II, Session III: Case Studies

DAY I, SESSION I: Experiences with Linked Data

Developing Integrated Administrative Data for Policy Analysis and Research

Maria Cancian, Institute for Research on Poverty at the University of Wisconsin-Madison

Abstract

A review of the policy evaluation and research supported by the longitudinal merged administrative data that comprise the Institute for Research on Poverty Data Core. The presentation includes a summary of the structure and contents of the data system, illustrative examples of research projects employing the data – from applied program analysis to more academic research—and a discussion of lessons learned for similar data construction efforts.

Illustrating the Potential Health Policy Uses of Linked Administrative Data and Survey Data: The Case of the Medicaid Undercount Study (aka SNACC)

Michael Davern, NORC at the University of Chicago; Jacob Klerman, Abt Associates;Kathleen Thiede Call, University of Minnesota

Abstract

Linked survey and administrative data have often been used to assess the quality of survey data self-reported public program enrollment information and income amounts. Another potentially strong contribution these kinked data files could make is to help evaluate how well state and local run public programs are serving their clients. Using a data file with the Current Population Survey's (CPS) cases linked to the Medicaid Statistical Information System (MSIS) we examine whether we detect state differences in reporting patterns of public program enrollment (Calendar years 2000-2003). We generate two logistic regression model predicting whether the CPS case was appropriately coded as having Medicaid in the CPS having Medicaid given that the MSIS shows enrollment. The second model examines whether CPS respondents linked to MSIS showing enrollment in Medicaid answer that they are uninsured in the CPS interview. Both models control for important variables including utilization of services, and state of residence. We feel the state variation in the first model estimating whether a case was correctly classified as having Medicaid could be caused by many things such as stigma and confusion over the exact plan name someone in enrolled in (e.g. ,SCHIP versus Medicaid). However, we feel an error to report any coverage at all is more serious as it shows people do not know they are enrolled in coverage that the State is providing (and therefore perhaps act more like they are uninsured and have health outcomes more like the uninsured). We demonstrate wide variability among the states in the probability of being coded in the CPS as being uninsured given that the person shows Medicaid enrollment. In this case "survey response error" can also be seen as a programmatic evaluation tool.

DAY I, SESSION II: Theory and Methods

Errors in survey reporting and imputation and their effects on estimates of Food Stamp program participation

Bruce Meyer, University of Chicago; Robert Goerge, Chapin Hall

Abstract

Benefit receipt in major household surveys is often underreported. This misreporting leads to biased estimates of the economic circumstances of disadvantaged populations, program takeup, and the distributional effects of government programs, and other program effects. We use administrative data on Food Stamp Program (FSP) participation in two states matched to American Community Survey (ACS) and Current Population Survey (CPS) household data. We show that nearly thirty-five percent of true recipient households do not report receipt in the ACS and fifty percent do not report receipt in the CPS. Misreporting, both false negatives and false positives, varies with individual characteristics, leading to complicated biases in FSP analyses. We then directly examine the determinants of program receipt using our combined administrative and survey data. The combined data allow us to examine accurate participation using individual characteristics missing in administrative data. Our results differ from conventional estimates using only survey data, as such estimates understate participation by single parents, non-whites, very low income households, and other groups. To evaluate the use of Census Bureau imputed ACS and CPS data, we also examine whether our estimates using survey data alone are closer to those using the accurate combined data when imputed survey observations are excluded. Interestingly, excluding the imputed observations leads to worse ACS estimates, but has little effect on the CPS estimates.

Non-response when linking survey data with administrative records

Julie Korbmacher; SHARE, MEA, University of Mannheim and Mathis Schroeder, German Institute for Economic Research

Abstract

The Survey of Health, Ageing and Retirement in Europe (SHARE) has collected retrospective life history data in its third wave (2008/2009) in thirteen European countries. In addition, about 900 cases have been linked with the respondents' consent to records of the German Pension Fund (Deutsche Rentenversicherung, DRV) in Germany. The main reasons for record linkage are the availability of more exact and reliable measurements of income data and job spell information in the administrative records on the one hand and the provision of context information from the survey data on the other. As the direct combination of these two datasets depends on the consent of the respondents, there may be a non-random selection into the sample of matched individuals. The consent process and the determinants of the consent decision are the main focus of this paper, because it is a common step also for other studies. Using the advantage of a full set of variables for all consenting and non-consenting respondents, we identify different areas which may influence the respondents' willingness to consent. As in other studies, we consider control variables on the individual and on the household level, but then augment our study by two additional areas of interest: the interview situation – where we consider measures such as comprehension of questions, length of answers and rates of missing answers – and interviewer variables – gender, age, and education. Using a multilevel approach, we then measure the amount of variation due to the interviewers. Our results suggest that interviewers play an important role in the consent decision and more so than variables on the individual or household level. As the interviewer is the integral part in data collection efforts and in addition, the direct link to gain a respondent's consent, this finding stresses the importance of interviewer training.

The role of statistics in comparative effectiveness research - Designing Observational Studies for Objective Causal Inference Using Propensity Scores

Lilly Q. Yue, Ph.D., Allen H. Heller, M.D., Donald B. Rubin, Ph.D., S. Stanley Young, Ph.D. and Elizabeth R. Zell, Ph.D.,; U.S. Food and Drug Administration, Bayer HealthCare Pharmaceuticals, Harvard University, National Institute of Statistical Sciences and Centers for Disease Control and Prevention

Abstract

Comparative effectiveness research is gaining increasing attention in US health care reform. The basic idea of this approach is to take advantage of existing data sets, which are primarily observational in nature, to infer the relative effectiveness of medical interventions. These data sets, however, typically have severe limitations for drawing reliable causal inferences about medical interventions. For example, administrative hospital insurance claims data may include billable charges for patients only in hospital and exclude patients who are not in hospital -- how can such data be used to compare outcomes for patients taking one drug versus another, even if the data for in-hospital patients are perfectly accurate? For another example, baseline characteristics and medical history are usually incomplete. Is it reasonable to assume that if a medical condition is not recorded, it is absent? And of course, observational data, even when accurately recorded, are problematic for drawing causal inferences and any study design requires careful consideration of underlying assumptions, which are often not stated in published literature based on such data sets. What are the consequences of the massive use of such data on future health care decisions?

DAY I, SESSION III: Building Integrated Data Sets

Exploiting administrative data to explore job churn in the Irish labour market

John Dunne, Statistical Methods and Quality Administrative Data Centre, Central Statistics Office, Ireland

Abstract

The paper will cover experiences from the Job Churn Explorer project at CSO. The project adapts and develops the underlying methodology outlined to date to the situation in Ireland to provide a detailed insight into the dynamics of job churn and its components as Ireland entered the current recessionary period. The analysis datasets used are derived from linking the following three sources – Business register – Employer tax returns – Social Protection records. The comprehensiveness of the resulting analysis dataset containing attributes on both workers and enterprises provides for significant new opportunities to inform policy and decision making with respect to the labor market. The presentation will also graphically present some of the analysis at both a worker and enterprise level to demonstrate the potential of this new information. The paper will discuss some of the challenges encountered and (partial) solutions implemented – with using admin data – adapting the underlying methodology – packaging large volumes of statistics for user consumption–reconciliation with system of Business Demography statistics.

Linking survey data with administrative employment data: The case of the IAB-ALWA survey

Manfred Antoni, German Institute for Employment Research (IAB)

Abstract

Numerous research questions in the social sciences can only be tackled by methods of inference that rely on rich data sets. Taken by themselves, either survey or administrative data have their respective advantages. The union of these data sources thus provides additional potential for methodological and substantive research. This study describes and evaluates the linkage of the German survey "Work and Learning in a Changing World" (Arbeiten und Lernen im Wandel, ALWA) with administrative data from the Institute for Employment Research (IAB), the research institute of the Federal Employment Agency. The ALWA data include information from more than 10,000 retrospective interviews with people aged 18 through 50. Longitudinal information was gathered on residential, educational, employment and partnership histories as well as on children and times of parental leave. This is complemented by a rich set of cross-sectional variables. Consent for data linkage was given by 92% of the respondents. Administrative data from the IAB contain daily information on employment and unemployment histories beginning with the year 1975. Information on transfer payments and wages are measured with high accuracy as they are related to social security contributions. Detailed longitudinal firm information is given for every employment spell included in the administrative data. The goal of the study is threefold. First, the potential for research is demonstrated by giving an account of the information available in the combined data set and by analyzing whether there is selectivity in the linked data compared to the overall survey population. Second, implications for survey design and field administration are shown by examining how the interviewer staff may be composed or how field management may be optimized to assert high and stable consent rates. Finally, insights for researchers linking survey data to register data are provided by showing what can be gained from different linkage techniques in terms of numbers of observations as well as whether and how they affect the selectivity of the linked sample.

School-based health centers: Cost-benefit analysis and impact on health care disparities

Jeff Guo, Terrance Wade & Wei Pan, University of Cincinnati; Kathryn Keller, Health Foundation of Greater Cincinnati

Abstract

We evaluated the impact of school-based health centers—which provide essential health care for students by aiming to eliminate many access barriers—on health care access disparities and conducted a cost–benefit analysis. We employed a longitudinal quasi-experimental repeated-measures design. Primary data sources included the Ohio Medicaid claims, enrollment file with race/ethnicity, and survey reports from parents and administrative staff. We used hierarchical linear modeling to control unbalanced data because of student attrition. We assessed quarterly total Medicaid reimbursement costs for 5056 students in the SBHC and non-SBHC groups from 1997 to 2003. We calculated net social benefit to compare the cost of the SBHC programs with the value that SBHCs might save or create. With SBHCs, the gap of lower health care cost for African Americans was closed. The net social benefits of the SBHC program in 4 school districts were estimated as $1 352 087 over 3 years. We estimated that the SBHCs could have saved Medicaid about $35 per student per year. Conclusions. SBHCs are cost beneficial to both the Medicaid system and society, and may close health care disparity gaps. We evaluated the impact of school-based health centers—which provide essential health care for students by aiming to eliminate many access barriers—on health care access disparities and conducted a cost–benefit analysis.

DAY I, SESSION IV: Research Access Roundtable

Robert Goerge (Moderator), Chapin Hall
John Abowd, Cornell University
Matthew Shapiro, University of Michigan

This Roundtable will present a discussion of administrative data utilization issues by three expert academic policy researchers. Topics will include the potential of linked data for academic research, research priorities for the federal statistical system and the collaboration needed to provide quality data linkages. Experts will also discuss challenges and successes in the access and availability of the data needed to conduct their research.

DAY II, SESSION I: Register Based Population Censes

Dutch Virtual Census

Lada Mulalic, Statistics Netherlands

Abstract

The Dutch Virtual Census of 2011 is planned to be conducted with a very similar methodology to the one used for the last virtual census of 2001. The Netherlands is using registers and other administrative sources, together with information from sample surveys, to provide census statistics. The acquired experience in dealing with data of various administrative registers for statistical use enabled Statistics Netherlands to develop a Social Statistical Database (SSD), which contains coherent and detailed demographic and socio-economic statistical information on persons and households. The one of the initial driving forces for this development was the virtual census of 2001. The importance of SSD for the virtual census of 2011 strongly increased comparing to the previous one, due to better quality and availability of more register data. A brief description of this database structure will follow. Additionally, strengths and opportunities, but also weaknesses and threats of the register based production of statistics and censuses will be discussed.

Use of administrative sources for census and demographic and social statistics

Lars Thygesen, Statistics Denmark

Abstract

Demographic and social statistics in Denmark, including censuses, have been mostly based on administrative registers since 1981. Surveys based on interviews or questionnaires are important supplementary sources in fields where suitable register data cannot be obtained. The statistics based on surveys and registers are closely connected in one coherent system, which is briefly described in this paper. The philosophy behind the system, as well as its merits, problems and challenges, are discussed. The system was basically developed by Statistics Denmark over approximately 15 years from the late 1960s as a consequence of a very clear management strategy. This paradigm combined increased reliance on administrative sources with a philosophy of a statistical system that was more flexible towards new and unforeseen statistical needs. The preconditions of this strategy were supplied by modernization of the Danish public administration and its registration practice, and by a good legal basis: The Act on Statistics Denmark (1966) foresaw that administrative registers should be used for statistical purposes and gave Statistics Denmark right of access to any data kept by public agencies. Later on, the data protection legislation contained provisions that recognized the needs of statistical production.

Combining registers into a fully register-based census - some methodological issues

Ingegerd Jansson, Dan Hedlin, Anders Holmberg, Statistics Sweden

Abstract

Statistics Sweden faces the challenge of conducting Sweden's first fully register-based census. Several registers, for example the existing Population register, the Real property register and the new Register of dwellings will be matched. There are number of methodological issues involved, such as statistical matching, disclosure control and evaluation of model assumptions. Missing data pose a particular problem. There will be individuals with no recorded dwelling in the population register, as well as dwellings with no residents according to the register of dwellings. Methods for matching individuals and dwellings will be discussed.

DAY II, SESSION II:

Working Toward Administrative Record Usage

Andy Teague, Office for National Statistics, United Kingdom
Martin Ralphs, Office for National Statistics, United Kingdom
Dave Dolson, Statistics Canada
Frank Vitrano, U.S. Census Bureau

This Panel features representatives from statistical census agencies from the United Kingdom, Canada and the United States. After representatives from each country present an overview of future census options and administrative record use in their country, there will be a topics-based discussion of the successes and challenges each country has had.

DAY II, SESSION III: Case Studies

The Scottish Longitudinal Study

Chris Dibben, University of St. Andrews

Abstract

The Scottish Longitudinal Study (SLS) is a large-scale linkage study which has been created by using data available from current Scottish administrative and statistical sources. These include Census data, Vital Events data (births, deaths, marriages), National Health Service Central Register (NHSCR) data (migration in or out of Scotland), Scottish Schools and NHS data (cancer registrations and hospital admissions). The SLS is a 5.3% representative sample of the Scottish population and began with data from the 1991 census. Approximately 274,000 SLS members have been identified from the 1991 census and information for these individuals has been linked from other datasets, including the 2001 and 2012 census, vital events and health information.

Crisis indicators: Triage tool for identifying homeless adults in crisis

Dan Flaming, Economic Roundtable

Abstract

The triage tool, or crisis indicator, identifies homeless individuals in hospitals and jails who have continuing crises in their lives that create very high public costs. This redesigned tool is four times more accurate than the earlier screening tool released in 2010. The tool is developed for use in jails, hospitals and clinics where homeless individuals with high levels of need and high public costs are most likely to be found. Discovery of the exceptionally high public costs for people in the 10th cost decile has led to interest in identifying these individuals and giving them high priority for access to permanent supportive housing. This group accounts for well over half of all public costs for homeless adults, and their costs decrease by 86 percent when they live in permanent supportive housing. The triage tool was developed based on two key propositions. The first proposition is that the greatest risk to homeless individuals is of continuing crises in their lives, particularly crises that cause encounters with hospitals and jails. The second proposition is that the most compelling basis for prioritizing access of homeless individuals to the scarce supply of permanently subsidized supportive housing is the public costs that will be avoided when they are housed. The triage tool is a system-based tool for identifying the one-tenth of homeless persons with the highest public costs, and the acute ongoing crises that create those high costs. This is the highest need segment of a much larger homeless population needing supportive housing.

Linking K-12 data to each other and to external data sources

Dorothyjean Cratty, U.S. National Center for Education Statistics

Abstract

This session is a demonstration and discussion of the types of policy-relevant research questions that can be addressed using existing state education data and a range of techniques that can be employed to turn administrative data files into research-ready datasets. Hundreds of North Carolina K-12 data files were linked with each other and external data sources to generate three powerful research models. Findings are discussed for statewide analyses of 3rd-12th grade college-readiness and dropout propensities as well as teacher matching and the role of income in class assignment.