SESSION 40

Data Integration II

chair Owen Abbott - Office for National Statistics, UK

Integrated Health Care Survey Designs: Data Quality Enhancements Achieved through Linkage of Surveys and Administrative Data

Steven B. Cohen, Agency for Healthcare Research and Quality, USA

A key feature of an integrated survey design is the direct linkage between sample members in the core survey with the larger host survey; administrative records; or follow-up surveys. The data quality and content of household specific health surveys may be substantially enhanced through the adoption of such integrated designs, which include data linkages between surveys and administrative data, and the conduct of follow back surveys to medical providers and facilities that have provided care to household respondents. With respect to health care expenditures collected from household respondents for their reported health care events, available linked medical provider level data is a more accurate source of information. The availability of such supplemental data on use and expenditures allows for the conduct of methodological studies to evaluate the accuracy of household reported data and informs adjustment strategies to household data in the absence of provider specific data to reduce bias attributable to response error.

The data integration model also allows for the cost efficient specification of a sampling frame for the core survey by utilizing an existing frame with detailed socio-demographic information to facilitate oversampling efforts and allow for dual frame designs. These attractive design features stand in clear contrast to the alternatives of new frame construction and/or independent screening interviews that characterize unlinked survey design efforts. In addition to utilizing existing databases as a sampling frame to support the sample design of the core survey, this prior information from the host survey or administrative records informs nonresponse and poststratification adjustments, facilitates imputation and serves as a data supplement to correct for item nonresponse. The detailed information available on demographic/socio-economic characteristics of both respondents/ and nonrespondents from the host survey or administrative records enhance the capacity of the specification of more accurate adjustments to correct for survey nonresponse. In the absence of an integrated survey design, the adopted nonresponse adjustment strategy would be constrained to socio-demographic and economic information that were available at the geographic level.

The design's capacity for data augmentation for a fixed time period, and the potential for longitudinal analyses over time through survey linkages are other attractive features of an integrated design framework In terms of data quality, household reported medical conditions can be evaluated for accuracy relative to provider specific records on medical conditions for the same patient and specific health events. In this paper, the capacity of integrated survey designs to improve the quality of resultant data, to achieve reductions in bias attributable to survey nonresponse and to enhance analytical capacity is discussed. Several examples are drawn from the Medical Expenditure Panel Survey (MEPS), sponsored by the Agency for Healthcare Research and Quality, which is characterized by an integrated survey design.

scohen@ahrq.gov

DATA INTEGRATION ◆ BIAS REDUCTION ◆ MEPS

The Relationship between Error Rates and Parameter Estimation in the Probabilistic Record Linkage Context

Nicoletta Cibella, Marco Fortini, Tiziana Tuoto - Istat, Italy

DATA INTEGRATION • RECORD LINKAGE • ERROR MATCH RATES

Nowadays data integration procedures are becoming extremely important in official statistical institutes. In particular, record linkage procedures, aiming at matching records referring to the same entities, both within a dataset and from two or more different data sources, improves the quality of information collected and enables to detail analysis, for instance removing duplication combining various information. A large number of linkage techniques are available and commonly used; in the field of the official statistics, the quality of the implemented procedures is crucial, especially because not only the criterion to estimate the accuracy of the procedures but also the one for evaluating the error match rates need to be established. The aim of this paper is assessing the probabilistic record linkage process quality by means of alternatives methods to estimate the parameters of the probability model; in other words, we aim to evaluate how the procedure accuracy is related and dependent on the choices adopted in the parameters estimation phase. Starting from Yancey (2004), the basic idea of this work is to achieve better parameter estimates considering different subsets of all the pairs candidate to the linkage, comparing different techniques for reducing the pairs search space including the conventional blocking criteria, the samples of suitable subsets and the recent mapping algorithms, that allow to map objects preserving the similarities and dissimilarities between them (Faloutsos C et al, 1995). Also for the parameters estimation phase, alternative methods have been analysed, performing both the EM algorithm in the classic probabilistic model, and the Bayesian approach; the related estimates are calculated on the set of all linkable pairs and on different subsets of the whole set, in order to evaluate the improvement due to the expected increase of the distinguishing power of some variables.

The results of the comparisons are evaluated and synthesized in terms of matching proportion, false match and false not-match rates. Generally, it's not easy to find automatic procedures to estimate these two types of errors so as to evaluate the record linkage procedures quality. So, finally also those errors are calculated via different methods, firstly starting from the known true matches status, but also through functions of the parameters themselves (Belin and Rubin, 1995; Torelli and Paggiaro, 1999).

The present study tests the alternative choices above described, exploiting the great amount of real data referred to the XIV Population Census and the relative Coverage Survey. Actually, the Census coverage rate is usually estimated on the basis of a post enumeration survey which needs to be linked with the Census data itself in order to estimate the unknown amount of the population, via dual system estimation model. This model assumes that the linkage procedure is error-free, so as a complex and accurate mixed (deterministic, probabilistic and clerical) procedure was performed; for this reason, it's possible to consider as known true match status of each unit the one performed in the census occasion so as to evaluate the quality of alternatives procedures in a simulation context.

References

Belin T.R., Rubin B. (1995), "A method for calibrating false-match rates in record linkage", Journal of the American Statistical Association, vol.90, n.430.

Faloutsos C., Lin K.-I. (1995) "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets". In M. J. Carey and D. A. Schneider, editors, SIGMOD, pages 163-174.

Fellegi I.P., Sunter A.B. (1969) "A Theory for record linkage", Journal of the American Statistical Association, 64, 1183-1210.

Jaro M.A. (1989) "Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida", Journal of the American Statistical Association, 89, 414-420.

Torelli N., Paggiaro A. (1999) "Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forse di lavoro" Working paper n.15 del progetto di ricerca cofinanziato MURST "Lavoro e disoccupazione: questioni di misura e di analisi", Dipartimento di Scienze Statistiche, Università di Padova.

Yancey W (2004), "Improving algorithm estimates for record linkage parameters", Research report series U.S.Census Bureau, http://www.census.gov/srd/papers/pdf/rrs2004-01.pdf

Winkler W.E. (1995), "Matching and Record Linkage", in Cox, Binder, Chinnappa, Christianson, Colledge, Kott (a cura di), Businness Survey Methods, Wiley & Sons, pp. 355-384.

Wolter K. (1986) "Some coverage error models for Census data", Journal of the American Statistical Association, 81, 338-346.

tuoto@istat.it

Improving of Household Sample Surveys Data Quality on Base of Statistical Matching Approaches

Ganna Tereshchenko - The Institute for Demography and Social Research of The National Academy of Sciences of Ukraine

In the present time for the official statistics bodies the problem of providing users of different levels by high-quality statistical information of the most essential demographic, social and socio-economic phenomena and processes becomes actual.

The socio-economic conditions of the population of Ukraine is defined on the basis of the information received from different sources – population census, the state sample surveys, labour statistics, demographic and social statistics, statistics of national accounts, etc.

The data from these sources mainly are used independently and aren't mutually coordinated, that results in essential reduction of efficiency of information using for statistical indicator estimation.

One of the most effective modern methodological approaches to the decision of this problem is the statistical matching of information from different sources.

Theoretical and methodological principles of statistical matching of information were developed mainly by such researchers as Rässler S., Rubin D.B., Marcello D'Orazio, Marco Di Zio, and Mauro Scanu.

Methodological principles of development and using of statistical matching methods in the state household surveys in Ukraine are presented.

Developed methodical approaches allow effective matching of different household sample surveys data received on samples of different design and for different periods of time, and also different household sample surveys used of harmonized indicators system.

By results of the executed researches it is established, that using of the statistical matching procedures for different population surveys can be the effective tool for increase of the indicators estimation efficiency as well as can provide the opportunity for the profound analysis of social and economic phenomena and processes.

Methodical maintenance for statistical matching of the labour force survey data, received on different designed samples is developed; it has allowed the essential improving of reliability level for employment and unemployment indicators estimates for regions of Ukraine.

The developed approaches are implemented into state statistics of Ukraine.

The methodological approach to statistical matching of separate data of labour force and household living conditions sample surveys is offered with use of harmonized indicators system. Procedures are testing at the data matching concerning status of household members' employment.

It has allowed analyzing the distribution of unemployed, defined by ILO methodology, on the level of average total per capita expenditures.

The information, which is given to users at practical using methods of statistical matching for estimation of indicators of household sample surveys, demands special attention.

It is necessary to give users the following information which describes quality of matching data: description of statistical matching methods, their settings and grounds of expedience of their using; ground of indicators reliability estimated on base of matched data; analysis of possible influence on the indicators estimates and their reliability level of the untaken into account consequences of using of statistical matching methods, etc.

a_tereschenko@ukr.net

Associated papers

Combined Register and Survey-based Data Gathering for Agricultural and Forestry Economics Statistics

Hannu Maliniemi, Paavo Väisänen - Statistics Finland

The purpose of Agricultural and Forestry Economics Statistics (AFES) is to measure, describe and analyse the formation of income from agricultural economic activity. The survey units are agricultural enterprises and the data collection concerns farming incomes subject to taxation, and expenditure, assets and debts of farms, as well as changes in the fixed assets of farms. The main sources of income data for the statistics are the Tax Register, the Farm Register and a statistical survey of farm enterprises. The new AFES replace two previous sets of statistics, the Agricultural Enterprise and Income Statistics and the Income and Taxation Statistics of Farms which have been compiled unchanged since 1973. The target population of the AFES is formed by combining the registers of the Ministry of Agriculture and Forestry and the Tax Register. Changes in the taxation data made it possible to compile statistics on farm incomes from total data. The previous method was based on a sample drawn from the Farm Register and the data were collected from the sample using statistical forms into which the Tax Authorities entered the taxable incomes of the selected farms. The reform of agricultural taxation opened the possibility to utilise the taxation data of all units taxed under the Farm Income Tax Act. Since 2004, the tax forms of all farms have been entered into the databases of the Tax Administration and the data are now available for the years 2004, 2005 and 2006. During this period the average number of agricultural taxation units was 145,000 and at the same time the Farm Register contained slightly under 70,000 farms. Since 2006, in addition to the forest owner farmers the target population has also included the forest owners living in cities.

The biggest challenge was to combine the two different registers with different register units. A new statistical unit, agricultural enterprise was defined to handle the combined farm and taxation register. The Farm Register uses a farm identification number but the Taxation Register uses an enterprise or person identification code.

The register data were supplemented by a survey in which incomes according to farm production were collected by questionnaire. The data collection was automated by building the questionnaire into accounting programs, which decreased the framers' response burden. The accounting program sends the data automatically from the farmer's computer to Statistics Finland via the Internet, which speeded up the data collection process and increased the response rate. The register and survey data are combined at the estimation phase, and the register data were additionally used to imput values to item non-response on the survey questionnaires. Register data were used to increase the precision of survey estimates by applying calibration techniques, in which the estimates of the totals of the register variables were benchmarked to the true values. An interesting question concerning the quality of the estimates is identification of the 70,000 farmers liable for agricultural tax whose farms are not included in the Farm Register and which in the previous years represented undercoverage of the target population.

The composite data opened new possibilities for research into agricultural economics by allowing detailed analyses of production lines, decisions on the activities of farms, and changes in the economic situation.