Conference on Small Area Estimation

March 26–27, 1998

U.S. Bureau of the Census

Abstracts

Mixed-Effect Logistic Regression Model for Household Mail Response in the Decennial Census

Eric Slud (e-mail: evs@math.umd.edu, University of Maryland and ASA/Census Research Fellow)

This talk presents results of fitting a logistic regression model with random block-group effects to household mail response to the census, using 1990 Delaware census data with demographic and geographic covariates at the block-group level. The random-effect models are motivated using residuals from a standard logistic regression, and are fitted as mixed-effect nonlinear regression models using Splus software.

A Transparent File For A One-Number Census

Cary Isaki (e-mail: Cary.T.Isaki@ccmail.census.gov), Michael Ikeda, Julie Tsay (U.S. Bureau of the Census), and Wayne Fuller (Iowa State University)

In the context of a population census employing sampling and estimation, we propose a method for constructing a census data file that is devoid of any evidence of sampling and estimation. Such a data file is said to be transparent. Using data from the 1995 Test Census, we construct a transparent file of short form items and compare the results with 1995 Integrated Coverage Measurement and synthetic estimates for small areas. Extensions to long form data are discussed.

Demographic Perspectives on Small Area Estimation of Population

Signe Wetrogan (e-mail: Signe.I.Wetrogan@ccmail.census.gov, U.S. Bureau of the Census)

The Intercensal Population Estimates Program, in the Population Division, U.S. Bureau of the Census produces population estimates for the nation, states, counties and places (cities, towns, and townships) as part of its program to quantify changes in population size and distribution since the last census. Required under Title 13, these estimates are used for: Federal and state funds allocation, survey controls, denominators for vital rates, administrative planning, and descriptive and analytical studies. The methodology for producing the intercensal estimates varies by geography. At the national, state, and county level, we use a demographic accounting method known as the cohort component technique. In this method, we estimate the separate components of population change—births, deaths, international migration and internal migration. For subcounty areas, we use a distributive housing unit method in which the estimated number of housing units for subcounty areas is used to distribute the county population to subcounty areas.

This presentation discusses the intercensal population estimates program and presents detail on the methodological approaches used to develop the national, state, county, and subcounty population estimates.

An Overview of the Census Bureau's Small Area Income and Poverty Estimates Program

Paul Siegel (e-mail: psiegel@census.gov, U.S. Bureau of the Census)

The Census Bureau has recently released estimates of income and poverty for all counties of the U.S. for income year 1993. The figures combine model and direct estimates, based on the Current Population Survey, to provide post-censal intelligence about income and poverty at the local (county) level. Heretofore estimates at this geographic level were available only decennially, from the census. These are the first of an

anticipated series of biennial estimates. The present paper provides some background for the project, and discusses the development of the models which underlie the county estimates.

Methods Used for Small Area Poverty and Income Estimation for Counties

Robin Fisher (e-mail: rfisher@census.gov) and Paul Siegel (U.S. Bureau of the Census)

Existing postcensal estimates of poverty and income at the county level are considered inadequate for various reasons: The Census is rapidly dated and the March CPS is not sufficiently reliable, especially for those counties which are not sampled by CPS. The goal of the Small Area Income and Poverty Estimates (SAIPE) project is to form these estimates. We modeled the number of poor in various age categories and median household and per capita income as a function of various variables taken from administrative records. We recognize two sources of "error"—sampling error and model error—and apply a shrinkage estimator to obtain estimates of number of poor or income by county. Finally, a ratio adjustment is used to make estimates consistent with the SAIPE state estimates. We describe the methods used to obtain these estimates and their standard errors in the January 15, 1998 release and present some empirical evaluations of the models.

Development of the 1993 SAIPE State Models

Bob Fay (e-mail: rfay@census.gov, U.S. Bureau of the Census)

This talk summarizes the methodology employed to estimate income and poverty characteristics for states in 1993 for the Census Bureau's Small Area Income and Poverty Estimation (SAIPE) project. Features of the model and estimation procedure may suggest wider application to similar problems. The talk also examines evidence from data for neighboring years about the properties of the estimation procedure and the consequences of alternative approaches.

Using Multiple Years of CPS Data in Estimating State Poverty Rates of School Aged Children

Bill Bell (e-mail: wbell@census.gov, U.S. Bureau of the Census) and Mark Otto (U.S. Fish and Wildlife Service)

The SAIPE project of the Census Bureau seeks to improve intercensal state and county income and poverty estimates. For states, Fay has developed regression models relating data from the Current Population Survey to variables drawn from administrative records and the decennial census, and used these models to produce improved state estimates via an empirical Bayes approach. We build on Fay's general model, extending it to incorporate multiple years of CPS data in various ways. We extend previous related work on small area estimation by using an explicit model for the CPS sampling errors. Our modeling framework allows us to investigate several interesting questions, including: (1) Are different regression parameters needed for the different years of CPS data? (2) What should be assumed about the relations over time of the true poverty rates? (3) Does it appear that using past years of CPS data can improve estimates in the current year? Preliminary results on these questions are obtained for state poverty rates of school-aged (5–17) children.

Small Area Estimation Using Loglinear Models for Complex Household Structures

Alan Zaslavsky (e-mail: zaslavsk@hcp.med.harvard.edu) and Elaine Zanutto (Harvard University)

The use of sampling for nonresponse followup (NRFU) in Census 2000 will create an unprecedented amount of missing data. Therefore, it is important to synthesize all available information to estimate the complete roster with acceptable accuracy. Potential sources of data for estimation of characteristics of nonrespondent households include survey data for households in the NRFU sample, administrative records data for some nonsample households, and data from respondents in the same block as the nonrespondents. Except in some special cases (which may include the design adopted for the 2000 decennial census), simple substitution of an alternative information source for the nonsampled households may produce estimates that are biased or inefficient. We propose a series of strategies for estimation of the characteristics of the nonsampled households using joint hierarchical loglinear models for those households and one or more of

the other data sources. We argue for evaluating these and other estimators using criteria that consider bias and variance at several different levels of aggregation of the data. We also present approximate formulae for the relationship between sample size and mean squared error that are useful in estimation of error of the estimators and in design of the NRFU sample.

Small Area Estimation in Microsimulation Analyses

Allen Schirm (e-mail: aschirm@mathematica-mpr.com, Mathematica Policy Research) and Alan Zaslavsky (Harvard University)

The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 has given states great flexibility to redesign their welfare programs. As they do so, policymakers need reliable estimates of the cost and distributional effects of alternative proposals. Microsimulation has been used extensively to obtain national estimates of the effects of proposed reforms to federal programs. However, current microsimulation models can often produce only imprecise state estimates because sample sizes are small. We will describe a model-based method for reweighting a microsimulation database to borrow strength and improve precision. A Poisson regression model is fitted to obtain an estimated prevalence in each state of every household in the database. This model is specified to control important aggregates at the state level, and the prevalences are expressed as a matrix of weights, with each household having a weight for every state. The set of weights calculated for a state makes the entire database look like that state rather than the whole country. Microsimulation estimates for a state are derived by passing through the microsimulation model all households in the database, not just the households actually observed in that state, and applying the appropriate weight for each household. We have estimated the biases and variances of the model-based estimates using the reweighted database and evaluated the model-based estimates relative to direct sample estimates for key estimands. We will present preliminary results from this evaluation.

Seasonal Adjustment of Small Area Estimates

Richard Tiller (e-mail: tiller_r@bls.gov, U.S. Bureau of Labor Statistics)

Statistical agencies routinely seasonally adjust large numbers of time series generated from periodic surveys. While these surveys produce highly reliable estimates for national aggregates, demographic and sub-national series are based on much smaller samples. Since the conventional time series decomposition process involves filters that are thought to effectively smooth the noise in the data, seasonal adjustment and trend estimation are rarely viewed in terms of a small area estimation problem. Obviously, small sample sizes do increase the variability of the data and conventional filters can substantially reduce this variation. However, the problem is more complicated because most periodic surveys have an overlapping design that induces strong autocorrelations in the survey error (SE). This latter feature creates a fundamental identification problem in the time series decomposition process. Without information on the correlation structure of the SE, it is not in general possible to identify key unobserved components of the series.

The time series approach to small area estimation provides a natural solution to this problem. SE is treated as an additional unobserved component of the time series, with the special advantage that it is objectively identified by design information. Given this information, a filter is constructed to suppress SE variation along with the seasonal and irregular noise in the population. Thus, seasonal adjustment and small area estimation for periodic surveys are closely related operations and may be solved by the same method.

Sub-national data from the Current Population Survey are used as examples. The effect that the survey design has on the conventional X-11 decomposition of the survey data and the gain from combining a small area estimator with conventional time series estimators are examined.

Empirical Bayes Estimation of Median Income of Four-Person Families by State using Time Series and Cross-Sectional Data

Gauri Datta (University of Georgia and ASA/Census Research Fellow), and Partha Lahiri and Tapabrata Maiti (e-mail: tmaiti@unlinfo.unl.edu, University of Nebraska)

The Department of Health and Human Services (HHS) uses estimates of the median income of four-

person families for all the fifty states and the District of Columbia to formulate its energy assistance program for low income families. Such estimates are provided by the U.S. Census Bureau on an annual basis.

A hierarchical time series model is considered to combine information from three relevant sources: (a) Current Population Survey (CPS), (b) Decennial Censuses and (c) Bureau of Economic Analysis. An empirical Bayes (EB) method is used to smooth the CPS estimates of the median income of four-person families for the states. The proposed method is an advancement over the EB method currently used by the U.S. Bureau of the Census in the sense that it uses a more realistic model, provides maximum likelihood and residual maximum likelihood method of variance components estimation and provides a valid measure of uncertainty of the proposed estimates which captures all different sources of variations. Compared to the corresponding hierarchical Bayes (HB) estimation, the method is very easy to implement and saves a tremendous amount of computer time. The proposed EB method is compared with rival estimators using the 1989 four-person median income figures obtained from the 1990 Census.

A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems

Gauri Datta (e-mail: Gauri@stat.uga.edu, University of Georgia and ASA/NSF/Bureau of Labor Statistics/Census Research Fellow) and Partha Lahiri (University of Nebraska)

We obtain a second order approximation to the mean squared error (MSE) and its estimate of empirical or estimated best linear unbiased predictor (EBLUP) of a mixed effect in a general mixed linear normal model which covers many important small area models in the literature. Unlike previous research in this area, we provide a unified theory of measuring uncertainty of an EBLUP for a complex small area model where the variance components are estimated by various standard methods including restricted or residual maximum likelihood (REML) and maximum likelihood (ML). It turns out that the mean squared error (MSE) approximations for the REML and the ML methods are exactly the same in the second order asymptotic sense. However, the second order accurate estimator of this MSE mean squared error based on the former method requires less bias correction than the one based on the latter method. This is due to a result in the paper which shows that the bias of the REML estimators of variance components is of lower order than that of the ML estimators. A simulation is undertaken to compare different methods of estimating the variance components and to study the properties of various estimators of the MSE of the mixed effect. In our context it is interesting to note that the residual likelihood is same as the conditional profile likelihood (CPL) of Cox and Reid (1987). Thus, this paper addresses an important open problem raised by Cox and Reid (1987) in small area prediction using the CPL method.

On the Measure of Uncertainty of EBLUP in Small-Area Estimation Problems

Ferry Butar (University of Central Florida) and Partha Lahiri (e-mail: plahiri@unlinfo.unl.edu, University of Nebraska)

Small-area typically refers to a small geographic area or a demographic group for which very little information is obtained from the sample surveys. An empirical best linear unbiased prediction (EBLUP) method uses sample survey data in conjunction with relevant supplementary data which are obtained from various administrative sources. The method has been found to be very useful in many applications of small-area estimation and related problems.

In this paper, a method based on bootstrap samples is proposed to measure the accuracy of the proposed EBLUP of a small-area characteristic. A simple approximation of the method which does not require any bootstrap simulation is also proposed. The model expectation of the proposed measure of uncertainty of the EBLUP is equal to the mean squared errors (MSE) of the EBLUP up to the order $1/m$. It is interesting to note that for a special case of our model, the measure is identical, up to the order $1/m$, with a measure of uncertainty previously proposed by Morris (1983). Since Morris' measure is an approximation of a hierarchical Bayes posterior variance, the proposed method enjoys both desirable frequentist and hierarchical Bayes properties.

Two well-known data sets are considered to compare the performance of the proposed method with some existing methods. The results from a simulated study are also presented to demonstrate the performances of the rival methods when m is moderately large.

Small Area Estimation: An Appraisal with Updates

J.N.K. Rao (email: JRAO@math.Carleton.CA, Carleton University)

Sample survey data can be used to derive reliable estimates for large areas, but sample sizes in small areas or domains are seldom large enough for direct estimates to provide adequate precision for small areas. This makes it necessary to employ indirect estimators that borrow strength from related areas. This overview focuses on model-based indirect estimators obtained from random effects models that relate a character of interest to auxiliary variables with known population characateristics. In particular, empirical Bayes (EB), empirical best linear unbiased prediction (EBLUP) and hierarchical Bayes (HB) will be considered under two types of models: area level and unit level. Methods for measuring the variability associated with the estimates will be compared. Design-consistent estimation under unit level models will be studied. Extensions to generalized linear mixed models, including logistic regression and probit regression for binary data, and time series models will be outlined. Several recent applications will be mentioned.

Generalized Linear Models for Small Area Estimation

Malay Ghosh (e-mail: ghoshm@stat.ufl.edu, University of Florida)

Small area estimation has become a topic of increasing importance in recent years. Hierarchical Bayesian methods have proved to be very effective for solving small area estimation problems as these provide a systematic connection of local areas through models. By now, there is an abundance of literature dealing with applications of normal theory hierarchical Bayesian models for small area estimation. However, often the survey data are discrete or categorical for which the normal theory analysis is not very appropriate. Generalized linear models are suitable for handling both discrete and continuous data. We shall review some of the existing literature on generalized linear models for small area estimation, and provide a few pointers towards possible application of such methods for more complex surveys.

Small Area Estimation for the National Household Survey on Drug Abuse: Solutions Based on Survey Weighted Logistic Mixed Models

Ralph Folsom (e-mail: folsom@rti.org, Research Triangle Institute)

In 1996 the Substance Abuse and Mental Health Services Administration published small area prevalence estimates for eleven drug use related attributes. The presentation will describe the methodology used to produce prevalence estimates and confidence intervals for selected States and Metropolitan areas using pooled data from the 1991 through 1993 NHSDA samples. Highlights of the methodology include survey weighted adaptations of Breslow and Clayton's penalized quasilikelihood (PQL) solution for logistic mixed models with age group specific, nested random effects, for States and MSA/County PSUs. Fixed regressors included block group, census tract, and county level predictors. MSA and State statistics were formed as population weighted averages of block group level prevalence estimates for 32 age by gender by race/ethnicity subpopulations. Confidence intervals were based on the empirical Bayes posterior variances and covariances among the parameters and a first order Taylor linearization of the logit transformed prevalence. Cross-validation results were used to judge model goodness-of-fit. The talk concludes with a discussion of methodology improvements planned for the ongoing 1994–1996 NHSDA small area estimation project. The improvements feature a survey weighted Monte Carlo Markov Chain (MCMC) solution for the logistic mixed model.

Small Area Estimation Using Data from the Third National Health and Nutrition Examination Survey (NHANES III)

Bill Davis (e-mail: wbd1@cdc.gov, Klemm Analysis Group) and Donald Malec (National Center for Health Statistics)

This paper presents a general methodology for making small area estimates with data from a national survey. We illustrate the methodology using data from the the third National Health and Nutrition Examination Survey (NHANES III), which was carried out during the period 1988 through 1994. The methodology is designed to estimate prevalence rates. We apply the methodology to estimate prevalence rates of the following two important health outcomes: overweight and high blood-lead levels. A generalized linear mixed model is used for the binary outcome. Our likelihood includes an adjustment for sample selection. This method allows the statistical weights to impact the parameter estimation. We use Markov-Chain Monte-Carlo (MCMC) to estimate the model parameters and make estimates by county and by demographic group—defined by age, race/ethnicity, and gender. This method allows us to make small area estimates with estimated precision by state and demographic group. For both outcomes, we show the state-level model-based estimates on a map and show the coefficient of variation as a function of sample size. The model-based approach preferentially uses state data. To validate our model-based estimates, we compare with design-based estimates at the national level and obtain excellent agreement for both outcomes. As a further check, we compare our model-based state estimates for overweight prevalence with synthetic estimates using data from NHANES II. This comparison shows the improvement we obtain by use of the county education as a covariate.

Bayesian Analysis of Mortality Rates for U.S. Health Service Areas

Balgobin Nandram (e-mail: balnan@wpi.edu, Worcester Polytechnic Institute), Joe Sedransk (Case Western Reserve University), and Linda Pickle (National Center for Health Statistics)

This talk summarizes our research on alternative models for producing age specific and age adjusted mortality rates for one of the diseases, all cancer for white males, presented in the Atlas of United States Mortality, published in 1996. We use Bayesian methods, applied to four different models. Each assumes that the number of deaths, d_{ij}, in health service area i, age class j has a Poisson distribution with mean $n_{ij}\lambda_{ij}$ where n_{ij} is the population at risk. The alternative specifications differ in their assumptions about the variation in the $\ln \lambda_{ij}$ over health service areas and age classes. We use expected predictive deviances, posterior predictive p-values and a cross-validation exercise to evaluate the concordance between the models and the observed data. The models captured both the small area and regional effects sufficiently well that no remaining spatial correlation of the residuals was detectable, thus simplifying the estimation. We summarize by presenting point estimates, measures of variation and maps.

Small Area Estimation and Visualization Using Nonparametric Regression

Gerald Whittaker (e-mail: gerryw2@econ.ag.gov, U.S. Department of Agriculture)

Survey data collected by the U.S. Department of Agriculture have traditionally been analyzed on the basis of political boundaries, and the results presented in tables. Nonparametric regression provides a method to analyze the data at a lower level of aggregation. The results of nonparametric regression of a variable as a function of geographic coordinates can be presented as surfaces or maps. These surfaces typically contain much more information than a simple map based on political boundaries. While some information may be lost by smoothing at the point location of the observations, the ease of interpretation of a surface compensates for the loss. The averaged shifted histogram (ASH) implementation of nonparametric regression was chosen for this work over several alternatives. The ASH is computationally very fast, and it is relatively easy to account for a complex survey design in estimation. The results of the ASH estimation are easily displayed via GIS software. The spatial analysis of covariates using the ASH has provided new insights into applications such as the relation of government payments to agricultural land values, spatial distribution of government payments, and environmental effects of agricultural nutrients, among others.

Hierarchical Bayes Estimation of Hunting Success Rates with Spatial Correlations

Zhuoqiong He (Missouri Department of Conservation) and Dongchu Sun (e-mail: dsun@stat.missouri.edu, University of Missouri)

A Bayesian hierarchical generalized linear model is used to estimate hunting success rates at the sub-area level for post-season harvest surveys. The model includes fixed week effects, random geographic effects, and spatial correlations between neighboring sub-areas. The computation is done by Gibbs sampling and adaptive rejection sampling techniques. The method is illustrated using data from the Missouri Turkey Hunting Survey Spring Season in 1996. Bayesian model selection methods are used to demonstrate that there are significant week differences and spatial correlations of hunting success rates among counties. The Bayesian estimates are also shown to be quite robust in terms of changes of hyperparameters.