Survey Methodology Research
FY 2011 - FY 2016 Awards List

Awards from this competition were jointly reviewed and supported by NSF's Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies represented by the Federal Committee on Statistical Methodology (FCSM). The following agencies provided direct financial support for these awards:

National Agriculture Statistics Service, DoA
Bureau of Justice Statistics, DoJ
National Center for Education Statistics, DoE
Bureau of Labor Statistics, DoL
National Center for Health Statistics, DHHS
Department of Transportation
Science Resource Studies, NSF
Economic Research Service, USDA
Social Security Administration
Energy Information Administration, DoE
U.S. Census Bureau, DoC

FY 2016

Title: Doctoral Dissertation Research: The Impact of Working Memory on Response Order Effects and Question Order Effects in Telephone and Web Surveys
Proposal ID: 1631994
PI: Robert F. Belli
Institution: University of Nebraska - Lincoln
Amount: $16,000
Duration: 12 Months (Estimated)

Abstract
This research project will investigate whether working memory has an impact on the responses participants provide on survey questionnaires. Working memory is a cognitive system that provides both the limited storage for relevant information and the temporary processing needed to perform ongoing mental tasks. When individuals participate in a survey, for each question presented, the participant must remember the question and response options while formulating a response. This research will extend the current understanding of the impact of working memory on participants' response selections in questions about attitudes. The research will contribute new knowledge about the cognitive aspects of survey methodology. The results may have implications for the design of questionnaires and could be of particular importance to large, nationally representative surveys like the Health and Retirement Study (HRS), where it is essential to measure cognitive functioning. As a Doctoral Dissertation Research Improvement award, support is provided to enable a promising student to establish a strong, independent research career.

This project will collect new data through administering a survey to a random sample of Nebraska residents. Three research questions will be addressed with this data: 1) Does working memory impact question order and response order effects? The project will assess whether these effects are consistent across high and low working memory capacity groups for younger and older participants. 2) Is the impact of working memory on question-order effects and response-order effects consistent across two modes of survey administration, telephone and web? To examine this question, the same set of questions will be administered in a telephone survey and web survey. 3) How do participants compare across two different sets of memory measures? A subset of memory measures from the Health and Retirement Study (HRS) will be included in the survey to compare with other questions designed to specifically measure working memory capacity. The data will be analyzed using statistical methods such as chi-square tests and logistic regression.

Additional Information:

Title: Improving Probabilistic Record Linkage and Subsequent Inference
Proposal ID: 1631970
PI: Jared Murray
Institution: Carnegie-Melon University
Amount: $225,000
Duration: 36 Months (Estimated)

Abstract

This research project will develop methods for linking records across databases in the absence of unique identifiers such as Social Security numbers and for making inference using the linked data files. Record linkage is a perennial and challenging problem across the social sciences, with important applications in areas such as demography, economics, public health, and official statistics. Plummeting costs of new forms of data collection and storage and the proliferation of "big data" have increased the need for merging such databases as researchers and statistical agencies struggle to integrate carefully curated datasets with messy and incomplete data from historical, administrative, and commercial sources. The methods developed in this project will facilitate the successful integration of different data sources, thus generating new resources for future research. These combined data sources may also provide some alternatives to expensive survey data collection in an era of declining response rates. Freely available software will be developed and stored in a public repository.

The increasing desire to deploy probabilistic record linkage has spurred significant research into various components of the process, such as how to compare records, how to reduce the number of record comparisons to keep the problem computationally feasible, how to quantify the weight of evidence for or against a link between records, and how to ultimately generate a merged database. Often these components are studied in isolation from each other and from the ultimate goal of making inferences using the merged files. This research project will take a more holistic view of the record linkage process in order to advance the state of the art. The project has two primary goals. The first goal is to develop new models for record linkage that incorporate the impact of preprocessing methods that reduce the total number of record pairs to be evaluated. While widely deployed and well motivated, these methods have effects on subsequent modeling that are not well understood. The second goal is to enhance understanding of uncertainty and error throughout the process and to develop imputation methods for propagating error due to uncertain record links and other missing data, such as item nonresponse in a survey. These methods will be designed with an eye toward large applications that require new computational approaches. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: A Spectral Framework for Network-Driven Sampling
Proposal ID: 1612456
PI: Karl G. Rohe
Institution: University of Wisconsin - Madison
Amount: $ 172,365
Duration: 36 Months (Estimated)

Abstract

Probability sampling drastically reduces the burden of research in various disciplines because statistical inference can extend conclusions from a sample to the entire population. However, classical sampling techniques require a sampling frame that lists each individual in the population and a way of contacting each individual. In many settings, a sampling frame is not available. In others, a sampling frame is too expensive to compile or only covers a biased subset of the population. Particularly with hard-to-reach populations, network-driven sampling provides one of the only ways to find members of the population. Leveraging a network to find a target population appears in many disciplines with a multitude of names: respondent-driven sampling, snowball sampling, web crawling, link-tracing, breadth-first search, co-immunoprecipitation, and chromatin immunoprecipitation. These disparate techniques all provide access to hard-to-reach and networked populations by essentially asking participants to refer friends. As a result, these are all network-driven techniques. Classical sampling theory does not apply to network-driven sampling because friends are similar; this induces dependence between samples that is influenced by the underlying social network. Preliminary research conducted by the investigator identifies a critical threshold that relates the structure of the social network to the referral rate in the sampling tree; beyond this critical threshold, standard network-driven approaches produce highly uncertain estimates. This research aims to produce new statistical techniques that continue to perform well beyond the critical threshold. Moreover, this project will study novel forms of network-driven data collection that incorporate additional information to produce more representative samples.

Classical sampling results are not applicable to network-driven sampling because friends are similar, inducing dependence between samples. Previous theoretical results show that some network-driven studies do not obtain square root n-consistent estimators. Whether a study obtains square root n-consistency depends on both (i) the spectral properties of the underlying social network and (ii) the growth of the sampling tree. This research aims to provide new estimators that correct for the dependence between samples. These dependence-corrected estimators can obtain square root n-consistency, even when current estimators do not. This project will also construct new diagnostics and new sampling designs for network-driven sampling. The new spectral framework will provide a suite of theory, methodology, and practices that will enable studies to obtain square root n-consistent estimators.

Additional Information:

Title: Doctoral Dissertation Research: Gaze Patterns During Video-mediated Interviews
Proposal ID: 1632015
PI: Michael F. Schober
Institution: The New School
Amount: $15,988
Duration: 12 Months (Estimated)

Abstract:

This research project will explore how and why the small self-view window (a window with a live video feed of oneself in the corner of the screen) in video-mediated communication (such as Skype) can affect people's disclosure of sensitive information during survey interviews. In an earlier study, respondents in the self-view condition disclosed more socially undesirable (that is, more embarrassing) behaviors and perceived the interview to be less sensitive. This project will expand on earlier research by tracking where survey respondents look on the screen while answering sensitive questions via video. Examining gaze patterns and how they might differ for different respondents or questions will provide a basis for understanding why people disclose more and find the interview less sensitive with a self-view window. The findings will increase understanding about how nonverbal behaviors correlate with disclosure. Video-mediated interviewing is a potentially cost-effective mode of data collection that could facilitate the collection of policy-relevant data on some hard-to-reach populations. The data collected may be useful for informing the design of new video-based data-collection interfaces. Project results will be presented at conferences and publication venues that maximize impact across several disciplines. As a Doctoral Dissertation Research Improvement award, support is provided to enable a promising student to establish a strong, independent research career.

Video-mediated interviewing offers a unique combination of features that differ from those in other face-to-face or audio interactions, including the ability to see oneself in a self-view window. This research project will compare the responses and gaze patterns of adult participants interviewed in a controlled laboratory setting over Skype. The experiment will use a set of sensitive and non-sensitive questions from U.S. government and social scientific surveys. Respondents will be randomly assigned either to a self-view (respondents can see themselves and the interviewer) or no self-view (respondents can see only the interviewer) condition. Measures of gaze will be recorded using an unobtrusive eye-tracking system. The data will be analyzed to determine when and how often self-view respondents look at the images of themselves, whether respondents who disclose more look at the self-view more, and how gaze location, duration, and frequency of glances differ for sensitive vs. neutral questions and for less socially desirable answers.

Additional Information:

Title: Semiparametric Estimation and Variable Selection in the Presence of Nonignorable Nonresponse
Proposal ID: 1612873
PI: Jun Shao
Institution: University of Wisconsin - Madison
Amount: $ 290,431
Duration: 36 Months (Estimated)

Abstract

This project aims to use the tools of science and technology studies (STS) to enhance empirical understanding of scientific collaboration across scales, disciplines, and international borders. The investigator will engage in an ethnographic inquiry of two international science projects seeking to integrate knowledge about microbes in the Brazilian Amazon, a global reservoir of biodiversity and frontier for international science partnerships. She will examine the collaborative practices of scientists using qualitative and visual methods to complete her research objectives, which are to identify concrete pathways to more effective knowledge integration and to advance theoretical understanding of the ontological and geopolitical dimensions of interdisciplinary science and international collaboration. The project will generate societal benefits by identifying best practices in interdisciplinary and international science. It will also enhance public understanding of new trends in molecular and microbial environmental science through humanistic inquiry, specifically art and visual medium. Expected outcomes of the project include the mentoring and training of two students in STS theory and methods; a white paper aimed at public funding agencies and science practitioners; several publications in high-impact journals; and a three-month museum visit with development of an associated website and educational materials. Dissemination plans include presentations at two international conferences, co-authored publications with students and collaborators from Brazil and Canada, and the development of a project website in English and Portuguese.

This ethnographic study of environmental science in the Amazon will identify pathways to more effective knowledge integration and will advance theories in STS regarding the ontological and geopolitical dimensions of interdisciplinary science and international collaboration. It will make important intellectual advances on three fronts: (1) New empirical knowledge regarding effective knowledge integration, including the barriers and best practices in collaboration across scales, disciplines, and cultures; (2) Novel research methods that visually explain how disciplinary cultures diverge in their approaches to framing and integrating objects of inquiry at multiple scales; and (3) Improved theories of science as visual practice, grounded in how individual subjects experience and are even transformed by encounters between differing scaling imaginaries and practices. With a focus on the Brazilian Amazon, the project will produce useful empirical insights about knowledge production in a region where scientific exploration, expertise, and knowledge extraction has been mired in political controversy for decades. Research that enhances systematic understanding of factors affecting collaborative knowledge production in response to environmental change is a recognized priority.

Additional Information:

FY 2015

Title: Collaborative Research: Multilevel Regression and Poststratification: A Unified Framework for Survey Weighted Inference
Proposal Number: 1534414
PI: Andrew Gelman
Institute: Columbia University
Amount: $91,332
Duration: 36 Months (Estimated)

Abstract

This research project will develop a unified framework for survey weighting through novel modifications of multilevel regression and poststratification (MRP) to incorporate design-based information into modeling. Real-life survey data often are unrepresentative due to selection bias and nonresponse. Existing methods for adjusting for known differences between the sample and population from which the sample is drawn have some advantages but also practical limitations. Classical weights are subject to large variability and can result in unstable estimators, while regression approaches present computational and modeling challenges. The new framework developed by these investigators will allow adjustment for selection bias and nonresponse as well as improvements in design-respecting inference. Using this approach, survey analysts will be able to properly account for non-ignorable design issues in the regression framework, and practitioners who conduct surveys in government, academic, commercial, and non-profit sectors will be able to construct statistically efficient survey weights in a routine manner. This new framework may be applicable to problems resulting from the newly emerging explosion of "big data," such as integration of surveys from multiple sources, analysis of streaming data, and respondent- driven sampling. The project will develop software that can be accessed by the general research community.

This research project will connect survey weighting with poststratification under the framework of MRP. In MRP, data are partially pooled during the modeling process and then local estimates are combined via poststratification to obtain the population inference. This smoothed estimation borrows information from neighboring poststratification cells and allows flexible multilevel modeling strategies that have the potential to be robust to model misspecification. The project generalizes MRP to handle weighting adjustments for regression, deep interactions, calibration for non-census variables, complex survey design, multistage sampling, multiple survey frames, and other complications that arise in real-world survey analysis. The new methods will be applied to two ongoing surveys, the New York Longitudinal Poverty Measure study and the Fragile Families and Child Wellbeing study. Computations will be performed using the open source Bayesian program Stan and will be freely disseminated. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Empirical Assessment of Respondent Driven Sampling from Total Survey Error Perspectives
Proposal ID: 1461470
PI: Sunghee Lee
Institution: University of Michigan Ann Arbor
Amount: $190,830
Duration: 36 Months (Estimated)

Abstract

This research project will provide an empirical investigation into the realities of respondent driven sampling (RDS) data collection. RDS is a new method for sampling rare or hidden populations. The method starts with the members of the target population and traces their social networks as well as the networks of those to whom they are connected. Although this method is growing in popularity, there is a scarcity of publicly available RDS datasets. As a result, methodological assessments of RDS are very limited. This research project will conduct a theory-drive methodological assessment of RDS. The project will provide a platform to develop design features for RDS studies that minimize potential errors and violations of critical assumptions. By developing appropriate inference strategies, rare or hidden populations will be more fully represented in the data collected by RDS methods. Absent these types of examinations, behavioral and social science data collected through RDS runs the risk of mischaracterizing rare or hidden populations in unknown ways. To the extent that these data inform public policies, rare or hidden populations may benefit from this research. New datasets will be generated and made publicly available to promote methodological research beyond this project.

This research project will conduct an empirical assessment RDS within the Total Survey Error (TSE) framework. TSE is a framework in survey methodology that allows for the systematic examination of errors. This project seeks to improve current RDS data collection and inference practices by examining sampling productivity, error properties, and replicability. The investigators will collect data on two rare populations in Los Angeles County for which external probability-based sample data are available. The probably-based data will be used as "gold standards" against which estimates from the RDS collected data will be compared, providing a unique opportunity to assess RDS not only as a sampling method but also as a data collection method. The data sets will empirically inform a simulation study that will examine the effects of various network structures and response propensities on sampling productivity and inferential errors. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Shape-Constrained Estimation and Inference for Surveys
Proposal ID: 1533804
PI: Mary Meyer
Institution: Colorado State University
Amount: $499,978
Duration: 48 Months (Estimated)

Abstract

This research project will provide a new class of tools for survey practitioners to improve the stability and precision of survey estimates. Surveys represent a key source of data for government, business, and academics. The cost of conducting surveys has increased dramatically during the last decade, however, while response rates have declined. The new methods developed by this project will take advantage of "soft" or vague information often available in practice in many different contexts but not directly used or used in only "ad hoc" ways. This information is referred to as "shape constraints" in statistics. The new shape-constrained survey estimation tools will represent an important advance in design-based estimation, which continues to be the primary approach used in large-scale government surveys. At the same time, this work will be the first application of shape constraints in the survey context. The new tools will be especially useful for large-scale survey agencies and have the potential to lead to considerably improved efficiency and to result in estimates that satisfy the qualitative properties expected by data users. The investigators will develop user-friendly software.

This research project will develop new survey estimation methods for a variety of different types of shape constraints that can be encountered in practice, together with statistical tools to assess whether the shape constraint holds in the survey population and to estimate the precision of the estimates. Shape-constrained estimation will be introduced in several important survey methodological areas of practical interest, including in estimation for sub-populations, post-stratification, adjusting for nonresponse and small area estimation. The research will result in new statistical algorithms for weighted constrained estimation, in statistical testing methods for the validity of the shape constraints for survey data, and in a new class of weight calibration methods that are based on shape constraints. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Collaborative Research: Multilevel Regression and Poststratification: A Unified Framework for Survey Weighted Inference
Proposal ID: 1534400
PI: Yajuan Si
Institute: University of Wisconsin-Madison
Amount: $328,667
Duration: 36 Months (Estimated)

Abstract

This research project will develop a unified framework for survey weighting through novel modifications of multilevel regression and poststratification (MRP) to incorporate design-based information into modeling. Real-life survey data often are unrepresentative due to selection bias and nonresponse. Existing methods for adjusting for known differences between the sample and population from which the sample is drawn have some advantages but also practical limitations. Classical weights are subject to large variability and can result in unstable estimators, while regression approaches present computational and modeling challenges. The new framework developed by these investigators will allow adjustment for selection bias and nonresponse as well as improvements in design-respecting inference. Using this approach, survey analysts will be able to properly account for non-ignorable design issues in the regression framework, and practitioners who conduct surveys in government, academic, commercial, and non-profit sectors will be able to construct statistically efficient survey weights in a routine manner. This new framework may be applicable to problems resulting from the newly emerging explosion of "big data," such as integration of surveys from multiple sources, analysis of streaming data, and respondent- driven sampling. The project will develop software that can be accessed by the general research community.

This research project will connect survey weighting with poststratification under the framework of MRP. In MRP, data are partially pooled during the modeling process and then local estimates are combined via poststratification to obtain the population inference. This smoothed estimation borrows information from neighboring poststratification cells and allows flexible multilevel modeling strategies that have the potential to be robust to model misspecification. The project generalizes MRP to handle weighting adjustments for regression, deep interactions, calibration for non-census variables, complex survey design, multistage sampling, multiple survey frames, and other complications that arise in real-world survey analysis. The new methods will be applied to two ongoing surveys, the New York Longitudinal Poverty Measure study and the Fragile Families and Child Wellbeing study. Computations will be performed using the open source Bayesian program Stan and will be freely disseminated. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information :

Title: Collaborative Research: Record Linkage and Privacy-Preserving Methods for Big Data
Proposal Number: 1534412
PI: Rebecca Steorts
Institute: Duke University
Amount: $265,579
Duration: 36 Months (Estimated)

Abstract

This research project will develop sound statistical and machine learning techniques for preserving privacy with linked data. Social entities and their patterns of behavior is a crucial topic in the social sciences. Research in this area has been invigorated by the growth of the modern information infrastructure, ease of data collection and storage, and the development of novel computational data analyses techniques. However, in many application areas relevant and sensitive information is commonly located across multiple databases. Data analysis is inherently impossible without merging databases, but at the cost of increasing the risk of a privacy violation. This research will address the problem of how to perform valid statistical inference in the presence of multiple data sources, data sharing, and privacy in the age of "big data." The investigators' new modeling construct for inference and uncertainty quantification will contribute to both statistics and the many disciplines for which statistics is a principal tool. The methods will have a wide range of applications in the social, economic, and behavioral sciences, including medicine, genetics, official statistics, and human rights violations. The investigators will collaborate with post-doctoral researcher and with graduate and undergraduate students. The statistical methods will be encapsulated in open-source software packages, allowing off-the-shelf use by practitioners while facilitating more detailed control and extensions.

This interdisciplinary research project will improve upon methods in record linkage and privacy using state-of-the-art techniques from statistics and machine learning. Record linkage is the process of merging possible noisy databases with the goal of removing duplicate entries. Privacy-preserving record linkage (PPRL) tries to identify records that refer to the same entities from multiple databases without compromising the privacy of the entities represented by these records. The research will focus on three aims: (1) development of new Bayesian methods for PPRL, where the error can be propagated exactly across the entire linkage process and into statistical inference, including new privacy measures to capture a tradeoff between utility and risk of any individual risk in a linked database; (2) development of new robust methods for realizing synthetic data releases post-linkage with differential privacy guarantees and its relaxations to address additional layers of privacy and support broader data sharing; and (3) exploration of "big data" methods such as variational inference to address scalability and latent cluster exchangeability issues existing within linkage and privacy, such that the new methods can scale to multiple and large databases. The new methods will be scalable and assess uncertainty throughout the entire linkage and privacy process and can be evaluated using Bayesian disclosure risk and Bayesian differential privacy. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

FY 2014

Title: Characterizing Nonresponse Error across General Population Survey Data Collection Modes
Proposal ID: 1424433
PI: Philip Brenner
Institute: University of Massachusetts Boston
Amount: $308,225
Duration: 36 Months (Estimated)

Abstract

This research project will advance knowledge about nonresponse error introduced by different modes of data collection for a wide variety of survey measures. The results of this research will contribute to the development of surveys employing multiple modes, thus enabling greater public participation and strengthening survey data quality. Population-based surveys are used to measure everything from unemployment to the health of the nation, guiding policy decision making at national, state, and local levels. But responses to surveys in general, and telephone surveys in particular, have declined significantly over the past two decades. As a result, researchers are exploring multiple modes of survey data collection. This project will provide additional information about the magnitude of nonresponse error among alternative modes of data collection.

This project will estimate nonresponse error directly rather than inferring it from demographic differences between those who respond and do not respond to different survey protocols. Nonresponse error will be measured for two modes of survey administration: computer-assisted telephone interviewing (CATI) and telephonic interactive voice response (IVR). To minimize mode effects, questions for the test surveys will be selected among those least likely to be impacted by mode. These questions will be chosen among questions included in several federal surveys, such as the Health Information National Trends Survey, National Household Travel Survey, the Behavioral Risk Factor Surveillance System survey, the National Crime Victimization Survey, the National Immunization Survey, and the Current Population Survey. Using an address-based sample frame, residents in diverse neighborhoods in the Boston metropolitan area will be randomly assigned to one of two experimental survey modes: CATI or IVR. At the end of survey field period, nonrespondents will be assigned to in-person interviewers. The interviewers will contact household members who did not respond to the survey, inviting them to complete the survey. Monetary incentives will be provided to encourage response. In-person interviews typically have the highest response rates among any survey administration mode. Data from CATI and IVR nonrespondents will be used to assess nonresponse error for key measures included in the survey. The resulting combined samples will provide direct estimates of the variables included in the survey with a minimum of nonresponse error against which to compare estimates from the two survey administration modes. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Adjusting for Unit Nonresponse That May Not be Missing at Random
Proposal ID: 1424492 PI:
Phillip Kott
Institute: Research Triangle Institute Amount:
$145,000
Duration: 24 Months (Estimated)

Abstract

This research project will further understanding of a statistical technique for measuring and reducing the potential for biases from non-response in sample surveys. Sample surveys are a principal tool for providing accurate information crucial for good economic and social decision making. Nonresponse in sample surveys conducted both privately and by government agencies is increasing. With that increase, the potential for serious biases in estimates derived from surveys and used in critical decision making grows. This research will address that problem, particularly in cases where non-response may be a function of variables having missing values. Extensions of this research will have broad application to multi-phase surveys, non-probability surveys, and other statistical applications plagued by potential selection bias whether due to nonresponse, coverage errors, or both. The project will develop powerful, publically available tools for implementing this approach.

Calibration weighting can be used to remove potential selection biases due to unit non-response or coverage errors. This research project will demonstrate the advantages in terms of reductions in non-response bias and increases in accuracy of having more calibration than model variables when non-respondents are determined not to be missing at random. In so doing, it will introduce a better method of calibration weighting in this context than currently exists in the literature. The research also will investigate new tests for determining whether a set of calibration variables can by itself serve as the model variables or whether a survey variable truly needs to be added to the response model to avoid selection bias. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Using Multilevel Regression and Poststratification to Measure and Study Dynamic Public Opinion
Proposal ID: 1424962
PI: Justin Phillips
Institute: Columbia University
Amount: $300,000
Duration: 36 Months (Estimated)

Abstract

This research project will develop techniques for using national survey data to estimate dynamic measures of public opinion across a variety of types of subnational units such as states, congressional districts, and state legislative districts. These techniques will allow researchers to generate accurate estimates of public opinion over time by fine-grained demographic-geographic-temporal subgroups. National surveys are designed to give good estimates of national public opinion at a particular point in time. They do not, however, necessarily give good estimates of opinion for subnational units. They also do not allow for understanding time trends of within these units. Recent advances have been made on estimating subnational opinion, but this work has yet to meet the special challenges of measuring opinion over time. This research will improve statistical tools, assess the accuracy of these tools, and create a set of guidelines for proper implementation. The researchers will employ these new tools to make substantive contributions to social science research. The newly developed techniques will be made available to a broad range of researchers and poll analysts through the creation and distribution of statistical software packages.

This project will develop a dynamic multilevel regression and poststratification (MRP) technique that will allow researchers to generate time-varying estimates of public opinion. MRP is a statistical technique that uses individual survey responses from national opinion polls coupled advances in Bayesian statistics and multilevel modeling to generate opinion estimates by demographic-geographic subgroups or "types." Dynamic MRP improves on standard MRP by taking advantage of additional information over time. Using this approach, the project will undertake a new investigation of the relationship between public preferences and policymaking at the state level. Specifically, the project will focus on two important issues: the death penalty and same-sex marriage. By tracking and tracing changes in opinion and policy over long periods of time, the investigators will gain causal leverage on the effects of public opinion that is lacking from existing research efforts. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

FY 2013

Title: Some Contributions to Sampling Theory with Applications
Proposal ID: 1327359
PI: Malay Ghosh
Institute: Westat Inc
Amount: $160,185
Duration: 36 Months (Estimated)

Abstract

This project will develop methods for small area estimation. Small area estimation generally requires the use of models, either explicitly or implicitly. These model-based estimates can differ widely from the direct estimates, especially for areas with very low sample sizes. While model-based small area estimates are very useful, one potential difficulty with these estimates is that when aggregated, the overall estimate for a larger geographical area may be quite different from the corresponding direct estimate, which is usually believed to be more reliable. One way to avoid this problem is the so-called "benchmarking approach," which amounts to modifying these model-based estimates so that one gets the same aggregate estimate for the larger geographical area. This research project will develop a general two-stage Bayesian benchmarking procedure using a single model. With this approach, for example, the state per capita income estimates would be benchmarked to the national per capita income estimate and the corresponding county estimates to the benchmarked state estimates without requiring two separate models. The researcher will develop Bayesian pseudo-empirical likelihood for estimating finite population distribution functions and the corresponding population quantiles. The approach will be extended to estimation of the population distribution function in the small area context where the goal is the same but for which one needs individual estimates for local areas, often involving very small sample sizes. This makes it necessary to "borrow strength" through linking models based on auxiliary information available from censuses or other administrative records. The third component of the research involves inference under informative sampling based on copula models. The copula model to be considered in this project allows dependence in modeling the selection probabilities in terms of the observed outcomes. This is in contrast to the currently available method, which assumes independence in this modeling.

The methods to be developed from this project have multiple applications and will be of value to a broad range of survey researchers. In particular, the research on two-stage benchmarks will be of relevance for many of the Federal statistical agencies. The work on empirical likelihood, with particular emphasis on estimation of small area poverty indicators, is an extremely timely topic. Finally, the research on informative sampling will be of value to researchers across many fields, including epidemiology and economics.

Additional Information:

Title: Fractional Imputation for Incomplete Data Analysis
Proposal ID: 1324922
PI: Jae-Kwang Kim
Institute: Iowa State University
Amount: $249,997
Duration: 36 Months (Estimated)

Abstract
Incomplete data frequently are encountered in survey data due to nonresponse, inaccurate measurement, and two-phase sampling, among other things. Any form of incomplete data can damage the representativeness of a sample, and a naive analysis with incomplete data can lead to biased estimation. Imputation is a process of assigning values to the missing items with the objective of reducing bias and improving the efficiency of the resulting estimators. This project will develop fractional imputation methods as a tool for handling incomplete data for general-purpose estimation. These methods will serve as important building blocks for the establishment of a complete statistical package for analysis of incomplete data that ultimately can be applied to problems in a variety of disciplines, including the social, behavioral, and economic sciences. In particular, the project will develop fractional imputation to address several important problems with incomplete data, including (1) likelihood-based inference, (2) robust estimation using fractional hot deck imputation, (3) synthetic imputation for survey integration, and (4) statistical matching technique.

Because fractional imputation is a relatively new approach for handling incomplete data, there is a critical need for theoretical and methodological development. The advantages of the fractional imputation approach lie in its computational simplicity, wide applicability, and its statistical validity. By using fractional weights, fractional imputation avoids the burden of iterative computation, such as Markov Chain Monte Carlo, for the evaluation of conditional expectation associated with missing data. The proposed approach can be used to estimate parameters consistently and efficiently. The fractional imputation approach can be applied to nonstandard situations such as measurement error models, regression analysis combining two different surveys, and causal inference from observational studies. The impact of the proposed research is therefore substantial because the proposed approach can be used as a general methodology for incomplete data. Because of the computational simplicity and statistical validity of the fractional imputation approach, the results of the proposed research should have wide applicability. It also should have a major impact in providing complete data sets for analysis and new data products combining information from different surveys.

Additional Information:

Title: National Historical Geographic Information System
Proposal ID:1324875
PI: Steven Manson
Institute: University of Minnesota – Twin Cities
Amount: $869,999
Duration: 36 Months (Estimated)

Abstract

The National Historical Geographic Information System (NHGIS) is the nation's most comprehensive source for statistical data, geographic data, and metadata describing spatial characteristics of the American population from 1790 to the present. With 265 billion data points, NHGIS is the largest publicly accessible social science database in the world and is used by thousands of researchers. This project will implement four major improvements to the infrastructure of NHGIS. NHGIS will create data tables for the 1790-1940 decennial censuses from a massive new individual-level database. The project will develop new GIS data identifying locations of incorporated places and MCDs for the 1790-1970 censuses. Census unit boundaries frequently change from one census to the next, making it difficult to analyze population changes over time and space. NHGIS will interpolate census statistics from 1980, 1990, and 2000 and the American Community Survey to produce estimates for 2010 census areas at several geographic levels, including census tracts, places, county subdivisions, and metropolitan areas. Finally, the project will sustain and expand NHGIS education and outreach efforts, offering individual user support and in-person workshops, online tutorials, and web-based community tools.

NHGIS democratizes access to the census, the fundamental source of data about the U.S. population. The proposed improvements will be used for academic research and also for social science training, journalism, policy research at the state and local levels, and private sector research. This infrastructure will allow thousands of investigators in disciplines from economics and ecology to environmental and health policy to address changes across the broad sweep of time at fine levels of spatial organization. NHGIS provides a unique laboratory for the spatial analysis of economic and social processes and offers the empirical foundation needed for developing and testing models of society. Creating new spatiotemporal data for the period 1790 through 1940 will open new opportunities to investigate profound social and economic transformations of American society, including industrialization, immigration, westward migration, and urbanization. Small-area census data are the primary source for investigating such critical issues as suburbanization, the decline and rebirth of central cities, residential segregation, immigrant settlement patterns, rural depopulation, agricultural consolidation, and population shifts from the rust belt to the Sunbelt. The improvements to NHGIS will allow researchers to apply new and powerful approaches to familiar problems by broadening the scope of local and regional analyses to explore variations across time and space simultaneously.

Additional Information:

Title: Reduction of Survey Length through Split Questionnaire Design: Consequences for Nonresponse and Measurement Error
Proposal ID: 1259985
PI: Andrey Peytchev
Institute: Research Triangle Institute
Amount: $580,000
Duration: 36 Months (Estimated)

Abstract

Much research in the social sciences and development of government policy relies on survey data, and the demand for survey data continues to grow. The need for more data has led to longer surveys, increasing the burden for survey respondents in terms of time and effort. Empirical evidence shows a positive correlation between survey length and survey nonresponse, which threatens the representativeness of the survey estimates. There also is evidence that measurement (reporting) error increases as respondents are asked to answer more questions in the survey. Collecting fewer variables may not satisfy a given study's objectives, however. This research project experimentally evaluates the ability to collect all desired data through a split questionnaire design in which respondents are asked only a subset of the questions. The project will use a multiple imputation method to complete the data in the sections that are not asked of particular respondents. The investigators' will extend current imputation methods to include semi-parametric and parametric models. The main hypothesis is that the split questionnaire design approach will yield estimates with less bias and even less total error compared to deploying the full questionnaire.

This project evaluates a method that essentially transfers part of the time and effort to complete the survey from the individual to the researcher. It also evaluates the ability to collect higher quality data as a result of this reduction in respondent burden. Finally, the study aims to extend the employed statistical methods to better preserve the properties of the data. The results will help to provide an alternative methodology for a wide array of surveys, improve split questionnaire design methodology itself, and provide information regarding the circumstances under which implementing such designs can be beneficial.

Additional Information:

Title: Mobile Devices for Survey Data Collection
Proposal ID: 1261340
PI: Roger Tourangeau
Institute: Westat Inc
Amount: $601,922
Duration: 36 Months (Estimated)

Abstract
This project will examine mobile devices, specifically smartphones and tablet computers, as vehicles for survey data collection. The appeal of these devices for survey researchers is clear. Because they are lightweight and relatively inexpensive, they make it easier to collect data using such existing survey modes as computer-assisted personal interviewing. The research will examine three issues raised by use of such devices. First, the input methods that these devices permit (such as touchscreen interfaces) are relatively unfamiliar to many users and may create response problems. Although these interfaces are sometimes used on laptops, tablets and smartphones require them, making usability concerns more central. Second, the screens on tablets and smartphones are considerably smaller than those on laptop or desktop computers. Experiments on web surveys demonstrate the importance of "visual prominence."
Any information that respondents need to use should be immediately visible to them without their having to perform any action (such as a mouse click) to make the information visible. Even the need for an eye movement may effectively render information invisible. Because of the small screens on mobile devices, it may be much harder to make all of the potentially useful information visible to respondents than it is with a laptop or desktop computer. The final issue is the perceived privacy of data collected on these devices. Respondents are willing to reveal sensitive information about themselves when a computer administers the questions, and web surveys seem to retain the advantages of earlier methods of computerized self-administration. But it is unclear whether respondents will display the same level of candor when the survey is administered over the Internet on a tablet computer or a smartphone. Two realistic field experiments and a usability study will examine these issues. Both experiments will be conducted in a single, face-to-face survey. The first experiment will compare laptop computers with tablets and smartphones and will examine the effects of both screen size and input method on breakoffs, missing data, completion times, and indicators of the quality of the responses. The second experiment will compare the same three data collection platforms as vehicles for collecting sensitive information. The experiment will ask respondents to assess the sensitivity of the questions, because item sensitivity may vary as a function of the device used to collect the data.

Surveys are a central tool for social scientists and policymakers in the United States, and survey research is a multi-billion dollar industry in the United States alone. Any set of technological advances, such as the widespread adoption of smartphones and tablet computers, is likely to have a major impact on how surveys are done. Although mobile devices will be widely used for surveys regardless of whether this research is done, the work will produce practical guidelines for using such devices to collect survey data and will alert survey researchers to some of the potential pitfalls of these devices.

Additional Information:

FY 2012

Title: The Validity of Markov Latent Class Analysis for Evaluating Measurement Errors in Complex Panel Surveys
Proposal ID: 1229222
PI: Paul Biemer
PI: Marcus Berzofsky
Amount: $175,000
Total Award Duration: 24 Months (Estimated)

Abstract

Markov latent class analysis (MLCA) comprises a broad class of models and techniques for analyzing categorical longitudinal data subject to misclassification. An important application area is exploring data quality issues in panel surveys. Because MLCA does not rely on gold standard or replicate measurements, it can be applied to virtually any panel survey. For data quality evaluations, MLCA has been used to compare interview modes and alternative questionnaire designs, estimate measurement bias, investigate the causes of misclassification, and investigate many other measurement error issues. Despite its many potential applications in survey work, MLCA has not enjoyed widespread use among survey methodologists because practical guidance on fitting MLC models to complex survey data is lacking. This project will: (1) evaluate the magnitude of the model bias when one or more MLCA assumptions fail when analyzing complex survey data under a wide range of conditions; (2) identify and evaluate the current strategies for diagnosing and repairing MLC model failure and misspecification; (3) address the limitations of current methods by developing improved strategies for diagnosing and repairing MLC model failure and misspecification, particularly in applications to complex surveys; and (4) apply the most effective diagnostic and remedial approaches to real panel survey data to demonstrate the range of modeling issues that can arise in practical applications as well as how to deal with them effectively. As part of the application of these approaches, at least 10 years of data from several national panel surveys will be analyzed to identify temporal trends in measurement error for key national statistics.

This research has important implications for MLCA in all branches of science where classification error is an issue, including social science, epidemiology, clinical research, educational testing, and psychology. The project's impact will be felt in at least four ways. The research has particular relevance for complex survey applications because of the emphasis in this research on modeling cluster-correlated data selected with unequal probabilities and subject to nonresponse and measurement error. It also has important implications for disadvantaged and minority populations whose data may be differentially affected by measurement error. In addition, the evaluation of error trends will provide important information on current and historical levels of measurement error for three important federal statistical programs. Finally, theories regarding the relationship between measurement error and survey participation will be formulated and tested. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Diagnosis and Reduction of Bias in Respondent-driven Sampling
Proposal ID: 1230081
PI: Krista Gile
Institute: University of Massachusetts Amherst
Amount: $199,938
Total Award Duration: 36 Months (Estimated)

Abstract

Respondent-Driven Sampling (RDS) is a type of link-tracing network sampling used to study hard-to-reach human populations. Beginning with a convenience sample, respondents are given uniquely identified coupons to distribute to other population members, making them eligible for enrollment. This is effective at collecting large diverse samples from many hard-to-reach populations. Despite often highly-effective sampling, statistical inference from RDS data is under-developed. Estimates are based on strong assumptions allowing the data to be treated as a probability sample. This project develops methodology for RDS in three ways: (1) Identifying sources of bias and elevated variance in real populations; (2) Extending the network model-assisted inferential framework to address additional sources of bias; and (3) Expanding the network model-assisted framework to arbitrarily large populations. In (1), the research will use simulated RDS samples from fully-observed network data to both explore features of real-world network structures that induce bias or elevated variance in estimates, and also to calibrate diagnostics of such features based on sampled data. Parts (2) and (3) involve advances of the network model-assisted estimator for RDS data. This is the most flexible estimator currently available for RDS data, but is quite computationally burdensome. In (2), the project will extend the existing network model-assisted estimator to adjust for additional features of the network and sampling process. In (3), the project will develop a new variant of this estimator that is more computationally practical in large populations.

Hard-to-reach populations are often at the low end of disparities in wealth, opportunity, and acceptance and at higher risk for negative social and health outcomes. These populations couple reduced social and health opportunities with a difficulty in collecting traditional probability samples. Thus, RDS is widely used and is of special interest to social and behavioral scientists as well as public health officials. This project will allow practitioners to confidently use RDS data in a wider variety of contexts. Specifically, researchers will (a) better understand which network and sampling conditions might induce bias into their estimators and have better tools to find evidence of these features in their samples, (b) have more options for adjusting their estimators to account for non-ideal network and sampling conditions, and (c) be able to compute the resulting estimators with greater efficiency, making the methods practical even in cases of very large target populations. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Interactional Influences on Survey Participation
Proposal ID: 1230069
PI: Nora Schaeffer
Institute: University of Wisconsin-Madison
Amount: $260,596
Total Award Duration: 24 Months (Estimated)

Abstract

This research focuses on the critical challenge of recruiting participants for social research surveys, a problem highlighted by the recent National Research Council?s Panel on the Future of Social Science Data Collection. The refusal component of nonresponse has grown steadily in recent years, threatening the ability of surveys to represent the populations from which the samples are drawn. The proposed research examines the moment-by-moment unfolding reciprocal influence between interviewers and sample members. For example, when an interviewer engages in persuasive actions, is it because she has somehow perceived cues that this sample member is persuadable? When the interviewer engages in actions that appear to be successful in persuading the sample member, is it simply that a persuadable sample member has provided her the opportunity to do so? The research uses new substantive and methodological approaches, drawing on theories of social exchange and reciprocity; conversation analysis and the interaction order; and content analysis. The empirical investigations exploit a unique collection of pairs of acceptances and declinations in a case-control design that uses observations from the Wisconsin Longitudinal Study that are matched on the sample members' propensity to participate. The matching of observations provides for comparison between calls that end in acceptances and declinations of the request for the interview. The project fills a gap in knowledge about whether sample members' voices provide interviewers with information that they can use to rapidly form accurate assessments of the likelihood that a call will lead to an interview. An acoustical analysis of the sample member's initial speec h provides variables such as pitch that can be used with the case-control design to estimate whether these properties of the sample member's voice predict acceptance of the request for participation. By selectively and strategically adding to the existing data transcripts of subsequent contacts and attempts to obtain cooperation, the research also expands previous descriptions of ways interviewers tailor and maintain interaction to describe whether empathic responsiveness and interactional responsiveness (as specific forms of tailoring) make a difference for acceptances of the request to participate.

The intellectual merit and importance of the project are rooted in its contributions to improving the accuracy of social scientific research, which is key to the nation's data infrastructure. The quality of information that describes an entire population can be compromised if participation is restricted to highly selected segments within a sample. This research expands theories of survey participation and their application to the development of practical methods for increasing participation in survey interviews for social research. The broader impacts of the research include identifying specific techniques for improving recruitment to participation in social research, disseminating information about these techniques in refereed journals and presentations at conferences, and training of both graduate and undergraduate students as they participate in the research effort. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: Doctoral Dissertation Research: Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys
Proposal ID: 1238612
PI: Richard Valliant
Institute: University of Michigan Ann Arbor
Amount: $15,153
Total Award Duration: 12 Months

Abstract

Sequential mixed-mode surveys use a mix of modes or data collection methods such as mail, telephone, in-person, and web to increase the number of people who respond to a survey. In sequential designs, there is usually no control in assigning subgroups of respondents to modes. As a result, nonrandom assignment of modes is an inherent characteristic of sequential mixed-mode surveys. This design is important since there are usually limited funds to probe people to respond. While the goal of using mixed modes is clear, one compelling research question is how the nonrandom mix of mode impacts survey data and how these effects should be handled in estimating survey population characteristics such as mean income, and health insurance coverage. To date, since the nonrandom mix of modes poses a challenge in evaluating the mode effects, the existing inference methods assume that mode effects can be ignored in sequential mixed-mode surveys despite their unknown impact on the quality of the survey estimates. This research develops and evaluates the statistical inference methods accounting for nonrandom mode effects to test the comparability of the survey estimates from the different modes. In parallel, this project also develops statistical inference methods accounting for both nonresponse and nonrandom mode effects in the presence of nonignorable mode effects. The public-use Current Population Survey (CPS), 1973, and Social Security Records Exact Match, and the nonpublic-use American Community Survey (ACS) data will be used to conduct empirical and simulation evaluations.

This research provides federal agencies, survey organizations, research centers, and other data producers assessment and inferential methods that adjust for both nonresponse and nonrandom mode effects in the context of sequential mixed-mode surveys. Some large surveys have employed some variation of mixed-mode surveys in order to meet budget constraints. On the other hand, in the presence of nonignorable mode effects, the bias properties for the survey population characteristics are not known and the existing assessment and inferential methods do not control for the nonrandom mode effects. This research produces sequential mixed-mode assessment methods which will test the ignorability of the mode effects which can be a threat for the quality of survey data. In parallel, this research also produces methods of inference which will yield higher quality survey estimates in the presence of nonignorable mode effects.

Additional Information:

FY 2011

Adjusting for Unmeasured Confounding Due to Cluster with Complex Sampling Designs
Proposal ID: 1115618
PI: Babette Brumback
Amount: $ 310,000
Total Award Duration: 36 months

Abstract

In social epidemiology, a geographic neighborhood or cluster is viewed as an important determinant of health behaviors, mediators, and outcomes. One may be interested in the effects of measured or unmeasured neighborhood characteristics or in turn on individual effects that have been disentangled from neighborhood effects. Analyses of nationally representative surveys, such as the National Health Interview Survey, provide a means of estimating these effects. This project will develop statistical methods that can account for the complex sampling design of such surveys at the same time as disentangling individual effects from neighborhood effects. These methods will be applied to analyze data from the National Health Interview Survey. Furthermore, in global health and several other fields, community randomized trials with complex sampling designs are used to estimate the effect of one or more interventions versus a standard or control condition. Scientific interest often focuses on the relative effects of community-level adherence to the intervention on individual-level outcomes. The project also will develop statistical methods that can account for the complex sampling design of such trials while simultaneously extracting the effect of intervention adherence, as opposed to that of intervention intention on individual-level outcomes. These methods will be applied to analyze data from randomized trials designed to study effects of school-level sanitation, water safety, and hygiene on individual education outcomes.

This project involves collaborative research across the fields of biostatistics, social epidemiology, and global health. The research will advance statistical methodology as well as improve the capability of researchers in social epidemiology, global health, and other fields to address important scientific questions. Disentangling individual-level effects from neighborhood-level effects will be useful in understanding the relative roles of the individual versus society and environment in health behaviors and outcomes, which will be useful in designing interventions. Going beyond simple comparisons of treated and untreated individuals in randomized clinical trials and estimating the effects of community-level intervention adherence on individual-level outcomes will further understanding of the effects of interventions in global health and other fields. Illustrating and communicating the new statistical methods for joint estimation of individual-level and neighborhood-level effects on outcomes using nationally representative surveys will provide the federal agencies that sponsor these surveys with enhanced options for creating public-use datasets to facilitate these analyses. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Collaborative Research: Responding to Surveys on Mobile Multimodal Devices
Proposal ID: 1026225
PI: Frederick Conrad
Total Award Duration: 36 months
Amount: $705,410

Proposal ID: 1025645
PI: Michael Schober
Amount: $258,335
Total Award Duration: 36 months

Abstract

Collecting survey data of national importance (for example, on employment, health, and public opinion trends) is becoming more difficult as communication technologies undergo rapid and radical change. Important basic questions about whether and how to adapt data collection methods urgently need to be addressed. This project investigates how survey participation, completion, data quality, and respondent satisfaction are affected when respondents answer survey questions via mobile phones with multimedia capabilities (e.g., iPhones and other "app phones"), which allow alternative modes for answering (voice, text) and can allow respondents to answer questions in a different mode than the one in which they were invited. Two experiments will compare participation, completion, data quality, and satisfaction when the interviewing agent is a live human or a computer and when the medium of communication is voice or text, resulting in four modes: human-voice interviews, human-text interviews, automated-voice interviews, and automated-text interviews. The first experiment randomly assigns respondents to one of these modes; the second experiment allows respondents to choose the mode in which they answer. Results will shed light on whether respondents using these devices agree to participate and answer differently to human and computer-based interviewing agents, and whether this differs for more and less sensitive questions. Results also will shed light on how the effort required to interact with a particular medium (e.g., more effort to enter text than to speak) affects respondents' behavior and experience, and whether the physical environment that respondents are in (a noisy environment, a non-private environment, a brightly lit environment with glare that makes reading a screen difficult) affects their mode choice and the quality of their data. Finally, the results will clarify how allowing respondents to choose their mode of response affects response rates and data quality.

These studies are designed to benefit researchers, survey respondents, and society more broadly. For researchers, the benefit is to allow them to adapt to the mobile revolution as they collect data that are essential for the functioning of modern societies, maintaining high levels of contact and participation while gathering reliable and useful data. For survey respondents, the potential benefit is the design of systems that make it more convenient and pleasant to respond and that enable them to choose ways of responding appropriate to their interactive style, the subject matter, and their physical environment. For society more broadly, it is essential that the survey enterprise is able to continue to gather crucial information that is reliable and does not place undue burden on citizens as their use of communication technology changes and as alternate sources of digital data about people proliferate. More fundamentally, the results will add to basic understanding of how human communication is evolving as people have expanded ability to communicate anytime, anywhere, and in a variety of ways. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information:

Title: "Ensuring Legacy Data Access & Dissemination: Occupational Coding In The General Social Survey"
Proposal ID: 1123510
PI: Michael Hout
PI: Peter V Marsden
Institute: National Opinion Research Center
Amount: $500,000
Total Award Duration: 24 Months

Abstract

In the past thirty years, changes in technology, business, and government practice have substantially altered the American occupational structure. Our project provides a foundation to understand the consequences of new occupations on the current economy and contemporary society, and to preserve unique data key to documenting these fundamental historical changes. Specifically, this project modernizes the occupational and industry data in the General Social Survey (GSS) from the 1970s to the present time.

The project has several goals. They dovetail recent key NSF recommendations that encourage large infrastructure data sources such as the GSS to facilitate increased data access and dissemination. This can be done by presenting data and metadata according to a well-defined protocol, which will allow desirable modes of data access, search, downloads, and documentation. The project also meets the NSF challenge to retrofit historical or legacy data and metadata to become machine readable. This will possibly open up vast amount of data for dissemination and analysis once issues of confidentiality and disclosure are resolved.

To accomplish this goal, this project will (1) retrieve GSS respondents? detailed verbatim descriptions of their work activities, occupations, and industries from the physical questionnaire manuscripts from early GSS waves, (2) convert them into machine-readable form, (3) recode them to reflect 2010 occupation and 2007 industry categories developed by the U.S. Census, and (4) attach external data such as socioeconomic scores and prestige assessments to the recoded categories.

The intellectual merit of digitizing occupational information and recoding occupational and industry categories in the process is that it enables researchers to use the full potential of the occupation and industry information recorded in the GSS over time. Doing so will enhance the value of the GSS as a resource for comparative and contemporary research on social inequality, mobility, and other fields and preserve its growing value as a historical database describing trends in U.S. society over two generations. Ensuring the longevity of such legacy data by converting hand-written text into machine-readable text, the project also develops an archive of verbatim descriptions that will allow future researchers to code them using other standards, including U.S. Census standards that may become available in upcoming decades.

Broader Impacts

The GSS is a public resource as well as a scientific one. Public media, especially newspapers, make extensive use of the GSS. By improving the quality of occupational and industry information in the GSS and ensuring that it is coded in a consistent way over time, this project will help journalists and citizens make sense of social trends and patterns. Also, high schools and colleges make extensive use of the GSS as a teaching tool. Teachers and students will get more out of these exercises from the new data products this project will produce when data reflect contemporary distinctions among occupations and industries as accurately and precisely as possible.

Additional Information:

Title: Collaborative Research: Best Predictive Small Area Estimation
Proposal ID: 1121794
PI: Jiming Jiang
Amount: $69,462
Total Award Duration: 36 months

Proposal ID: 1122399
PI: Jonnagadda S. Rao
Institute: University of Miami School of Medicine
Amount: $78,632
Total Award Duration: 36 months

Proposal ID: 1118469
PI: Thuan Nguyen
Institute: Oregon Health and Science University
Amount: $85,800
Total Award Duration: 36 months

Abstract
Surveys usually are designed to produce reliable estimates of various characteristics of interest for large geographic areas or socio-economic domains. However, for effective planning of health, social, and other services and for apportioning government funds, there has been a growing demand to produce similar estimates for small geographic areas and subpopulations, commonly referred to as small areas. This research project aims at developing a new method of small area estimation that potentially will lead to a dramatic improvement in accuracy over the traditional methods in practical situations. Model-based small area estimation utilizes statistical models, such as mixed effects models, to "borrow strength." In particular, the empirical best linear unbiased prediction (EBLUP) is a well-known model-based method that has had dominant influence in small area estimation. From a practical point of view, however, any proposed model is subject to model misspecification. When the proposed statistical model is incorrect, EBLUP is no longer efficient or even effective. In such cases, a new method, known as observed best prediction (OBP), may be superior. This project involves several important research topics on OBP, including theoretical developments, assessment of uncertainties under weak model assumptions, and implementation of the OBP via user-friendly software. The research largely will expand the results of our earlier studies, and contribute to making the OBP method more effective, practical, and easy to use.

The research introduces a completely new idea and method to model-based statistical methods in survey sampling. It is expected to impact other scientific areas where statistical methods have been used for prediction problems. The project will develop and freely disseminate R code to implement the OBP method. The education component of the project will introduce the OBP method into courses at the investigators' institutes. These courses are expected to draw students and researchers from statistics, biostatistics, genetic epidemiology, animal and plant sciences, educational research, social sciences, and government agencies. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Additional Information: