Statistical Challenges in Combining Information from Big and Small Data Sources

Trivellore Raghunathan *
Survey Research Center,
Institute for Social Research
University of Michigan
Ann Arbor, 48106

Abstract

Social Media, electronic health records, credit card transactional and administrative data, web scraping, and numerous other ways of collecting information have changed the landscape for those interested in addressing policy-relevant research questions. During the same time, the traditional sources of data, such as large-scale surveys, that have been a stable source for policy-relevant research have suffered setbacks due to large nonresponse and increasing data collection costs. The non-survey data usually contain detailed information on certain behaviors on a large number of individuals (such as all credit card transactions) but very little background information on them (such as important covariates to address the policy-relevant question). On the other hand, the survey data contains detailed information on covariates but not so detailed information on the behaviors. Both data sources may not be perfect for the target population of interest. This paper develops and evaluates a framework for linking information from multiple imperfect data sources along with the Census data to draw statistical inference. An explicit modeling framework involving selection into the big data, sampling and nonresponse mechanism in

*Paper Presented at the Expert Panel Meeting at the National Academy of Sciences, November 19, 2015. Author email address: teraghu@umich.edu

the survey data, distribution of the key variables of interest and certain marginal distributions from the Census Data are used as building blocks to draw inference about the population quantity of interest.

1 Introduction

The digital revolution, though at least 50 years old, is coming to fruition now due to, in large part, ever increasing computational infrastructure and inexpensive storage. Social media, computerized or electronic records and many other digitized archives have changed the landscape of data. Statisticians have witnessed such changes in the landscape in the recent past. The advent of powerful desktop and server machines in the late eighties and early nineties made it possible to fit many statistical models that were impractical to implement a few years earlier and many old algorithms such as Metropolis (Metropolis et al (1953)), Hastings-Metropolis (Hastings (1970)), Gibbs sampling (Geman and Geman (1984)) and other Markov Chain Monte Carlo methods were no more theoretical exercises or relegated to main frame computers but became a common practice, so much so that, complex statistical model building has become quite routine.

The statisticians are at the cusp of the next stage of revolution where data from many sources and in many forms are becoming available and beckoning them to rise up to the challenge of integrating these data sources to construct inference about the population, their primary goal. The new challenge also includes the art and science of processing huge data sets that are not necessarily in the familiar rectangular format with rows for subjects and columns for variables.

During the same time period, the probability sample surveys, the traditional bread-and-butter tool for researchers has been facing challenges due to declining response rates. Many surveys conducted by survey research firms with tremendous perseverance and costs range between 40% to 50% and some telephone surveys much less. The government surveys are still eliciting larger response rates but at the enormous cost of nonresponse follow-up. Public fatigue, privacy and confidentiality concerns and costs will continue to affect the surveys. Hence, the surveys are relying more and more on post-survey adjustments using scant variables available on respondents and nonrespondents.

The real task for the statistical community is to face the challenge of

declining response rates and the rising costs of conducting surveys with an increasing opportunity afforded by non-survey data sources without deviating from the principal objective: representative or "valid" inference about the target population of interest. There is a need for discovering a new set of tools or reshaping the old tools to leverage these two kinds of data sources. This can be done through refining the design of surveys and statistical models for combining information from multiple sources.

The goal of this paper is to lay out certain statistical framework for combining information from multiple data sources using the statistical modeling and imputation framework. Clearly lay out the assumptions needed to pool information from multiple sources and use those assumptions to construct "synthetic" or "plausible" data sets representative of the target population interest. This will enable the research community to broaden the scope of questions that can be asked and answered.

2 Big Data versus Survey Data

Declining response rates and increasing costs of traditional surveys and the advent of big data may tempt us to consider big data as the primary (or the only?) source for inferring about the population. To delve into the consequence of this possibility, consider the problem of estimating the prevalence rate, θ, of a certain attribute. Define a binary variable where $X = 1$ is for subjects with the attribute and $X = 0$, otherwise. A simple random sample survey of size n_S results in an estimate $\widehat{\theta}_S$, the sample proportion. The sampling variance of this estimate is $\theta(1-\theta)/n_S$. For now assume that there is no nonresponse.

Suppose that the same variable is captured in a non-survey data of size n_A, resulting in an estimate $\widehat{\theta}_A$, the proportion computed based the elements in the non-survey data. Suppose that $A = 1$ denotes that the person is captured in the non-survey data. Generally no information is available for the subjects not captured in the non-survey data. Nevertheless, let $Pr(A = 1|X = 0) = \pi$ and $Pr(A = 1|X = 1) = \rho\pi$ be the respective probabilities of capturing persons without and with the attribute. That is ρ is the rate of capturing a person with the attribute in the non-survey data relative to those without the attribute.

Suppose that we apply the same binomial model. Note that this a subjective model without the probability sampling framework as in the case of

$\widehat{\theta}_S$. Some of the early references where such models were considered for non-probability samples are Smith (1983), Rubin (1987) and Deville (1991). The basic idea is to model the selection as a function of outcome and covariates and then lay out the conditions under which the observed sample can be used to project or predict the nonsampled part of the population. The response propensity models are examples of such subjective probability models that allows for post-survey adjustments (Little (1982)).

It follows that $Pr(X = 1|A = 1) = \theta\rho\pi/(\theta\rho\pi + (1-\theta)\pi) = \theta\rho/(\theta\rho + (1-\theta))$. The bias in the estimate $\widehat{\theta}_A$ is $-\theta(1-\rho)(1-\theta)/[1-(1-\rho)\theta]$. Thus the mean square error of $\widehat{\theta}_A$, under the assumed binomial model, is

$$MSE(\widehat{\theta}_A) = \frac{\theta(1-\theta)}{n_A}\frac{\rho + n_A\theta(1-\theta)(1-\rho)^2}{(1-(1-\rho)\theta)^2}$$

The relative efficiency of the estimate from the non-survey data relative to the random sample estimate is

$$RE_{A|S} = \frac{n_A(1-(1-\rho)\theta)^2}{n_S(\rho + n_A\theta(1-\theta)(1-\rho)^2)}.$$

Note that, this relative efficiency is not always greater than one even is n_A is very large compared to n_S. Let n_A be very large relative to n_S and the above equation simplifies to,

$$\frac{(1-(1-\rho)\theta)^2}{n_S\theta(1-\theta)(1-\rho)^2}.$$

An interesting question is when does the estimate from the big data become less efficient than the survey data. It can be shown the above equation is less than 1 for ($n_A >> n_S$), when

$$n_S \geq \frac{(1-(1-\rho)\theta)^2}{\theta(1-\theta)(1-\rho)^2}.$$

To get some perspective, suppose that $\rho = 1.2$ (that is, people with the attribute are 20% more likely to be captured in the non-survey data than those without the attribute) and the true prevalence rate is $\theta = 0.1$, then the non-survey data is less efficient whenever $n_S \geq 289$. Suppose that $\rho = 1.05$ and for the same θ, the threshold simple random size is $n_S \geq 4,489$. That is, the squared bias term tends to dominate even with the modest differential

inclusion probabilities with respect to the outcome of interest in the non-survey data. It is not hard to imagine some differential inclusion probabilities related to the outcome of interest when the non-survey data are constructed for special purposes (Marketing companies, particular banks etc).

Of course, the real surveys rarely employ simple random sample design but typically involve unequal probabilities of selection, stratification and clustering. Thus, n_S could be interpreted as effective sample size adjusted for design effect.

The simple analysis suggests that selection bias can have a big impact on the inferences from the non-survey data and could not be even checked without having a reliable survey or some external data to check against or to calibrate. That is, "big data" need to be free of coverage errors, especially for government statistics, academic research where the generalizability of the results is the norm.

On the other hand, if an estimate, $\widehat{\rho}$, of ρ were available (say, based on a substudy: a small carefully designed survey or experiment) then one could construct a bias corrected estimate, $\tilde{\theta}_A$, by equating

$$\widehat{\theta}_A = \frac{\widehat{\rho}\tilde{\theta}_A}{1-(1-\widehat{\rho})\tilde{\theta}_A},$$

yielding,

$$\tilde{\theta}_A = \frac{\widehat{\theta}_A}{\widehat{\rho}+(1-\widehat{\rho})\widehat{\theta}_A}.$$

A pooled estimate combining the survey and non-survey data can be derived as

$$\widehat{\theta} = (v_S^{-1}+v_A^{-1})^{-1}(\widehat{\theta}_S/v_S + \tilde{\theta}_A/v_A)$$

where $v_S = \widehat{\theta}_S(1-\widehat{\theta}_S)/n_S$ and $v_A = \tilde{\theta}_A(1-\tilde{\theta}_A)/n_A$. An implicit Bayesian model is to treat $\theta|A \sim N(\tilde{\theta}_A, v_A)$ as the prior distribution and $\widehat{\theta}_S|\theta \sim N(\theta, v_S)$ as the sampling distribution.

3 Strategies for Estimating Selection Bias

It is critically important to assess and estimate the selection bias term ρ. Fortunately, the modeling framework provides for laying out the assumptions and some approaches for estimating the selection bias. Suppose that Z is a covariate with k categories such that $Pr(A=1|X=1,Z=j) = Pr(A=$

$1|X = 0, Z = j) = Pr(A = 1|Z = j)$, $j = 1, 2, \ldots, k$. This is akin to missing at random assumption in the missing data framework (Rubin (1976)) conditional on Z.

Note that

$$Pr(A = 1|X = 1) = \sum_j Pr(A = 1|Z = j)Pr(Z = j|X = 1)$$

and

$$pr(A = 1|X = 0) = \sum_j Pr(A = 1|Z = j)Pr(Z = j|X = 0)$$

. Writing

$$Pr(A = 1|Z = j) = Pr(Z = j|A = 1)Pr(A = 1)/Pr(Z = j),$$

we obtain

$$\rho = \frac{Pr(A = 1|X = 1)}{Pr(A = 1|X = 0)} = \frac{\sum_j Pr(Z = j|X = 1)Pr(Z = j|A = 1)/Pr(Z = j)}{\sum_j Pr(Z = j|X = 0)Pr(Z = j|A = 1)/Pr(Z = j)}$$

Thus, to implement this method we need estimates of the marginal and various conditional distributions of the covariate, Z.

- From the non-survey data we need estimates of $Pr(Z = j|A = 1)$.
- The Census or the population data may provide $Pr(Z = j)$
- A sample survey or a pilot study may provide $Pr(Z = j|X = l), l = 0, 1$.

The categorical nature of the covariates makes these building blocks as aggregate data that producers of non-survey data may be able to provide without violating privacy and confidentiality. For example, if the non-survey data source is a bank, for example, and Z is the categories of total "volume", then the bank may be able to provide the marginal distribution of based on its customers.

What are some of the options for constructing Z? Suppose that the non-survey and survey data have some common covariates U. Suppose that $\widehat{\beta}_S$ is the estimated regression coefficient, in a logistic regression model with X as the dependent variable and U as independent variables, obtained

from the survey data. Let $Z = [1 + \exp(-U^t \widehat{\beta}_S)]^{-1}$ be the predicted probability. The same regression coefficient, $\widehat{\beta}_S$, is then applied to the non-survey data to construct Z. That is, Z is the (counterfactual) prediction of X for the subjects in the non-survey data that would have been obtained had they been in the survey data. The underlying assumption is that conditional on having the same prediction under the survey data, the actual attribute status is not related to the selection into non-survey data. The predicted variable, Z, can be categorized to create classes.

The second approach is to use some common variables between the non-survey and the sample frame data sets. Some examples of such variables are block or block group characteristics. Suppose that $S = 1$ indicates a sampled subject and $S = 0$ indicates a non-sampled subject. Let U be the frame variables also available in the non-survey data (or can be attached to non-survey data). Let $\widehat{\beta}_S$ denote the regression coefficient from the logistic regression model predicting S from U. Apply this estimated regression coefficient to the non-survey data. This covariate represents the likelihood of subjects in the non-survey data for being predicted to be in the sample. Again, this covariate could be categorized to form classes.

A final example of a strategy for constructing the covariate Z is the propensity of being in the survey data. Specifically, append the non-survey and survey data and define $D = 1$ for the survey subjects and $D = 0$ for the non-survey subjects. Estimate the propensity score by using a logistic regression model with D as the dependent variable and all the common covariates in the two data sets. The categories can be created based on the propensity score. The rationale underlying this strategy is that if the subject in the non-survey data matches to subject in the survey data then the labeling of subjects as survey/non-survey is completely at random.This strategy was used to correct for discrepancies between the self-report and clinical measures of chronic conditions such as hypertension, diabetes etc by pooling data from the National Health Interview Survey and National Health and Nutrition Examination Survey as described in Schenker, Raghunathan and Bondarenko (2010).

The central theme of all these approaches is to balance or match the non-survey data with the survey or population data through propensity scoring. Within the matched sets, selection bias is assumed to be non-existent or at least negligible. Note that, the bias corrected non-survey data estimate will have very small mean square error relative to survey based estimate (if the bias correction is successful). Thus, the the survey goal could be just

to provide enough data to permit bias-correction. A smaller scale survey with high response rate could possibly be mounted with lower cost and thus leveraging the information in the larger non-survey data.

Obviously, the survey data is subject to nonresponse but several studies (Groves et al (2010)) have shown that even with high nonresponse rate, the survey estimates suffer from lower nonresponse bias, if at all. This may be due to chances of completely missing a significant section of the population in a probability sample survey is quite rare. For example, in the big data from Apple's Research Kit (or Twitter or Facebook or any other data source) will have only iphone users (Twitter or Facebook or some specific attributes). However, in probability survey even with higher propensity of response for the iphone users compared to non-iphone users, a few non-iphone users might participate and provide relevant data. Using auxiliary variables collected on respondents and nonrespondents (through proper planning at the design stage and collected during the conduct of the survey), post-stratification techniques one can derive unbiased estimates from the biased survey data. For an interesting example, see Wei et al (2015).

Some sources of auxiliary variables include interviewer observations, contextual or geographical data estimated from a variety of sources, commercial data etc. To some extent, the survey world did not creatively plan the collection and the use of auxiliary variables with an anticipation of the steep decline in the response rates. One of the reasons is that survey inference, as a field, was less embracing towards the use of statistical modeling in the inferential activities where as the non-survey inference world fully embraced and exploited the modern statistical modeling and computational advances to its great advantage. The quote "All models are wrong and some are useful", attributed to George Box, a famous statistician summarizes the attitude needed: Carefully craft the model that captures the important features of the data being analyzed, perform proper diagnostics to assess the model fit and then proceed with the inference about the population, fully incorporating the uncertainties in the non-observed data conditional on the model. An assessment of sensitivity of the inferences to the model assumptions needs to be a standard feature in all inferential activities.

The design-based inference paradigm adopted the notion that all models are wrong, therefore, no model should be used. Instead, it made numerous "algorithmic" assumptions (such hot-deck, cold-deck, editing rules, pooling strata, combine PSUs etc) without any framework for checking these assumptions. However, the model assumptions were adopted in some cases such as

small area estimation. This schizophrenic application of statistical thinking needs to change.

4 Going Beyond Each Through Combining

Consider a situation were a data source A provides variables (U, X, Y), the data source B provides (U, X, Z) and data source C provides (U, Y, Z). If the data sources A, B and C are representative of the same population then vertically appending the data creates a traditional missing data problem (missing Z in the data set A, missing Y in the data set B and missing X in the data set C). Existing technology such as multiple imputation can be applied to create completed data sets that allows joint analysis of (U, X, Y, Z). Note that, such leveraging extends the utility of each data source beyond what it was intended to be. This strategy could be used by the data repositories, Federal agencies to use the variety of data already collected, harmonize the variables and link it spatially and temporally.

An example of one such project is to consider the 1940 census which is now available electronically to create a cohort of individuals and then try to link (deterministic or probabilistic) first to all available digitized information such as the Current Population Surveys, American Community Surveys, various other surveys, Administrative records, mortality files etc. This requires a concerted efforts working across agencies within the confines of secured environment, such as Census Bureau Research Data Center. This first stage effort will provide a data set with considerable holes (missing information). The investigation of missing portions will then lead to sampling of non-digitized records such as later year census data for digitization and incorporation into the data set.

Obviously, the cohort formed from the 1940 census is not a representative for the later years. Thus, sub-sampling and digitization of subjects in the later census years and and attaching available survey data to them will improve the representativeness and provide better temporal picture. Once all reasonable efforts have been made to fill-in as much information as possible through deterministic or probabilistic linking then one can adopt a statistical approach for multiply imputing the missing portions of the data set. Thus creating a retrospective observation based longitudinal data entirely by leveraging the existing data resources.

The goal is not to create an actual data set, but a plausible data set

that matches the population in various respects. Just like an imputed data for any one survey is not an actual data set but a plausible data set. The reasonableness of such a data set can be assessed by comparing the inferences from this data set to the inferences from the actual data set for a given time period and given set of variables. For example, one can check whether the plausible data set so constructed yield descriptive and analytical inferences for, say the year 1990, yield similar to the one based on, say 1990 long form. Such calibration of the plausible data increases the confidence in the inferences constructed from it.

Returning the example with three data sources, A, B and C, suppose that each one them may be subject to selection bias. Usually, the selection bias is not be known. The unknown information are the conditional distributions, $[Z|U, X, Y, A]$, $[Y|U, X, Z, B]$ and $[X|U, Y, Z, C]$.

Suppose that a small representative survey is conducted to collect data D, on (U, X, Y, Z), appropriately weighted and imputed for missing values using the design variables, paradata and other auxiliary variables. The goal is not make this survey a primary vehicle for drawing inference about the population but enough to estimate the quantities needed to leverage the large data sets A, B and C.

The following strategy could be used to achieve our goal of creating a plausible data set from the population:

1. Append all four data sets (vertically concatenate). Create a categorical variable V with three levels, with $V = 1$ for data A, $V = 2$ for data B and $V = 3$ for data C. Set V to be missing for all subjects in the data set D. When this variable is imputed the observed data is being used allocate subjects in the data set D to one of the three data sources.

2. Impute the missing values in Z for data A by applying the restriction that model be fit and predicted values be generated by sub setting the data with $V = 1$.

3. Impute the missing values in Y in the data B by restricting model fit and imputation to $V = 2$

4. Impute the missing values in Z in the data set C by restricting the model fit and imputation to $V = 3$.

5. The final step is assign weights to subjects in the data sets A, B and C commensurate with their representation in the population. For ex-

ample, the post-stratification based on the population characteristics (for example, the census data or estimated from large surveys such as the American Community Survey, the Current Population Survey or the National Health Interview Survey). The second option is to use the imputed data set D to estimate the representation of the subjects like those in A, B and C. Suppose that p_A, p_B and p_C be the weighted estimated of proportion for categorical variable V in the data set D. Let m_A, m_B and m_C be the sizes of data sets A, B and C, respectively with $m = m_A + m_B + m_C$. Assign each subject in the data set A the weight of m_A/mp_A. Similarly, m_B/mp_B and m_C/mp_C for the data sets B and C, respectively. All subjects in the data set D receives the original survey weight.

There are many refinements of the procedure described above. For example, suppose that conditional distributions derived from data set A do not match the conditional distributions in the data sets D. This implies certain level of uncertainty in the actual population distribution. Imputation approach can incorporate these uncertainties by refining the model assumptions or by creating imputations under different model assumptions. Thus, the modeling principles provide a concrete infrastructure for leveraging data from multiple sources.

To incorporate the uncertainty in the imputations, the above steps can be repeated several times to create a set of multiply imputed plausible data sets. Standard multiple imputation combining rules (Rubin (1987), Little and Rubin (2002), Raghunathan (2015)) can be applied to create inferences. Of course, this strategy extends to many variables with arbitrary pattern of missing data and more than three data sources that can be pooled to create large plausible data set from the population adjusted for selection bias.

One of the ongoing project involves creating an infrastructure to develop an understanding of relationship between demographic and socio-economic factors (X), health conditions (D) and medical expenditures (E). Each of these variables are multivariate. Unfortunately, there is no single data source that provides comprehensive information on all three domains for the entire population. However, there are several data sets measuring a subset of these domains. For example, the Medicare Current Beneficiary Survey, National Health Interview Survey, National Health and Nutrition Examination Survey, Health and Retirement Study, Medical Expenditure Panel Survey, National Comorbidity Survey etc are some of the representative surveys provide data

in some of these dimensions. Through calibration, post-stratification and imputation plausible data set is being created for four age segments of the population: Age 65 and above, 45 to 64, 18 to 44 and under 18 years of age. The work has been completed for Age 65 and above for the period 1999-2009, primarily using MCBS, NHANES and CMS claims as data sources (Cutler et al (2015)).

5 Discussion

Combining survey and non-survey data sources provides unique opportunities to extend the usefulness of each data source and pose challenges in terms of the methodology to be used to harness information from these sources. The declining response rates in sample surveys and potential selection bias in the non-survey data sources makes the task as that of pooling information from imperfect sources.

Even with low response rate surveys, through auxiliary variables and post-stratification, it is possible to adjust for bias and by reducing the sample size, more efforts can be devoted for increasing the response rate or reducing the nonresponse bias. This smaller high quality survey can then be used to correct for potential selection bias in the non-survey data.

The central theme of this paper is that task of combining information from multiple imperfect data sources can be accomplished through proper development of statistical models with reasonable assumptions that be directly or indirectly tested or validated. The current missing data framework, modeling and software can be modified to achieve this goal. Some simple examples given in this paper are just for kindling the imagination for this line of research to be undertaken by the scientific community.

There are several limitations. The data sources could be collected under different contexts, some are self-reports and others could be record based. It is possible that some data were collected on web, some on telephone, some through mail and some through in-person interview. The mode differences may make the measurement not comparable. There may design differences across the surveys being pooled.

Non-survey data may also differ in important ways. For example, Twitter, Facebook caters to different audiences. Privacy concerns of people who use these social media sites may be different from those who do not. Thus any variable highly patterned by the privacy concern is subject to biased

estimation. It is important to understand the purpose for which the data are collected and the context in which the data is provided.

All these limitations are challenges that require thoughtful small scale experiments and incorporation of results through modeling. There is no doubt that the landscape for the data analysis has changed and will continue to change. This reality should propel us to think creative ways to harness the information from survey and non-survey data sources.

References

[1] Deville, J. (1991), A theory of quota surveys, *Survey methodology*, 17, 163-181.

[2] Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayeisan restoration of images, *IEEE Transactions on Patern Analysis and Machine Intelligence*, 6, 721-741.

[3] Groves, R. M. (2008). Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, 70, 646-675.

[4] Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

[5] Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C, Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.L. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys. *Journal of Royal Statistical Society (Series A)*, 173, 389-407.

[6] Little, R. J. A. (1982). Models for Nonresponse in Surveys, *Journal of American Statistical Association*, 77, 237-250.

[7] Little, R. J. A. and Rubin, D. B. (2002). **Statistical Analysis with Missing Data**, New York: Wiley.

[8] Metropolis, N. Rosenbkuth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast compouting machines. *Journal of Chemical Physics*, 21, 1087-1092.

[9] Raghunathan, T. E. (2015). **Missing Data Analysis in Practice**. CRC Press: Boca Raton, Florida.

[10] Rubin, D. B. (1987). **Multiple Imputation for Nonresponse in Surveys**, Wiley: New York.

[11] Schenker, N. and Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health, *Statistics in Medicine*, 26, 1802-1811.

[12] Schenker, N., Raghunathan, T. E., and Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey, *Statistics in Medicine*, 29, 533-545.

[13] Smith T. F. M. (1983). On the validity of inferences from Non-random samples, *Journal of the Royal Statistical Society*, 146, 394-403.

[14] Wei, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections from nonrepresentative polls, *International Journal of Forecasting*, 31, 980-991.