Challenges Faced In The Daily Modeling Of Survey Responses

Reid Rottach, Mahdi Sundukchi, Norilsa Toribio
US Census Bureau, 4700 Silver Hill Road, Washington DC 20233

Abstract

Recently, a system was developed at the US Census Bureau to upload each day's Computer Assisted Personal Interviewing (CAPI) data from interviewer laptops and analyze the results using regression models. The intent was to track the possible effect of changes in management at the bureau's Regional Offices on survey responses to key variables. We discuss this application of daily modeling in light of lessons learned from a completed project. Furthermore, we discuss a new system being developed to track components of survey error across sample design changes in the 2010 Redesign of demographic surveys.

Key Words: statistical modeling, total survey error, exploratory analysis, categorical data analysis, generalized regression

1. Introduction

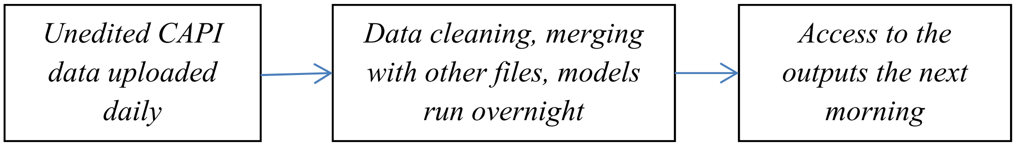
In 2011, the Census Bureau began developing measures to study demographic survey data as it was uploaded from interviewer laptops on a daily basis. The purpose of this work was to track responses during a realignment of Census Bureau Regional Offices (ROs) that occurred in 2012. The surveys in scope were the American Community Survey (ACS), the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), the National Crime Victimization Survey (NCVS), the National Health Interview Survey (NHIS), and the Consumer Expenditure Quarterly (CEQ) Interview Survey.

In general, the intent of modeling survey outcomes daily is to follow trends across the interview period for a more complete picture of what's occurring in the field, and be able to anticipate changes early on. Some issues that this type of study could shed light on are an especially low response rate, estimates that spike in a given period, a change in the level of precision in an estimate, or sample sizes that are higher or lower than expected. The data used for the applications described in this paper comes from Computer Assisted Personal Interviewing (CAPI). CAPI refers to the instrument used for interviewing rather

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

than necessarily the mode, so in some cases the interviews may be conducted by telephone. The CAPI data is loaded onto Census Bureau servers nightly, which makes analysis on a small time scale feasible.

In our applications, the statistical modeling is embedded in a system where the data from one day's input is used in outputs that are available the following morning. One of the significant constraints of this type of analysis is the necessity for everything to be automated from the time the data streams in until the time outputs are viewed.


Unedited CAPI data uploaded daily
Data cleaning, merging with other files, models run overnight
Access to the outputs the next morning

There are two applications discussed in this paper: the first was a study of the effect of the realignment of Census Regional offices on survey estimates, and the second is a study of the effect of recent changes that are part of the 2010 Sample Redesign of our demographic surveys.

2. Application to the Study of the RO Realignment

In 2012, the number of ROs was reduced from twelve to six. The realignment involved a change in the supervisory structure, which could possibly have led to changes in workloads for the Field Representatives (FRs), for example. There was concern that the changes might have an effect on survey outcomes, so the purpose of the study was to measure the effect of the realignment on our estimates.

Under the new supervisory structure, the country is divided into 48 geographic areas, each one managed by a Survey Statistician in the Field (SSF). Each of the six ROs have eight SSFs assigned. The realignment occurred in seven waves across 2012; the first wave was in January, and the seventh was in November. Each of the ROs that remained at the end of the year had one new SSF assigned at the start of each wave, except in the seventh wave when the last two were assigned. Figure 1 below is a representation of this phase-in.

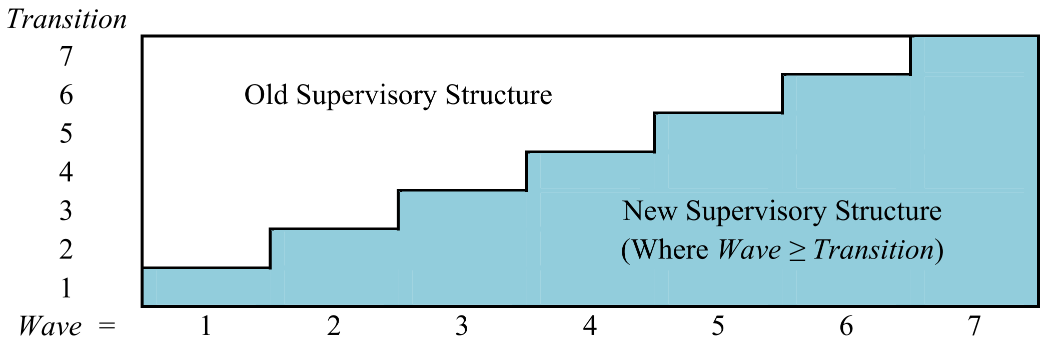

Transition
7
6
5
4
3
2
1
Wave =
1
2
3
4
5
6
7
Old Supervisory Structure
New Supervisory Structure
(Where Wave ≥ Transition)

Figure 1: Representation of the transition to the new supervisory structure across seven waves. The variable *Transition* is a geographic variable representing the wave at which that region transitioned to the new supervisory structure.

The table in Figure 1 inspires an approach to modeling the survey variables, with two main effects of *Transition* and *Wave*. As an example, the tables that follow show standardized residuals from loglinear models of the proportion of respondents (Figure 2) and unemployed (Figure 3) from the CPS. The loglinear models have the form:

$$log(m_{ij}) = \mu + \alpha_i + \delta_j + e_{ij} \tag{1}$$

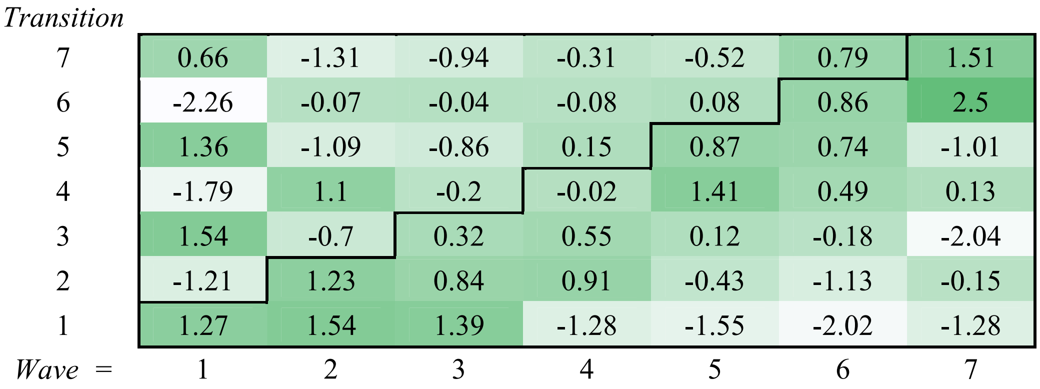
Where m_{ij} is the cell count of the characteristic, μ is the overall effect, α_i is the effect of *Wave* i, and δ_j is the effect of *Transition* j

The residuals in the tables are standardized by the square root of the model prediction. The sum of squares of the standardized residuals is the Pearson Chi-squared statistic that may be used to test independence (Agresti, 1990). The standardized residuals are:

$$e'_{ij} = e_{ij}/\sqrt{\mu + \alpha_i + \delta_j} \tag{2}$$

Transition							
7	2.21	1.05	0.12	-1.15	-0.5	-2.37	-1.03
6	-2.89	-0.62	1.13	0.31	0.5	1.44	1.42
5	3.39	0.15	-1.69	-1.45	-0.29	-0.5	-1.02
4	1.41	1.92	-2.1	-1.49	-0.97	-0.61	0.62
3	-1.07	-0.53	1.26	0.77	-0.55	1.22	-0.45
2	-1.06	-0.21	0.35	1.46	0.32	0.26	-0.28
1	-3.43	-2.29	0.94	2.13	1.85	1.7	1.52
Wave =	1	2	3	4	5	6	7

Figure 2: Standardized residuals for the distribution of total respondents



Transition							
7	0.66	-1.31	-0.94	-0.31	-0.52	0.79	1.51
6	-2.26	-0.07	-0.04	-0.08	0.08	0.86	2.5
5	1.36	-1.09	-0.86	0.15	0.87	0.74	-1.01
4	-1.79	1.1	-0.2	-0.02	1.41	0.49	0.13
3	1.54	-0.7	0.32	0.55	0.12	-0.18	-2.04
2	-1.21	1.23	0.84	0.91	-0.43	-1.13	-0.15
1	1.27	1.54	1.39	-1.28	-1.55	-2.02	-1.28
Wave =	1	2	3	4	5	6	7

Figure 3: Standardized residuals for the distribution of respondents coded as unemployed

The impact of the realignment is an interaction effect that was not included in the model. The study of a possible effect using these tables is exploratory, and motivates the use of graphical aids, such as differentiating between the old supervisory structure and the new (the dark line along the diagonal) and shading cells relative to the size of the standardized residual. This approach is discussed in light of lessons learned following the actual study in 2012, which involved a more complex model. The possible effect of the realignment had been studied using hypothesis tests. In particular, the approach was to examine the significance of a level shift following the assignment of each new SSF. Most models were of the form:

$$logit(p_{ijk}) = \mu + \alpha_i + \delta_j + \gamma I_{\{i \geq j\}} + X_{ijk}\beta_{ijk} \tag{3}$$

Where p_{ijk} is the expected value of the binary response variable, μ is the overall effect, α_i is the effect of *Wave i*, δ_j is the effect of *Transition j*, $\gamma I_{\{i \geq j\}}$ is the management effect, with $I_{\{.\}}$ being an indicator function, and the remaining terms, $X_{ijk}\beta_{ijk}$, are to control for possible confounding variables

The significance tests were performed on the coefficient γ. When these coefficients became significant, it often led to requests for exploratory analyses of the data. In the sample redesign application, described in the next section, the focus will be on exploratory analyses rather than hypothesis testing.

A further complication to daily modeling was the changing sample sizes from day to day. Early in the year, there were problems with overfitting which led in many cases to absurd regression coefficients. Penalized regression may have helped, but it was found that these procedures took far more computing time, and likely would not have been feasible due to the number of variables that were tracked. Another option would have been to distribute cases from 2011 into the regression, randomly assigning a “1” to the new management indicator. In this case, on the first day of 2012, the coefficient would have had expected value zero and a wide confidence interval.

3. Application to the 2010 Sample Redesign

The 2010 Redesign of demographic surveys conducted by the Census Bureau had many substantial changes. Historically, the surveys sampled every ten years following each Decennial Census. The sample was supplemented with listing operations in an area frame, and a sample of new housing units using updated lists of building permits. In the new design, sample will be selected each year from the most recently updated Master Address File (MAF). The new frame allows for new methods, including annual sampling rather than every ten years. See Nguyen and Gerstein (2011) for an overview of the 2010 Redesign.

The upcoming study using daily CAPI data is not completely determined at this point, but two uses of the data are being developed. The first is to track key variables, including the number of respondents with different outcome codes, using categorical data analysis of the transition from the old design to the new. The second is to track response propensities within demographic groups.

3.1 Modeling the Transition to the New Sample Design

The SIPP will be the first survey to introduce 2010 Redesign samples, in January 2014. One of the many changes in SIPP’s design is a change in the survey instrument (Walsh and Fields, 2011). This complication leads us to plan the SIPP study shortly after the start of the new design, so the discussion in this paper will focus on the second survey to introduce 2010 sample, which is the CPS.

With the CPS, the phase-out of 2000 Redesign sample and phase-in of 2010 Redesign sample occurs across 16 months beginning in April, 2014. Figure 4 is a representation of the CPS rotation chart during the phase-in/phase-out (PIPO) period. In any given month, there are eight panels interviewed for the survey, which can be seen in the chart by noting that there are eight shaded cells in any row. Due to the rotation pattern of the CPS, in any given month there will be one panel having its first interview, another having its second, another having its third, and so on. The index used to identify the interview count is the *month-in-sample*, or MIS (US Census Bureau, 2006).

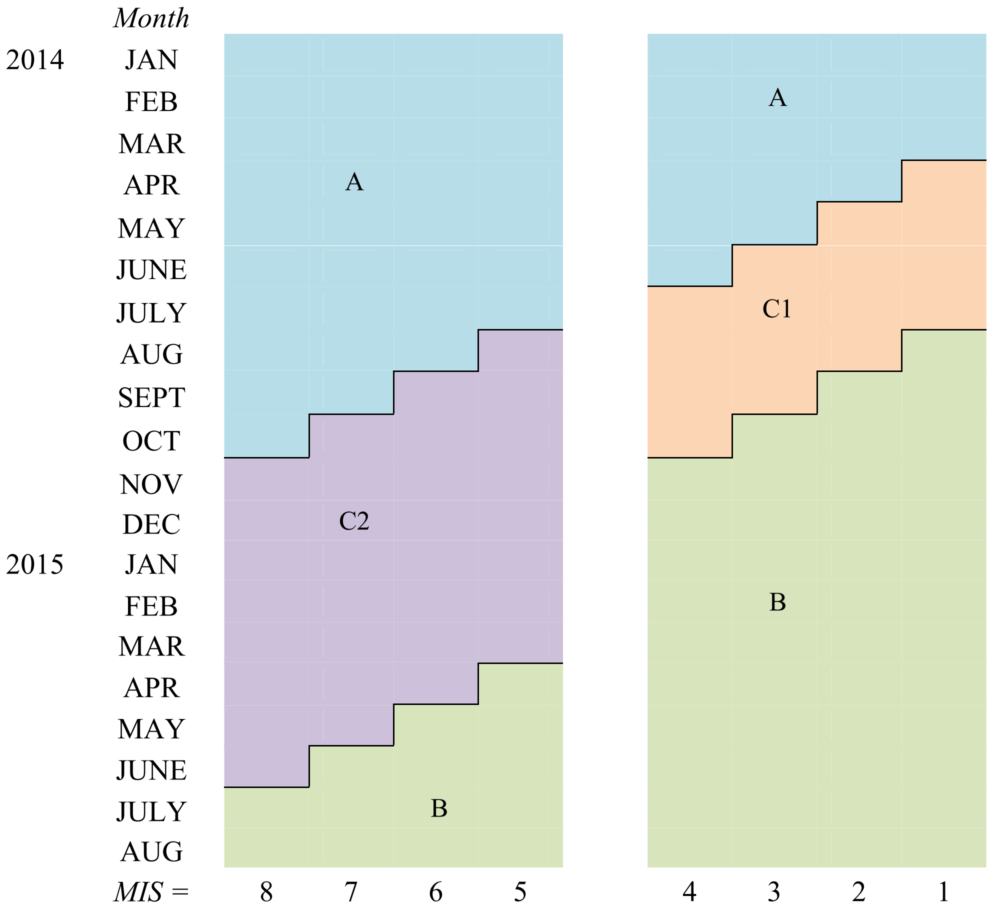


Figure 5: Visualization of the CPS PIPO. Columns represent panels and rows represent months. Blue cells (region A) represent sample from the 2000 design, green cells (region B) represent sample from the 2010 design, and the orange (region C1) and purple (region C2) cells represent sample that is a mixture of the two designs

As with the categorical data analysis of the RO realignment, a loglinear model may be the basis for exploratory analysis of an effect due to the sample design change. The eight panels in a given month are independent replicates of the second stage sample, although it is well known that the number of monthly interviews a housing unit has had will affect the expected value of its response (US Census Bureau, 2006), so MIS is an important predictor of cell proportions. The underlying model is:

$$log(m_{ij}) = \mu + \alpha_i + \delta_j + e_{ij} \tag{4}$$

Where m_{ij} is the cell count of the characteristic, μ is the overall effect, α_i is the effect of *Month i*, and δ_j is the effect of *MIS j*

A chart similar to the ones in figures 2 and 3, of standardized residuals will highlight the possible effect that the sample design changes are having on survey responses.

3.2 Modeling response outcomes across the interview period

The daily data we receive will allow analysis of response outcomes across the interview period. Refusals and noncontacts are of special interest, as well as number of contact attempts.

$$log(m_{ijk}) = \mu + \alpha_i + \gamma_j + \theta_k + e_{ijk} \tag{5}$$

Where m$_{ijk}$ is the cell count of the characteristic, μ is the overall effect, α_i is the effect of *Month i*, γ_j is the effect of *Day j*, and θ_k is the effect of *Demographic group k*

4. Future Research

The application to 2010 Sample Redesign is still in the early stages of development. It would be of particular interest to develop a model that tracks a partition of total survey error across time for different key variables. The partition of error may include those due to stages of sampling, interviewer effects, and errors due to nonresponse, coverage, and population controls.

The ACS may provide a valuable resource for all the surveys due to its large sample size, low nonresponse rates, and the fact that it provides reliable estimates at the county level. The new frame for our demographic surveys is the MAF. Before sampling takes place, many units on the MAF are filtered out based on the likelihood that they would not be valid housing units (Loudermilk and Li, 2009). The filter rules are the same for all of the Census Bureau surveys that sample from the MAF, so the ACS may be used to assess coverage errors for the other surveys.

A general framework for tracking survey error components involving the generalized regression estimator (GREG) as a foundation (Sarndal, Swensson, and Wretman, 1992) is described below. The GREG can be expressed in different ways, but for our purposes consider the following expression:

$$\tilde{Y}_g = \hat{Y}_{HT} + (\breve{\boldsymbol{X}} - \hat{\boldsymbol{X}}_{HT})'\hat{\boldsymbol{\beta}} \tag{6}$$

Where $\tilde{Y}_g$ represents the GREG, $\hat{Y}_{HT}$ represents the Horvitz-Thompson (HT) estimator of the characteristic of interest, $\breve{\boldsymbol{X}}$ represents a vector of control totals, with $\hat{\boldsymbol{X}}_{HT}$ being the HT estimator of those controls, $\hat{\boldsymbol{\beta}}$ is an estimated vector of regression coefficients from a regression of the variable of interest onto the control group indicators

Assume the HT estimators have the form:

$$\begin{aligned} \hat{Y}_{HT} &= Y + \hat{a}_{y,i} + \hat{b}_{y,i,j} + \hat{c}_y \\ \widehat{\boldsymbol{X}}_{HT} &= \boldsymbol{X} + \widehat{\boldsymbol{a}}_{x,i} + \widehat{\boldsymbol{b}}_{x,i,j} + \widehat{\boldsymbol{c}}_x \end{aligned} \qquad (7)$$

Where Y and $\boldsymbol{X}$ represent the population parameters; $\hat{a}_{y,i}$ and $\widehat{\boldsymbol{a}}_{x,i}$ represent the errors due to the first stage sample (i indexes the set of PSUs selected); $\hat{b}_{y,i,j}$ and $\widehat{\boldsymbol{b}}_{x,i,j}$ represent the errors due to the second stage sample (i and j together index the set of housing units selected); $\hat{c}_y$ and $\widehat{\boldsymbol{c}}_x$ represent the coverage error

Since the coverage errors are not assumed to have expected value zero, the HT estimator is biased in this expression. The coverage terms can be decomposed further into components for housing unit and person coverage. The housing unit component would be due to coverage errors on the MAF, and the person component would be due to errors in listing eligible persons within each respondent household.

Each January, the population controls are updated using more complete data sources, so there are often discontinuities in the monthly population controls between December and January. Smoothed estimates of the population controls in months other than January would use the following January's controls. Assuming these smoothed estimates are close to the true parameters, they can be used to estimate population control errors.

Assume the vector of control totals has the form:

$$\breve{\boldsymbol{X}} = \boldsymbol{X} + \widehat{\boldsymbol{d}}_x \qquad (8)$$

Where $\widehat{\boldsymbol{d}}_x$ is the error in the population controls

Under this general framework, the errors in the GREG can be partitioned as:

First stage sampling error is $\hat{a}_{y,i} - \widehat{\boldsymbol{a}}_{x,i}{}'\widehat{\boldsymbol{\beta}}$
Second stage sampling error is $\hat{b}_{y,i,j} - \widehat{\boldsymbol{b}}_{x,i,j}{}'\widehat{\boldsymbol{\beta}}$
Coverage error is $\hat{c}_y - \widehat{\boldsymbol{c}}_x{}'\widehat{\boldsymbol{\beta}}$
Population control error is $\widehat{\boldsymbol{d}}_x{}'\widehat{\boldsymbol{\beta}}$ (9)

When the x-variables represent cells in a contingency table, the terms involving $\widehat{\boldsymbol{\beta}}$ can be expressed in the following form:

$$\widehat{\boldsymbol{f}}'\widehat{\boldsymbol{\beta}} = \sum_k \varphi_k \hat{p}_k \tag{10}$$

Where $\hat{p}_k$ is the proportion with the characteristic of interest within cell k, and φ_k is a coefficient that depends only on the x-variables and not on the response variable

In equation (10), each coefficient φ_k summarizes a component of error in cell k. The set of coefficients, defined for each cell, would provide a measure for comparing the relative importance of different sources of error for different populations.

References

Agresti, A. (1990). Categorical Data Analysis, New York: John Wiley & Sons, Inc.

Loudermilk, C., and Li, Mei. (2009), "A National Evaluation of Coverage for a Sampling Frame Based on the Master Address File (MAF)," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC.

Nguyen, T., and Gerstein, A. (2011). "Sample Design Research in the 2010 Sample Redesign," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC.

US Census Bureau (2006). "Current Population Survey: Design and Methodology, Technical Paper 66," October 2006.

Walsh, R., and Fields, J. (2012). "Reducing Respondent Burden: Evaluating the Progress of the SIPP-EHC," Federal Committee on Statistical Methodology Conference, Washington, DC.