Small Area Estimation Applications in the US Census Bureau

Yang Cheng[1,3], Bac Tran[1], Carma Hogue[1], and Partha Lahiri[2]
[1]US Census Bureau, Washington DC, USA
[2]JPSM, University of Maryland, College Park, USA
[3]Corresponding author: Yang Cheng, e-mail: yangcheng@census.gov

Abstract

We first discuss the problem of small area estimation that arises in the context of the U.S. Census Bureau ---- Annual Survey of Public Employment and Payroll (ASPEP) and Current Population Survey (CPS). The direct survey-weighted estimates of employment for small domains are highly variable. In this talk, we illustrate the application of a small area estimation methodology to estimate employment by combining ASPEP with the previous census records using an empirical best prediction (EBP) methodology. The ASPEP data are usually subject to skewness and heteroscedasticity and thus the well-known EBP methodology based on unit level linear mixed normal model does not fit well. In order to get around the problem, we apply a unit level linear mixed normal model on the log-transformed employment. We evaluate different competing estimates using the census data.

Keywords: Borrow Strength, EBLUP, Heteroscedasticity, Linear Mixed Model

1. Introduction

Over the last few decades, the U.S. Census Bureau has pioneered in developing innovative small area methodologies in different programs. In one of the most cited papers in small area estimation literature, Fay and Herriot (1979) developed a parametric empirical Bayes method to estimate per-capita income of small places with population less than 1000 and demonstrated, using the census data, that their method was superior to both direct design-based and synthetic methods. More recently, researchers at the Census Bureau implemented both empirical and hierarchical Bayes methodologies in the context of Small Area Income and Poverty Estimates (SAIPE) and Small Area Health Insurance Estimates (SAHIE) programs; see Bell et al. (2007) and Bauder et al. (2008).

Besides the Census Bureau's well-known SAIPE and SAHIE programs, researchers in the Governments Division and the Current Population Survey (CPS) branch of the Demographic Statistical Methods Division are actively pursuing state-of-the-art small area estimation techniques to improve the current estimation methodologies. In particular, the CPS Branch is in the process of extending and evaluating the well-known triple-goal estimation method, first proposed by Shen and Louis (1998), using the CPS data and administrative records. The triple-goal method is being pursued to meet the needs of multiple users interested in using estimates for different purposes, including ranking small areas in terms of a parameter of interest, identifying small areas with parameters above or below certain thresholds, and estimating parameter of individual small areas. Research findings from this project on multi-goal small area estimation will be reported in an upcoming ISI-IASS satellite meeting on small area estimation to be held in Bangkok, on September 1-4, 2013.

In this paper, we report our preliminary work on small area estimation for an important establishment survey involving government units. The Governments Division of the Census Bureau conducts a census of about 90,000 state and local government units every five years, in part to collect data on the number of full-time and part-time state and local government employees and payroll. Between two consecutive censuses, the Governments Division also conducts the Annual Survey of Public Employment and Payroll (ASPEP), a nationwide sample survey covering all state and local governments in the United States, which includes five types of governments: counties, cities, townships, special districts, and school districts. The first three types of governments are referred to as general-purpose governments, because they generally provide multiple government activities. Activities are designated by function codes. School districts cover only education functions. Special districts usually provide only one function, but can provide two or three functions. ASPEP is the only source of public employment data by program function and full-time/part-time detail. Data on employment include the number of full-time and part-time employees and gross pay, as well as hours paid for part-time employees. All data are reported for the government's pay period covering March 12. Data collection begins in March and continues for about seven months. For more information on the survey, we refer to http://www.census.gov/govs/apes.

In 2009, ASPEP was redesigned and the old sample design was replaced by a systematic stratified probability proportional-to-size (PPS) modified cut-off sample design in order to reduce sample size and respondent burden for small townships and special district governments. At the same time the goal was to improve the precision of the estimates and data quality. The sample design was implemented in multiple steps. First, a state-by-governmental type stratified PPS sample was selected, where size was taken as the total payroll (the sum of full-time pay and part-time pay) from the employment component of the 2007 Census of Governments. In the second stage, a cut-off point was constructed to distinguish small and large government units in municipal and special district strata. Lastly, the strata with small-size government units were subsampled using a simple random sampling design.

The ASPEP survey is designed to produce reliable estimates of the number of full-time and part-time employees and payroll at the national level and for large domains (e.g., government functions such as elementary and secondary education, higher education, police protection, fire protection, financial administration, judicial and legal, etc., at the national level, and states aggregates of all function codes). It is also required to estimate the parameters for individual function codes within each state. This requirement prompted us to explore small area estimation methodology that borrows strength from previous census data as an alternative to collecting expensive additional data for small cells. We refer to Rao (2003) and Jiang and Lahiri (2006) for a comprehensive account of small area estimation theory and applications. In Section 2, we briefly describe our method. In Section 3, we present our findings from our data analysis.

2. Proposed Method

Let y_{ij} denote the number of full-time employees for the jth governmental unit within the ith small area ($i=1,\cdots,m;\ j=1,\cdots,N_i$). In this paper, we are interested in estimating the total number of full-

time employees for the ith small area given by $Y_i=\sum_{i=1}^{N_i} y_{ij}\ (i=1,\cdots,m)$. An estimator of Y_i is given by: $\hat{Y}_i = N_i\left[f_i\bar{y}_i + (1 - f_i)\hat{\bar{Y}}_{ir} \right]$ (1)

where $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ is the sample mean of the i[th] small area; $f_i = n_i/N_i$, N_i and n_i are the sampling fraction, number of government units in the population and sample for area *i*, respectively; $\hat{\bar{Y}}_{ir}$ is a model-dependent predictor of the mean of the non-sampled part of area *i* $(i = 1, \cdots, m)$.

In this paper, we obtain $\hat{\bar{Y}}_{ir}$ using the following nested error regression model on the logarithm of the number of full-time employees at the government unit level:

$$\log(y_{ij}) = \beta_0 + \beta_1 \log(\bar{X}_i) + v_i + \varepsilon_{ij}, \tag{2}$$

$$v_i \overset{iid}{\sim} N(0, \tau^2) \text{ and } \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \tag{3}$$

where $\bar{X}_i$ is the average number of full-time employees for the i^{th} small area obtained from the previous census; β_0 and β_1 are unknown intercept and slope, respectively; v_i are small area specific random effects. The distribution of the random effects describes deviations of the area means from values $\beta_0 + \beta_1 \log(\bar{X}_i)$; ε_{ij} are errors in individual observations $(j = 1,..., N_i;\ i = 1,..., m)$. The random variables v_i and ε_{ij} are assumed to be mutually independent. We assume that sampling is non-informative for the distribution of measurements y_{ij} $(j = 1,..., N_i;\ i = 1,..., m)$. A similar model without logarithmic transformation can be found in Battese et al. (1988). The logarithmic transformation is taken to reduce the extent of heteroscedasticity in the employment data. A similar model using unit level auxiliary information was considered by Bellow and Lahiri (2012) in the context of estimating total hectare under corn for U.S. counties. We use the following model-based predictor of $\bar{Y}_{ir}$:

$$\hat{\bar{Y}}_{ir} \approx \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \log(\bar{X}_i) + \hat{v}_i + \frac{1}{2}\hat{\sigma}^2 \right], \tag{4}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{v}_i$, and $\hat{\sigma}^2$ are obtained by fitting (2) using PROC MIXED of SAS. We obtain our estimate of total number of full-time employees in area *i* using equations (1) and (4).

3. Data Analysis

For our data analysis, we first created a dataset by including only those government units that overlap between the 2002 and 2007 Census of Governments units reporting strictly positive numbers of full-time employees. The analysis covered 49 states, excluding Washington, D.C, and Hawaii because we collected data from the entire population of governments in those states.

We drew a sample from the 2007 Census of Government units and computed the following estimates of total full time employees for each of the 29 function codes available for all the local governments: direct Horvitz-Thompson estimate (denoted by HT), EBLUP estimate of Battese, Harter, and Fuller (1988) (denoted by BHF), and our proposed estimate (log transformation). For a given function code, we compute: Percent Relative Error = 100(Estimate – True)/True (denoted by PRE), where truth is the 2007 Census full-time employees for that function code. There are 1,298 function codes across 49 states; only 241 of them (18.6 percent) show HT having a better. In these cases the sample sizes are relatively large with a large number of certainty cases. Table 3.1 displays these percent relative errors for the three estimates for California only. From this table, it is clear that our proposed estimates are significantly better than the BHF estimates for all function codes. Moreover, in 24 out of 29 function codes, our estimators are better than the HT estimators. We observe that our proposed estimates are generally better than both the methods for function codes with small sample size (n=2) like Gas Supply.

Figures 3.1 and 3.2 display our residual analysis for our proposed model and BHF model for California. As you can see, the residual QQ-plot of our model is better than the BHF model.

Figure 3.1: Residual Plot for California: Proposed Model

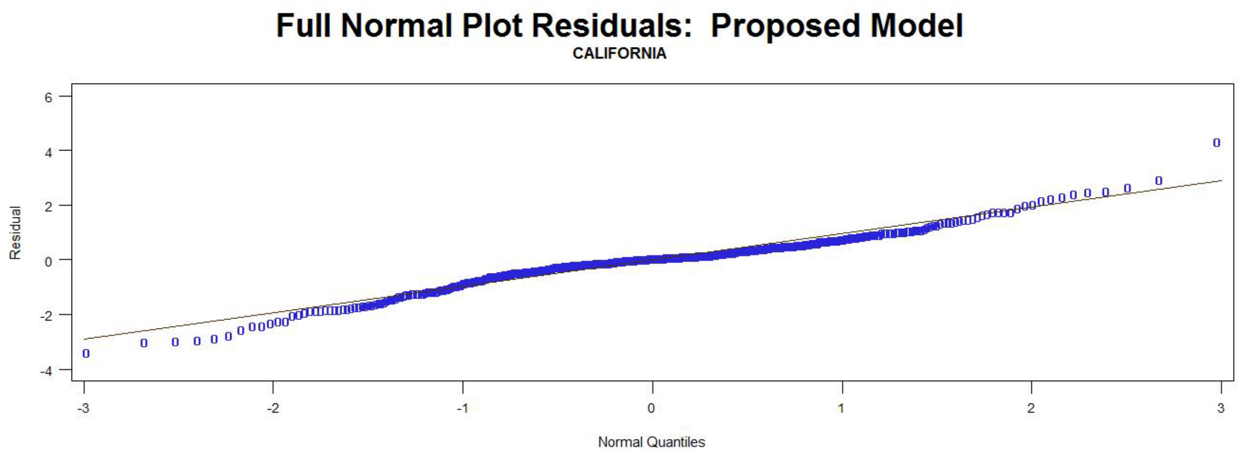
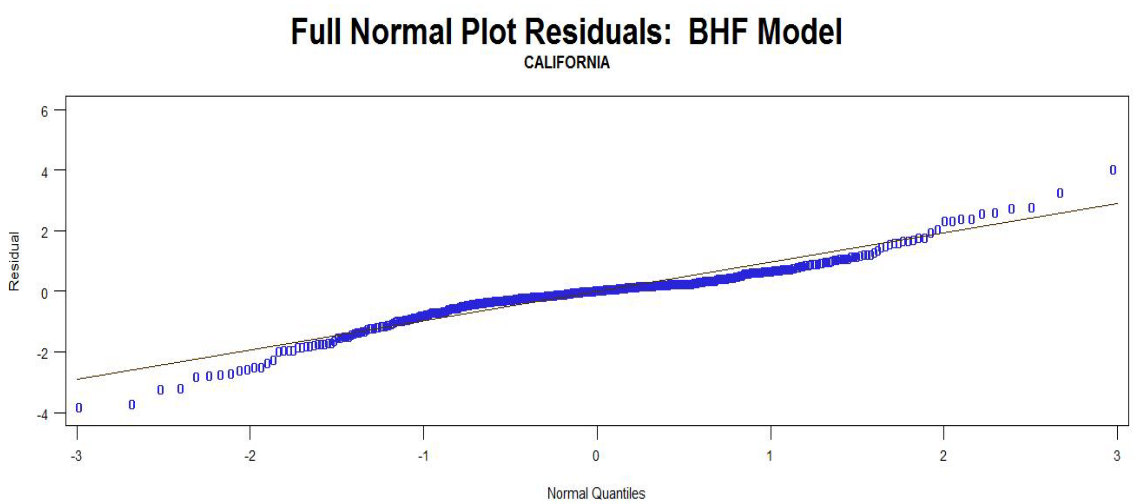


Figure 3.2: Residual Plot for California: BHF Model



We computed benchmarking ratios (BR) for both our model and the BHF model. The BR isdefined as $|(\sum(est - HT)/\sum HT)|$. The BR indicates how close the estimate (est) is to the HT when considering large areas. We defined a cell to be considered as small size if the sample was smaller than 50 units. We estimated the BR for all the states by size. Table 3.2 summarizes the benchmarking ratios of the proposed model and the BHF model.

Table 3.1: Percent Relative Errors for Different Estimates of Full Time Employees-(California, in percentage)

Function	HT	Proposed	BHF
Airports	4.34	-0.49	-2.49
Correction	0.71	0.17	-3.46
Elementary and Secondary - Instruction	-1.52	-4.08	-27.7
Higher Education - Other	5.72	-0.19	-9.97
Higher Education - Instructional	4.48	0.86	-9.14
Financial Administration	-1.58	-0.65	-12.0
Firefighters	2.91	-1.36	-19.5
Judical & Legal	0.39	0.82	-2.21
Other Government Administration	-1.95	-0.12	-16.2
Health	-2.97	-0.08	-6.26
Hospitals	4.77	-0.71	-5.81
Streets & Highways	-3.36	0.11	-19.7
Housing & Community Development (Local)	-5.18	-2.11	-27.6
Local Libraries	5.72	-0.06	-10.6
Natural Resources	-3.74	-2.46	-25.0
Parks & Recreation	2.14	-2.11	-19.3
Police Protection - Officers	0.07	-0.21	-14.4
Welfare	-1.67	-0.14	-3.30
Sewerage	3.57	-1.91	-20.9
Solid Waste Management	3.73	-1.58	-12.3
Water Transport & Terminals	34.14	-1.64	-15.4
Other & Unallocable	-0.29	-1.65	-14.5
Water Supply	1.18	-7.20	-30.5
Electric Power	-1.28	-0.30	-4.87
Gas Supply	41.60	-11.8	-30.6
Transit	-1.37	-1.18	-8.49
Elementary and Secondary - Other Total	-0.87	-2.92	-22.6
Fire - Other	-9.03	-1.23	-10.1
Police-Other	2.04	-0.12	-11.3

Table 3.2: Comparison of Benchmarking Ratios

Size	BR of the proposed model	BR of the BHF model
< 50	1.5	1.7
> 50	1.1	1.4

Conclusion

The HT estimator performed poorly in cells where sample sizes are relatively small. The BHF model failed because the data are so skewed. Our proposed estimator outperformed the HT on small areas and on some large areas as well. We will follow up with further research to see if the result is consistent with all the states. We will also compare the result with the composite estimator. Lastly, we will group cells into areas and apply the estimator that produces the least bias.

Acknowledgements

The last authors' research was supported in part by Census-BAE subcontract #41-1016588.

References

1. Battese, G.E., Harter, R.M., and Fuller, W.A. (1988) "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 83, 28-36.

2. Bauder, M., Riesz, S., and Luery, D. (2008) "Further Developments in a Hierarchical Bayes Approach to Small Area Estimation of Health Insurance Coverage: State-Level Estimates for Demographic Groups," Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, 1726-1733.

3. Bell, W., Basel, W., Cruse, C, Dalzell, L., Maples, J., O'Hara, B., and Powers, D. (2007) "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties," Census Report.

4. Bellow, M. and Lahiri, P. (2010) "Empirical Bayes Methodology for the NASS County Estimation Program," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

5. Fay, R.E. and Herriot, R.A. (1979) "Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data," *Journal of American Statistical Association*, 74, 269-277.

6. Jiang, J., and Lahiri, P. (2006) "Mixed Model Prediction and Small Area Estimation (with discussions)," *Test,* 15, 1, 1-96.

7. Rao, J.N.K. (2003) *Small Area Estimation*, New-York, John Wiley & Sons, Inc.

8. Shen, W. and Louis, T. (1998) "Triple-goal estimates in two-stage hierarchical models," Journal of Royal Statistical Society, Series B, 60:455–471.