EVALUATION OF THE SYNTHETIC ASSUMPTION - 1990 POST-ENUMERATION SURVEY

Jay Kim, Robert Blodgett - U.S. Bureau of the Census
Alan Zaslavsky - Harvard University
Jay Kim, U.S. Bureau of the Census, Washington D.C. 20233

Key Words: Homogeneity, Census Adjustment, Post-Enumeration Survey

1. BACKGROUND

1.1 Introduction

The Census Bureau's methodology for census adjustment requires: 1) poststratifing the sample persons on the post-enumeration survey (PES), 2) calculating dual system estimates for the poststrata, 3) calculating adjustment factors, 4) smoothing adjustment factors and 5) applying or "carrying down" the smoothed adjustment factors to the Census counts by block in the appropriate poststratum. This procedure, called "synthetic adjustment", assumes that the probability of being missed by the census is the same for all people in a poststratum.

The design of PES poststrata tried to achieve homogeneity within a poststratum with respect to the expected undercount rate. The homogeneity assumption states that the undercount rate does not vary for subgroups within a poststratum. This paper investigates the validity of this homogeneity, or synthetic, assumption, focusing on homogeneity between states within a poststratum. Two different approaches were taken for this test. The first analyzes five surrogates for the undercount rate from the 1990 Census: the allocation rate, mail return rate, multi-unit structure rate, mail universe rate and substitution rate. These rates were calculated for each sampled block part. A "block part" is the intersection of a block and a poststratum.

The second approach analyzes influence statistics from the PES obtained by linearizing the effect of each block on the adjustment factor. Also, the explanatory power of the state for the adjustment factor is compared with the explanatory power of the poststratum group.

1.2 Data Sets

The analysis involves both an extract of 1990 census data and the PES data. The census data are a stratified cluster sample, similar to PES but with 204,394 blocks, or about 125,000 block clusters. It was selected from the 1990 census using the 100% Edited Detail File. For more detail on the sampling, see reference (1). This file contains rates for the surrogate variables for each block part. These variables are analyzed by logistic regressions.

The PES data, with 12,144 blocks has two data sets, one for the E-sample and the other for the P-sample. Correct enumerations and E-sample total counts are on the E-sample file and match and P-sample total counts are on the P-sample file. From these, we will derive undercount rates and influence statistics. The undercount rates and influence statistics are analyzed with a linear model.

2. DATA ANALYSIS

2.1 Analysis of Census Data

2.1.1 Methods of Analysis

The logistic regression model takes the following form:

$$\log [P_{ij} / (1-P_{ij})] = A + B_i + C_j$$

where P_{ij} is the rate for a surrogate variable, in the i^{th} poststratum or poststratum group and j^{th} state, A is the intercept, B_i is the i^{th} poststratum effect and C_j is the j^{th} state effect. The models used only the 99 poststratum groups astride two or more states. Also, separate models were built for the 99 poststratum groups and for each of nine divisions. The maximum likelihood procedure of SAS's PROC CATMOD estimated the parameters and provided the Wald statistics (Chi-square values).

The test statistics assume simple random sampling. However, the data was collected with a cluster sample, and thus the test statistic must be adjusted. We estimate a design effect of the following form:

$$\hat{D}_{ij} = 1.25 \frac{\Sigma n_{ijk}(\hat{p}_{ijk}-\hat{p}_{ij})^2}{I\hat{p}_{ij}(1-\hat{p}_{ij})}$$

where $\hat{p}_{ijk}$ is the rate for the i^{th} poststratum (or poststratum group), j^{th} state and k^{th} block part; n_{ijk} is the size of the block part; I is the number of block parts in the area of our interest and $\hat{p}_{ij}$ is the rate for the i^{th} poststratum (or poststratum group) and j^{th} state. This fraction is the ratio of the observed between block variance to that expected under binomial sampling.

The arbitrarily chosen factor, 1.25, accounts for the fact that the poststratum group has more clustering than the whole area it covers. Within the poststratum group blocks were selected and they tend to be homogeneous by race and tenure.

2.1.2 Data Analysis

Table I summarizes the findings with respect to state effects. Nationally, the State effect is significant at $\alpha = .05$ for 95% of 99 poststratum groups for the allocation rate, 93% of 99 groups for the mail return rate, 96% of 99 groups for the multiunit structure rate, 84% of 43 groups for the mail universe rate and 88% of 84 groups for the substitution rate. The reason for different number of poststratum groups is that for some surrogate variables models can not be fit for some poststratum groups.

Table I. Number of Poststratum Groups with Significant (α=.05) State Effect (Results of Logistic Regression)

Div.	No. Grp	Alloc	Mail Ret	Mult Str	Mail Unv	Sub
1	5	5	5	5	1(1)	3(4)
2	12	11	11	12	7(10)	12
3	16	15	16	16	3(3)	12(12)
4	8	8	8	7	5(6)	5(8)
5	10	10	9	10	4(4)	7(8)
6	15	15	13	15	5(7)	15
7	9	8	9	9	4(4)	8(8)
8	7	7	7	7	2(3)	6(6)
9	17	15	14	14	5(5)	6(12)
Sum	99	94	92	95	36(43)	74(84)

The numbers in ( ) are the number of poststratum groups for which Chi-square values are available when less than the number of groups.

Four poststratum groups were selected to check whether or not each of the poststrata in these groups has a significant state effect. Interstate homogeneity was tested for the poststrata with respect to the surrogate variables which proved important in the model for smoothing the adjustment factors. The poststratum groups were selected to have a range of test statistics for state effects from a chi-square of 3933 with 7 degrees of freedom to a chi-square of 124 with 4 degrees of freedom.

These poststratum groups indicate that when a poststratum group has highly significant state effect, most of the poststrata within the poststratum group have significant state effect. However, if the significance level for the poststratum group is not high, not as many poststrata show significance for the state effect.

2.2 Analysis of PES Data

2.2.1 Methods of Analysis

Assuming the substitution rate is negligible, the adjustment factor ($\hat{R}$) for poststratum is

$$\hat{R}=\frac{WCE/WE}{WM/WP},$$

and the undercount rate is

$$1-1/\hat{R},$$

where WE and WP are the estimated population sizes from the E and P-sample, respectively. WCE is the weighted number of correct enumerations and WM is the weighted number of matches.

The statistic for the influence of the i^{th} block part on the adjustment factor or undercount rate is

$$I_i=\hat{R}\left(\frac{WCE_i}{WCE}+\frac{WP_i}{WP}-\frac{WE_i}{WE}-\frac{WM_i}{WM}\right),$$

where WCE_i, WP_i, WE_i and WM_i are similar to the above for the i^{th} block part.

A Linear model tested for the state effect on the influence statistics. Under the null hypothesis, the states have all the same undercount rate and expected mean of the influence statistics for each state is 0 within each poststratum group. The influence statistic can be analyzed with either one way analysis of variance (ANOVA) within a single poststratum group or two way ANOVA for all poststratum groups within a division.

Undercount rates for state parts were analyzed with two way linear model. This technique compares the size of the poststratum group and state effects in aggregate.

2.2.2 Data Analysis

Table II summarizes significance tests on influence statistics at the division level. The models include both poststratum group and state effects. Table II shows the F value for state effect adjusted for the poststratum group effect. These tests show no significant state effect except in Division 2 (Mid-Atlantic).

Table II. Influence Statistics at the Division Level

Div.	F value
1	.57
2	4.64
3	.43
4	.64
5	.66
6	.60
7	.39
8	.62
9	.77

Table III summarizes the significance tests on influence statistics at the poststratum group level. The models include only state effect. The tests reveal significant heterogeneity between states in 9 out of 99 groups at the 5 percent significance level and 15 out of 99 groups at the 10 percent level.

Table III. Influence Statistic at the Poststratum Group Level

Div.	No. Grp	α=.05	α=.10
1	5	1	1
2	12	2	2
3	16	2	2
4	8	0	2
5	10	0	1
6	15	2	3
7	9	0	0
8	7	0	1
9	17	2	2
Sum	99	9	15

These significant results are concentrated in poststratum groups for rural areas (place type 7, 8 and 9). 9 out of 32 such groups show significant interstate heterogeneity at the 10 percent level (see Table IV).

Table IV. Influence Statistic by Place Type

Type	No. Grp	Significant
0	11	3
1	23	0
2	10	1
3	7	0
4	0	0
5	6	2
6	6	0
7	7	2
8	10	3
9	10	4
2,5	1	0
2,7	1	0
3,7	1	0
4,5	1	0
7,8,9	4	0
8,9	1	0

A two-way ANOVA was fitted to undercount rates for state parts (intersections of a state and poststratum group). Table V shows the ratio of the sum of squares due to poststratum groups to that due to states within a division. The ratio is always greater than one and in Division 9 it is 40.28. It shows much larger effects for poststratum group than for state. This supports the decision to make use of poststratum rather than state as the cell for undercount estimation and adjustment.

Table V. Sum of Squares Ratios

Div.	SS(Grp)
	--------- 	
	SS(State)
		No. Grp	No. States

1	4.51	5	6
2	4.88	12	3
3	12.69	16	9
4	8.73	8	4
5	8.17	10	4
6	7.67	15	5
7	2.78	9	7
8	1.31	7	8
9	40.28	17	5

States include D.C.

3. Comparison of Methods

The two methods considered give different results. The method with surrogate variables and Census data shows significant state effects. The other with PES data does not. There may be two reasons for this discrepancy. Thus, significant state effects for surrogate variables do not necessarily imply significant state effects for undercount.

First, the sample size of the two data sets was very different. The Census data had 204,394 blocks and the PES data had 12,144 blocks.

Second, the correlation between the surrogate variables and undercount is not perfect. Table VI gives the correlations.

Table VI. Correlation

Var.	Cor.
Alloc.	.44
Mail Ret	-.57
Mult Str	.39
Mail Unv	.08
Sub	.47

4. Summary

This paper evaluates the homogeneity, or synthetic, assumption.

The evaluation used 1990 Census data and 1990 PES data. Surrogate variables from the 1990 Census tested for significant homogeneity among states within the poststratum or poststratum group. At the poststratum group level, state effect was significant ($\alpha = .05$) for 84%-95% of its poststratum groups for the various surrogate variables.

When the poststratum group showed strongly significant state effect, every one of the poststrata within the poststratum group also showed significant state effect. When the poststratum group showed marginally significant state effect, part of the poststrata within the group showed significant state effect.

The analysis of variance on the influence statistics at the division level showed a significant state effect only for Division 2. The same test at the poststratum group level showed significant ($\alpha = .10$) state effects for 15 out of 99 poststratum groups. The significant results were concentrated in the poststratum groups in the places of types 7, 8 and 9. Nine out of 32 such poststratum groups had significant state effects.

The findings are different between the two approaches taken. The reason for the difference is that the two data sets had different sample sizes and the surrogate variables were not perfectly correlated with undercount.

This paper reports the general results of research undertaken by the authors. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or Harvard University.

5. References

(1). Bateman, D., STSD 1990 Coverage Studies and Evaluation Memorandum Series #N-1, Revision 1, Sample Selection Procedures for Performing Evaluation Study P12, October 3, 1990.

(2). Bateman, D., 1990 Coverage Studies and Evaluation Memorandum Series #N-4, Final Report for PES Evaluation Project P12:Evaluation of Synthetic Assumption, July 11, 1991.