Investigation of the Relationship between Census Mail Response and the Partnership and Marketing Program

Mary H. Mulry and Jay K. Keller[1]
U.S. Census Bureau, Washington, DC 20233

Abstract

The U.S. Census Bureau is analyzing data regarding the influence of the Partnership and Marketing Program on the Census 2000 mail return rate with the goal of optimizing the use of limited funds for the 2010 Census communications campaign. Evaluations and analyses conducted during and after Census 2000 did not provide the U.S. Census Bureau with specific information about how a dollar's worth of advertising spending affected public cooperation or translated into dollar savings, nor how this played out by market segment. Still, it is incumbent upon the U.S. Census Bureau to try and develop such business models to provide guidance to the 2010 Census communications campaign. Two methods using multivariate analyses are discussed.

Keywords: Census 2000, logistic regression, segmentation, return on investment

1. Introduction

The U.S. Census Bureau is analyzing data regarding the influence of the Partnership and Marketing Program (PMP) on the Census 2000 mail return rate with the goal of optimizing the use of limited funds for the 2010 Census communications campaign. Census 2000 was the first time the U.S. Census Bureau paid for advertising in a census, and the cost was $167 million. "Previous censuses had relied on public service announcements." The mail response rate in Census 2000 was viewed as a success at 67 percent, much higher than the forecasted 61 percent (U.S. Census Bureau 2000).

However, evaluations and analyses conducted during and after Census 2000 did not provide the U.S. Census Bureau with specific information about how a dollar's worth of advertising spending affected public cooperation or translated into dollar savings, nor how this played out by market segment. Still, it is incumbent upon the U.S. Census Bureau to try and develop such business models to provide guidance to the 2010 Census communications campaign. Two methods using multivariate analyses may aid in developing such business models.

The approaches discussed in this paper are observational, but still appear worthwhile to pursue. Analyses of advertising and marketing during a census are plagued by the fact that conducting an experiment to estimate their effect is impossible. Not advertising the census in a major city to be able to evaluate the effect of advertising is unacceptable for a once-a-decade population count. Even if it were acceptable, residents would have other means of seeing census advertisements, such as the Internet, and that would confound the measurements. Experiments are possible during census tests, but since the tests are not the real census, the behavior of the public does not represent the behavior during a census.

Logistic regression models that relate the log-odds of a mail return to a linear function of demographic variables and variables for the awareness of the census promotional communications provide the basis for one approach. The respondents to a survey that measured the public's awareness of census and the promotional communications were linked to the census databases to determine if the census received a mail return for them. We use the logistic regression model for the log-odds of a mail return to estimate the probability of returning a form for those who have not seen any of the communications.

[1] The authors thank Pat Cantwell, Jennifer Marks, and Rolando Rodriguez for their helpful comments on earlier versions. This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Next we estimate the probability for those who have "average" scores on seeing communications. Then multiplying the difference in the probabilities by the population size produces an estimate of the number of additional returns.

In the other approach, cluster analysis partitions census tracts into segments using variables correlated with being hard to count in the census. The logistic regression models are used to classify the tracts by high, medium, and low in probability of response, increase in probability of response due to advertising and marketing, and increase in number of responses due to advertising and marketing.

Section 2 discusses the advertising and marketing program for Census 2000 and its evaluations. Section 3 describes the methodology for using logistic regression models. The data and the results are discussed in Sections 4 and 5, respectively. Section 6 describes how the logistic regression models can be used in combination with cluster analyses to form segments for use in advertising and marketing. Section 7 contains a summary.

2. Awareness of PMP

The method for estimating the increase in the Census 2000 mail return rate due to the influence of the PMP uses logistic regression models that estimate the probability of mailing back a census form. The models were fit using data collected in a survey to assess the public's awareness of the census in combination with a check in the census databases to determine if the respondents returned a census form by mail. The mail return rate equals the proportion of the occupied housing units that mailed back a census form. The survey and logistic regression modeling were done by National Opinion Research Center as part of the Census 2000 Evaluation Program and described in a report by Wolter et al. (2003).

The U.S. Census Bureau's strategy for promoting the census had five components:

- Partnerships with businesses, governments, and organizations
- Paid advertising campaign
- Media relations campaign
- Promotions and special events, including Census in Schools and the Census Road Tour in addition to other materials
- Direct mail pieces, including the advance letter, questionnaire, and reminder postcard

The paid advertising campaign for Census 2000 was planned and conducted by Young and Rubicam under contract with the U.S. Census Bureau. The campaign included messages for an English-speaking audience with additional advertising focused toward populations who speak other languages and populations believed to be hard-to-count. The campaign had three phases:

- The educational phase began November 1, 1999 and lasted until January 30, 2000. The goal was to teach the public about the census.
- The motivational phase began February 28, 2000 and lasted until April 9, 2000. The bulk of the advertising was during this period. The goal was to encourage participation in the census.
- The nonresponse followup phase began on April 17, 2000 and lasted until mid June. The ads informed the public that enumerators would be visiting to collect census data.

The U.S. Census Bureau hired NORC to evaluate the PMP, the name for two of the components in the U.S. Census Bureau's promotion strategy, the partnerships and paid advertising. For the evaluation, NORC conducted three surveys, called waves, at three different times to collect data about the public's awareness of the census and the communication vehicles being used to promote the census. The surveys asked awareness of communications vehicles classified under the partnerships and paid advertising as well as awareness of some that could be classified under the media relations category and the promotions and special events category. The sample design included four different samples for each wave. The core sample covered the total population and allowed separate analyses of Hispanic, non-Hispanic black, and non-Hispanic white populations. In addition, a separate sample was selected for each of three other populations, Asians, Native Hawaiians, and Native Americans.

- The Wave 1 survey collected data from September 1, 1999 until November 13, 1999 and completed 3,002 interviews.

- The Wave 2 survey collected data from January 17, 2000 until March 11, 2000 and completed 2,716 interviews of which 1,193 were in a sample of the entire population, known as the core sample, and eligible for a mail return.
- The Wave 3 survey collected data from April 17, 2000 until June 17, 2000 and completed 4,247 interviews, 1944 of which were in a sample of the entire population, known as the core sample, and eligible for a mail return.

The respondents in Waves 2 and 3 were linked to the census databases to determine if a mail return was received for them. The link for Wave 1 was attempted but proved problematic. The nonresponse follow cut-off date that the U.S. Census Bureau used to calculate the mail return rate was April 18. The definition of a mail return was different in the NORC study. For Wave 2, a form "was classified as a mail return if it had a valid census mail return date that was prior to the nonresponse followup interview date (NRD) provided on the U.S. Census Bureau file. For Wave 3, a mail return must have occurred before the NORC interview date and the NRD" (Wolter et. al 2003 p. 87). The implication is that some of the Wave 2 or Wave 3 respondents designated as mailing back a form may have been sent for a nonresponse followup interview. The processing for nonresponse followup in 2000 did not allow for retrieving an address if a mail return was received after the cut-off date for sending the address to the field.

3. Methodology

Wolter et. al. (2003) fit logistic regression models relating the log-odds of a mail return to a linear function of exogenous variables with both Wave 2 and Wave 3 data. They used a stepwise procedure to arrive at a set of variables that produced the "best" model although the stopping criteria for the algorithm were not reported. The general form of the model that relates the log-odds of a mail return to the exogenous variables is

$$log(\frac{p}{1-p}) = \beta_0 + \sum_i \beta_i X_i \quad \text{where}$$

p = probability of mailing back form

X_i = the ith exogenous variable

β_i = estimated coefficient of ith variable

Once the logistic regression model is fit, we can solve for the probability of mailing back a census form

$$p = \frac{e^{\beta_o + \sum_i \beta_i X_i}}{1 + e^{\beta_o + \sum_i \beta_i X_i}} \quad .$$

The exogenous variables considered for the model include race/ethnicity, language spoken at home, household income, education, age, sex, and tenure (Wolter et. al. 2003, p. 97). Most of the variables have the form of indicator variables, but the variables concerning the PMP are continuous variables. Geographic variables were not considered and would have been hard to define, particularly because of the variation in the communications through partnerships and the contributions of local officials.

To estimate the probability p_0 of mailing back if there was no PMP, the variable concerning awareness of PMP is set to a value that indicates that a person has not heard or seen anything about the census from any of the communication modes. This assumes the probability of mail return for those with no awareness of PMP equals the probability if there were no PMP. To estimate the probability p_{PMP} of mailing back with the PMP, the variable concerning awareness of PMP is set equal to the average awareness observed in the sample. Using the average awareness in this way does not imply that every person has the same awareness but rather attempts to account for some not being aware of PMP even while it is in progress and others having vary degrees of awareness.

Then the increase in the number of mail returns M for a group as defined by the variables in the model may be estimated by the product of the group's population size N and difference in the probability of a mail return with and without the PMP

$$M = (p_{PMP} - p_0)N \; .$$

4. Data and models

Wave 2 appears to be the wave best suited for the purpose of estimating the increase in the mail return rate due to the PMP. The timing of Wave 2 has advantages and disadvantages for this application. Wave 2 was conducted in winter 2000 before the census forms were mailed, but Wave 3 was conducted after the census forms

should have been mailed back. A person's awareness in May or June may not have been the same awareness in April at the cut-off for mail returns. However, since the advertising and promotion of the census increased closer to Census Day, the assumption that no awareness is similar to no PMP may be more reasonable for Wave 2 than Wave 3.

The estimation of the increase in the mail returns due to the PMP for Hispanics, non-Hispanic Blacks, and non-Hispanic whites uses the model based on the Wave 2 core sample shown in Table 1 while each separate model is used for Asians, Native Americans, and Native Hawaiians shown in Tables 2 through 4. The estimation for the remaining races, including multiple races, uses the overall average of the probability of increase for Asians, Native Americans, and Native Hawaiians. The reason for using several models is an anomaly with the Wave 2 core sample model that makes it unsuitable for estimating the probability of an increase in mail response for others.

The variables concerning PMP were defined using questions regarding whether the respondent had seen or heard of the census through each type of communication, such as television, radio, things children brought home from school. The respondents answers were recorded on a three-point scale with "1" for not seeing or hearing anything through the communication type, "2" for hearing or seeing a little through the communication type, and "3" for hearing or seeing a lot through the communication type. The value of the variable for a respondent for mass-media communications (MM) was the simple mean of the responses for those modes of communications. Similarly, the value of the variable for a respondent for community-based communications (CB) was the average of the responses for those types of communications. When the correlation of 0.70 between the MM and CB variables appeared problematic during the model fitting with the Wave 2 core sample, Wolter et al. (2003) defined two new variables SUM = MM+CB and DIFF = MM-CB, which are uncorrelated.

The final model for the core sample in Wave 2 fit by Wolter et al. (2003) and shown in Table 1 includes SUM as a variable. Although SUM is not significant, the interaction of SUM and the variable for speaking a language other than English or Spanish at home, denoted in Table 1 by LANG-other-language, is significant. The model results for the group that speaks a language other than English or Spanish at home could be considered counter-intuitive. The coefficient of the variable LANG-other-language is highly positive at 5.86. However, the coefficient of the SUM variable is 0.43 and the coefficient of the interaction (SUM*LANG-other-language) is substantially negative at –2.06. When LANG-other-language = 1, if the SUM variable increases by the log-odds increases by (0.43-2.06)SUM, or –1.63SUM. This is the same as saying that increasing SUM by 1 when LANG-other-language = 1 multiples the odds of returning a form by exp(-1.63) = 0.20. The implication is more exposure to the advertising reduces the odds of returning a form. We do not have a good explanation for this result.

5. Results

Table 5 shows the estimated increase in the mail return rate and numbers due to PMP derived by applying the models by race/ethnicity group and using the average awareness of PMP measured in Wave 2. The percentage increase for the remaining races is the pooled rate for Asians, Native Hawaiians, and American Indians. The number of occupied housing units and the number in each of the categories defined by tenure, race/ethnicity, and language spoken at home are based on the characteristics of the householder from Census 2000 long form data.

Overall, the approach estimates an increase of 4.3 million mail returns due to the PMP. Since there are 104.7 million occupied housing units in the mailout/mailback universe, this implies that PMP increased the mail return rate by a 4.1 percentage points. The estimate is based on the average awareness of PMP measured in Wave 2. For Hispanics, non-Hispanic whites, non-Hispanic Blacks, the average awareness used was calculated by language spoken at home. For Asians, Native Hawaiians, and American Indians, the average awareness was the one measured for each group.

The average awareness of PMP measured in Wave 2 is probably lower than the awareness at the time the census forms were being mailed back since most of the interviews were conducted before the bulk of the motivational advertising and promotions. Therefore, the average of the awareness measured by Wave 3 probably is closer

to the awareness at the time the forms were being mailed back.

Table 6 shows the results obtained under the assumption that the awareness is the mean of the average awareness in Wave 3 and that the models are suitable for this inference. The estimated increase in mail returns rises about 1.3 million to 5.6 million. The estimated increase in the percentage of mail returns rises about 1.2 percent to 5.3 percent.

These results are corroborated somewhat by the increase in mail response observed by a nationwide test of the American Community Survey (ACS). During the months January through March, the mail response was 5 to 9 percentage points higher in 2000 than in those same months in 2001 when there was no advertising by the U.S. Census Bureau (Bentley, Trancreto, and Hill 2006). While the estimates of the increase based on the models are on the low end of that range, the higher rate of mail response observed may be random variation as opposed to being influenced by the advertising and promotions for Census 2000. One difficulty in comparing the ACS and Census 2000 response is that Census 2000 had both a long and short form while ACS questionnaire is very similar to the census long form, which had lower rates of return (Stackhouse and Brady 2003).

6. Segments for 2010

The U.S. Census Bureau's Tract-Level Planning Database with Census 2000 Data (Bruce and Robinson 2006) contains data that may be used in cluster analysis to partition census tracts into segments using variables correlated with being hard to count in the census. The logistic regression models may then be applied to classify the tracts by high, medium, and low for the increase in probability of response due to promotions and marketing and for the increase in number. Other levels of geography, such as counties, also are available on the Planning Database and may be used in cluster analyses.

To illustrate how segments may be classified, we use groups defined by race/ethnicity groups and some variables in the logistic regression models as shown in Tables 7 through 9. For this illustration, we classify the groups by the amount of increase in the probability of mail return due to PMP in Table 10 and by the increased number of mail returns due to PMP in Table 11.

Information such as shown in Tables 10 and 11 would be helpful in planning and implementing an advertising and marketing campaign. The group with the largest increase in probability of mail return due to PMP is not necessarily the group that will produce the largest increase in number of mail returns. For the larger groups a small increase in the probability of mailing back a census form produces a large number of forms. For example, the estimated percentage increase for 57.2 million non-Hispanic white owners is only 3.3 percentage points, but the estimated increase in number of mail returns is 1.8 million.

7. Summary

As the 2010 Census approaches, the U.S. Census Bureau is attempting to build business models to be able to measure the return on investment for the advertising and marketing campaign. The approach of estimating the increase in mail return rate using logistic regression provides a way of estimating the gain. In addition, the gain can be measured by segment of the population.

References

Bentley, M., Tancreto, J., and Hill, J. (2006) "Research on Census Response Rates and Partnerships and Marketing". Internal report dated September 6, 2006. Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC.

Bruce, A. and Robinson, J.G. (2006). "Tract-Level Planning Database with Census 2000 Data." U.S. Department of Commerce, U.S. Census Bureau: Washington, D.C.

Stackhouse, H. F. and Brady, S. (2003) "Census 2000 Mail Response Rates". Census 2000 Evaluation A.7.a. U.S. Census Bureau. Washington, DC.

Wolter, K. et al. (2003) "Partnership and Marketing Program Evaluation". Census 2000 Evaluation D.1. Prepared by NORC. U.S. Census Bureau. Washington, DC.

U.S. Census Bureau (2000) **"**Well Done, America!". Press Release dated September 19, 2000. U. S. Census Bureau. Washington, DC.

Table 1. Model for core sample from Wave 2

Independent Variables	Est. Coeff.	Estimated St. Error	p-value	
Intercept	1.27	0.97	0.19	
RACEETH				
Hispanic	-0.91	0.46	0.05	*
Black non-Hisp	-0.63	0.45	0.17	
Other	-0.77	0.84	0.36	
LANG				
Spanish	0.84	1.32	0.52	
Other languages	5.86	2.6	0.02	*
TENURE				
Renter	-1.74	0.32	0.00	*
SUM	0.43	0.38	0.26	
LANG*SUM				
Spanish	0.01	0.48	0.99	
Other languages	-2.06	0.91	0.02	*

Source: Table 82 in Wolter et al. (2003).
1,070 interviews.

Table 2. Model for Asians from Wave 2

Independent Variables	Est. Coeff.	Estimated St.Error	p-value	
Intercept	3.35	0.79	0.00	*
AGE				
Low	-1.91	0.49	0.00	*
Medium	-1.12	0.46	0.01	
GRADE				
Low	-0.76	0.41	0.06	*
Medium	-0.39	0.37	0.29	*
TENURE				
Renter	-0.98	0.33	0.00	*
MM'	0.36	0.43	0.40	
CB	-0.66	0.71	0.35	

Source: Table 86 in Wolter et al. (2003).
391 interviews.

Table 4. Model for Native Hawaiians from Wave 2

Independent Variables	Est. Coeff.	Est. St.Error	p-value	
Intercept	0.58	0.83	0.48	
TENURE				
Renter	-1.46	0.43	0.00	*
MM	1.00	0.67	0.14	
CB	-0.43	0.99	0.67	

Source: Table 94 in Wolter et al. (2003).
454 interviews.

*Statistically significant

Table 3. Model for American Indians from Wave 2

Independent Variables	Est. Coeff	Est. St. Error	p-value	
Intercept	1.94	1.11	0.08	*
INCOME				
1st Quartile	-2.84	0.82	0.00	*
2nd Quartile	-2.55	1.41	0.07	*
3rd Quartile	-2.15	1.28	0.10	*
LANG				
Other languages	6.98	0.97	0.00	*
SEX				
Male	-1.30	0.62	0.04	*
MM	1.26	0.68	0.07	*
CB	-0.57	1.22	0.64	

Source: Table 90 in Wolter et al. (2003). n = 77

Table 5. Results of using Wave 2 awareness of PMP to estimate increase in mail returns using models from Wave 2

Race/ethnicity of householder	**Occ. units MOMB (000's)**	**Est. gain (000's)**	**Est. % gain**
White non-Hisp.	78,431	2,894,	3.7%
Hispanic	9,160	592	6.5%
Black non-Hisp.	11,829	694	5.9%
Asian	3,125	14	0.4%
Native Hawaiian	99	6	6.1%
Am. Indian	626	38	6.1%
Remaining races	1,432	24	1.7%
total	104,705	4,265	4.1%

Table 6. Results of using Wave 3 awareness of PMP to estimate increase in mail returns using models from Wave 2

Race/ethnicity of householder	**Occ. units MOMB (000's)**	**Est. gain (000's)**	**Est. % gain**
White non-Hisp.	78,431	3,645	4.6%
Hispanic	9,160	870	9.5%
Black non-Hisp.	11,830	906	7.7%
Asian	3,126	28	0.9%
Native Hawaiian	99	12	12.1%
Am. Indian	627	57	9.1%
Remaining races	1,432	39	2.7%
total	104,706	5,552	5.3%

Note: Remaining races includes mixed race.

Table 7. Race/ethnicity and tenure results of using Wave 3 awareness of PMP to estimate increase in mail returns using models from Wave 2

Race/Ethnicity of Householder	Occ. HU headed by Owners (000s)	Est. % gain due to PMP	Estimated gain due to PMP (000s)	Occ. HU headed by Renters (000s)	Est. % gain due to PMP	Est. gain due to PMP (000s)
White non-Hisp.	57,156	3.3%	1,861	21,275	8.4%	1,785
Hisp. (any race)	4,212	5.1%	216	4,948	13.1%	649
Black non-Hisp.	5,597	5.5%	306	6,233	9.6%	601
Native Hawaiians	46	9.0%	4	54	14.5%	8

Table 8. American Indian results of using Wave 3 awareness of PMP to estimate increase in mail returns using models from Wave 2

American Indian Householders	Occ. HU headed by Men (000s)	Est. % gain due to PMP	Estimated gain due to PMP (000s)	Occ. HU headed by Women (000s)	Est. % gain due to PMP	Est. gain due to PMP (000s)
Income 1st quartile	60	7.4%	4	84	10.8%	9
Income 2nd quartile	54	9.0%	5	46	10.8%	4
Income 3rd quartile	104	10.7%	11	60	9.9%	5
Income 4th quartile	162	6.6%	11	58	2.3%	1

Table 9. Asian results of using Wave 3 awareness of PMP to estimate increase in mail returns using models from Wave 2

Asian Householders	Occupied HU	Estimated % gain due to PMP	Estimated gain due to PMP
Age - young	905,735	0.66%	5,965
Age - middle	1,523,870	0.46%	6,993
Age - older	699,530	0.24%	1,653

Table 10. Categories for increase in probability of a HU mailing back a form using Wave 3 awareness of PMP in models from Wave 2

very high (>= 10%)	high (< 10% & >= 8%)	medium (< 8% & >= 3%)	low (< 3%)
•American Indian women with income in the 1st & 2nd quartile	•American Indian women with income in the 3rd quartile	•American Indian men with income in the 1st & 4th quartile	•American Indian women with income in the 4th quartile
•American Indian men with income in the 3rd quartile	•American Indians with income in the 2nd quartile		
•Hispanic renters	•Black (non-Hispanic) renters	•Hispanic owners	•Asians
•Native Hawaiian renters	•Native Hawaiian owners	•Black (non-Hispanic) owners	•Remaining races
	•White (non-Hispanic) renters	•White (non-Hispanic) owners	

Note: Remaining races includes mixed race.

Table 11. Categories for increase in number of HUs mailing back a census form using Wave 3 awareness of PMP in models from Wave 2

very high (>=1.7 million)	high (<1.7 million & >=600,00)	medium (<600,000 & >=140,000)	low (<140,000)
•White (non-Hispanic) renters	•Black (non-Hispanic) renters	•Black (non-Hispanic) owners	•Asians
•White (non-Hispanic) owners	•Hispanic renters	•Hispanic owners	•American Indians
			•Native Hawaiians
			•Remaining races

Note: Remaining races includes mixed race.