SELECTING VARIABLES FOR POSTSTRATIFICATION AND RAKING

Golam M. Farooque, Inez I. Chen, Bureau of the Census
Golam M. Farooque, Bureau of the Census, Washington, DC 20233

KEY WORDS: Logistic regression; Interaction terms; Raking dimensions; Post Enumeration Survey.

ABSTRACT: This article applies logistic regression models to the 1990 Post Enumeration Survey (PES) data for California and determines the important variables to form alternative poststratifications and raking matrices. The person level indicator variable for capture in the census is used as the dependent variable. This paper finds that age/sex, race/Hispanic origin, tenure, household composition, and urbanicity variables are the most important variables for forming alternative poststratifications and raking matrices. The first order interaction terms of significant independent variables are found insignificant when they are input to the logistic regression models with their main effects.

1. Introduction

For Census 2000, a major goal of the Census Bureau is to reduce the undercount, especially the differential undercount in different segments of the population. Generally, undercounts tend to be higher for minorities, especially, for Blacks, Hispanics, Asian, and nearly all nonowners (Hogan 1993, Robinson et al. 1993). The Bureau has been using Dual System Estimation (DSE) (capture/recapture technique) with post-stratification to produce Census counts at various geographical levels in order to correct for coverage errors. The DSE assumes that the probabilities of enumeration are the same for all members of the population. The past research showed that the probability of being enumerated in the census varies by race, age, sex, tenure, and geographical areas, hence, the homogeneous probabilities' requirement for DSE is not met. A considerable number of studies have been conducted to provide improved estimates of persons missed by both initial enumeration and the PES enumeration (Hogan 1993, Alho, et al. 1993). Yet, a residual heterogeneity remains which may cause correlation bias in the DSE. Alho et al. (1993) and Murly et al. (1997) used logistic regression analysis to select variables to reduce heterogeneity. Currently, the Bureau is conducting research on how to identify a set of variables that will be used to form poststrata for the year 2000.

This paper conducts a preliminary research to identify the important variables, from a set of variables to form alternative poststratifications and raking matrices. It is expected that this research will aid the 2000 postratification research. By applying the logistic regression models, this paper identifies the important variables from the 1990 PES data for California. This research is similar to Haines and Hill (1998) study. However, it incorporated the two-way interaction terms in addition to main effects in the logistic regression models. To determine the statistical significance of the variables, it used global tests and deviance tests in addition to Wald tests.

A literature review is presented in Section 2. Section 3 presents the methodology which includes definitions of variables, logistic regression model, model inferences, and variable selection techniques. The results of logistic regression models are presented in section 4. Section 5 concludes the paper.

2. Literature Review

The Census Bureau has been using the logistic regression modeling as an analytical tool to analyze the effects of some geographic and demographic variables in a multi-variate modeling of a dichotomous dependent variable (Alho et al. 1993, Mulry et al. 1997, Haines and Hill 1998). Haines and Hill applied logistic regression models on race/origin, age/sex, tenure, urbanicity, household composition (HH comp), household marital status (HH marital status), percent nonowners, mail response rates (MRR), percent minority, vacancy rates, household size, and relationship. Using the 1990 PES data for California, they identified that race/origin, age/sex, tenure, urbanicity, MRR, and relationship are important variables for raking matrices.

Hogan (1993) anticipated that many of the 1,392 poststrata adjustment factors would have very high variances and suggested to form fewer poststrata for the 1990 PES data. The adjustment factors were calculated by dividing the estimated true population by the census

The authors are mathematical statisticians in the Decennial Statistical Studies Division. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

count. Assuming homogeneity exists in fewer poststrata, he regressed observed adjustment factors by poststratum on indicator variables such as race, Hispanic origin, age, tenure, census division, place/size category, and some two-factor interaction terms. He forced race, age, and tenure indicator variables in the model and the other variables were selected using their predictive power.

Alho et al. (1993) applied the conditional logistic regression models to the 1990 PES data for minorities to estimate the probability of being captured in a census. The independent variables were both continuous and categorical. They used age, female, Black, Hispanic origin, renter, household size, related, married, % renter, %Black, %Hispanic, %multi-unit, %vacant, four census regions, metropolitan, and some interaction terms in their models. The dependent variable consists of persons who are captured only in the E-sample, only in the P-sample, and in both E and P-samples. They excluded the unresolved cases from the data. Their results reveal that household size, Black, Hispanic, renter and relationship are significant variables, in at least one of the census regions.

Mulry et al. (1997) examined the heterogeneity of census coverage error for small areas. Using a similar set of variables such as used by Alho et al., Mulry et al. built four separate logistics regression models for minorities, non-minorities, owners, and renters.

Using the 1994 National Assessment for Educational Progress (NAEP) data, Wallace and Rust (1996) studied the performance of raking and poststratification. Their estimated regression model consisted of age, race, regions, school types, metro status, and two-way interactions of these variables. Using a SAS regression procedure they developed the initial models. Since SAS procedure assumes simple random sampling, Wallace and Rust further estimated the selected initial models by using WesVarPC software.

3. Methodology

The methodology section comprises several subsections. In section 3.1, the data and variables are discussed. Sections 3.2, 3.3, 3.4, and 3.5 are on logistic regression modeling, odds ratios, model inference, and variable selection techniques, respectively.

3.1 Data and Variables

The study uses the 1990 PES data for California. The PES consisted of E-sample and P-sample. The E-and P-samples are overlapped. That is, both samples are from the same census blocks and housing units within blocks. The P-sample estimates the number of people missed by the original enumeration and the E-samples estimates the number of enumerations that are erroneous. This paper uses the resolved PES data (i.e., E- and P-samples matches, E-sample non-matches, and P-sample non-matches). Alho et at. (1993) argue that logistic regression procedures yield invalid parameter estimates if they are fitted to unresolved data. An unresolved case may arise due to a fictitious individual or the available information about the person is ambiguous.

The dependent variable in the model is a person-level indicator. It is assigned the value 1 for matched persons between E- and P- samples and non-matched E-sample persons, and 0 for non-matched P-sample persons.

The Bureau has been forming poststrata based on age, sex, race/origin, tenure, census region, and urbanization variables. Historically, it is known that these variables are good predictors of the probability of a person being included in the census.

Following previous research, in particular, Haines and Hill (1998), this paper applies the logistic regression to a set of independent variables- 5 categories of race/Hispanic origin (Non-Hispanic White and Other, Black, Non-Black Hispanic, Asian & Pacific Islander, and American Indians on Reservations), 7 categories of age/sex (under 18, 18-29 Male, 18-29 Female, 30-49 Male, 30-49 Female, 50 + Male, and 50 + Female), 2 categories of each of the following variables: tenure(owner or renter), HH comp (easy to enumerate or hard to enumerate), HH marital status (spousal or non-spousal), relationship(related or not related), urbanicity (urban or non-urban), percent owners (low or high), MRR (low or high), percent minority (low or high), and household size (one or (two & more)). These variables seem to be good predictors of the probability of a person being included in the census.

This paper tested two definitions for household composition and urbanicity variables. Household composition 1 is a household level variable and household composition 2 is a person level variable. Urban 1 consists of all urbanized areas against non-urban areas. Urban 2 contrasts large urban areas to all other areas. The results reported in the paper are based on urban 1 because urban 2 was found insignificant. Hence, urban 1 is referred as urban for the rest of the paper.

In addition to the main effects, the following two-factor interactions of significant main effects are also tested: race/Hispanic*age/sex, race/Hispanic*tenure, race/Hispanic*household composition 2, race/Hispanic*urban, age/sex*tenure, age/sex*household composition 2, age/sex*urban, tenure*household composition 2, tenure*urban, and household composition 2*urban. Interactions with household composition 2 are tested only because it is found more significant then household composition 1.

All independent variables are converted to indicator variables. For variables with more than two categories, the number of indicators equal to number of categories - 1; for variables with two categories, 1 indicator variable is created.

3.2 Logistic Regression Model

The model used in this paper is a binary model. This model is used to determine the probability of an individual with a given set of attributes will make one choice rather than the other.

Due to many assumptions required for the ordinary least square method, it cannot be applied to analyze data with binary dependent variable because (a) the relationship between the dependent variable and the predictors are nonlinear, (b) the error term and the predictors are correlated, and (c) the presence of heteroscedasticity in error term.

This study has chosen the logistic regression model over other binary choice models because it has been used in the Bureau's research. This paper develops the following logistic regression model with the situation where the independent variables are dichotomous.

$$(1)\ \Pi(Y|D) = e^{\alpha+\beta D} / (1 + e^{\alpha+\beta D})$$

Where, $\Pi(Y|D) = P(Y = 1|D)$, the probability that the person is captured in the Census for a given set of characteristics and $1 - \Pi(Y|D) = P(Y = 0|D)$, the probability that the person is not captured in the Census for a given set of characteristics. D is a vector of independent indicator variables that defines the characteristics of a person. β is a set of parameters associated with variables, Ds, and α is the intercept.

The parameters α and βs are estimated by applying the maximum likelihood estimation technique on equation (1). Estimation proceeds by finding estimates for α and βs that maximizes the likelihood function (2) for the set of values $y_1, y_2, ..., y_n$ with given probability defined by D_i. n is the sample of independent observations.

$$(2)\ l(\alpha,\beta) = \prod_{i=1}^{n} \Pi(D_i)^{y_i}\left[1 - \Pi(D_i)\right]^{1-y_i}$$

3.3 Odds and Odds Ratios

For large samples, the parameter estimates tend to follow a normal distribution. By substituting the estimated parameters α and βs in equation (1) one can easily calculate the probability of a person captured in the census. Once the probability of a person being captured is known, then the odds of that person being captured in the census can be computed as $O = \Pi(Y|D) / \{1 - \Pi(Y|D)\}$ for a given set of characteristics, D. The odds ratio is the ratio of the odds.

The logit transformation of $\Pi(Y|D)$ is $\log\left[\frac{\Pi(Y|D)}{1 - \Pi(Y|D)}\right]$. With the relationship expressed in equation (1), this transformation can be expressed as an additive function of independent variables such as equation (3).

$$(3)\ \log O \ = \alpha + \sum_{i=1}^{K} \beta_i D_i + \varepsilon$$

Where, log O is the log odds of being captured in the census given the explanatory variables. K is the number of independent variables included in the model. $\hat{\beta}_i$ is the estimated regression coefficient of the ith independent variable. It estimates the change in the log odds of being in categories of interest on the response for a one-unit change in the ith independent variables in the model.

3.4 Model Inferences

The statistical significance of variable(s) is assessed using the Wald-test and deviance test. The Wald test is computed by squaring the standard normal test when $\beta = 0$. It is approximately distributed as chi-square. The deviance test is the likelihood ratio statistic comparing the reduced model to the full model; it is the statistic for testing the hypothesis all parameters that are in the full model but not in reduced model are equal to zero. The global test is conducted to determine the adequacy of the model.

The calculated Wald test statistics, deviance test statistics, and global test statistics are compared with an adjusted critical value of chi-square. The critical values of $\chi^2(\nu)$ are multiplied by a design effect (DE) to reflect the sample design difference between PES and simple random sampling. All models in this paper are estimated by using the SAS PROC LOGISTIC procedure. This procedure assumes that data is produced by simple random sampling. For California, DE = 20.2 was used. The DE is the ratio of the variance of the 1990 PES undercount and the variance under simple random sampling. For the 2000 poststratification research, SUDAAN software will be used and SUDAAN accounts for complex sample design.

3.5 Variable Selection Methods

This paper first determines the significance of main effects and then tested the significance of two-factor interactions of significant main effects. There are several variable selection methods available. However, none of these methods are appropriate for this study because we forced race/Hispanic origin, age/sex, and tenure variables in the model. Historically, these variables were found significant and used to form poststrata. Thus, instead, this paper used the following two variable selection methods to select the important variables. In method 1, age/sex, race/Hispanic origin, and tenure are always forced in the model and the other independent variables tested individually. Haines and Hill (1998) used this method to select the variables.

In method 2, we also forced age/sex, race/Hispanic origin, and tenure variables in the model. However, unlike method 1, in method 2, the variables which were found significant in the previous steps were kept in the model while testing another new variable. A variable is considered significant if the calculated χ^2 exceeds the adjusted critical values of χ^2 at least at 10 percent levels of significance.

4. Results

Table 1 shows the estimated logistic regression results which are obtained by using variable selection method 1. Based on the results of Table 1, we find that household composition 1, household composition 2, urbanicity, and MRR are statistically significant. However, household composition 2 is significant at 1 percent level, household composition 1 and urbanicity variables are significant at 5 percent level, and MRR is marginally significant at 10 percent level. This decision is reached based on the Wald test statistics and deviance tests reported on this Table. As mentioned before, the critical values for these tests are adjusted for a design effect =20.2. Using the Global test, one can also conclude that the model with household composition 2 in presence of age/sex, race/Hispanic origin, and tenure performs better than any other models.

Table 1. Logistic Regression for Method 1 (Bolded Variables were Forced)

Independent Variables	Wald Test	Odds Ratio	Global Test	Deviance Test
1. race, age/sex, tenure	-	-	725.68(11)*	-
2. HH Comp 1	87.41(1)**	1.801	820.16(12)*	94.48(1)**
3. HH Comp 2	174.79(1)*	2.240	916.55(12)*	190.87(1)*
4. HH marital status	25.07(1)	1.467	751.74(12)*	26.06(1)
5. relationship	1.189(1)	0.942	726.87(12)*	1.19(1)
6. urban	91.42(1)**	2.136	805.06(12)*	79.38(1)**
7. % nonowners	4.42(1)	1.116	730.12(12)*	4.44(1)
8. mail response rates	54.10(1)***	0.706	778.18(12)*	52.51(1)
9. % minority	12.65(1)	0.828	738.34(12)*	12.67(1)
10. household size(HHS)	17.36(1)	0.697	741.96(12)*	16.28(1)

Note: On all Tables *, **, and *** mean significant at 1, 5, and 10 percent levels of significance, respectively.

The odds ratios of estimated coefficients of logistic regression models can be used to explain the odds of a person being captured in the census. For example, in Table 1 (model 3), the odds ratio for easy to enumerate persons versus hard to enumerate persons is 2.240. This means that odds of being captured for easy to enumerate persons are about 2.240 times larger than hard to enumerate persons. Similarly, the odds of urban people (model 6, Table 1) being captured in the census is 2.136 times higher than non-urban people.

Table 2 displays the odds ratios, deviance test, and global test statistics for logistic regression models by using the variable selection method 2. A variable is considered significant if its estimated test statistic is less than a pre-specified significance level of 10 percent. The importance of each other variable is evaluated by keeping the previous significant tested variable(s) in the model.

Since the variable selection method 1 indicates that HH comp 1 and HH comp 2 both are significant, two separate models are estimated: one with HH comp 1 and another with HH comp 2. The results are on Table 2. On this Table, the first set of statistics corresponding to each variable represents the test statistics produced from models with HH comp 1 and the second set in brackets under the first set denotes the test statistics obtained from models with HH comp 2. Like variable selection method 1, this method also shows that HH comp and urbanicity variables are significant. However, HH comp 2 is more significant than HH comp 1.

We also test whether different order of entering the variables in the models makes any difference in results. The findings are similar to those of Tables 1 and 2. Results are available from the authors upon request. Thus based on these findings we conclude that race/origin, age/sex, tenure, HH comp 2 and urbanicity are important variables to form poststratification and alternative raking matrices for California.

Next, we proceed to determine the significant two-way interactions for significant main effects. A significant interaction of two variables indicates that these two variables should be cross-classified and used in the same dimension of a raking matrix. For example, if 6 age/sex*tenure interactions are jointly significant by deviance test, then age/sex and tenure would lead to form one dimension for raking by cross-classifying age/sex and tenure. Results from interaction testing also help develop collapsing rules for combining the variables for poststratification. Here, we regressed the dependent variable on age/sex, race/origin, tenure, urban, HH comp 2, and two-factor interaction terms of these main effects.

Table 2. Logistic Regression Results for Method 2 (Bolded Variables were Forced in the Models)[1]

Independent Variables	Wald Test	Odd Ratio	Global Test	Deviance Test
1. race, age/sex, tenure	-	-	725.68(11)*	
2. HH Comp 1	87.41(1)**	1.801	820.16(12)*	94.48(1)**
3. HH Comp 2	174.79(1)*	2.240	916.55(12)*	190.87(1)*
4. HH martial status	5.63(1) [0.22](1)	1.205 [1.038]	825.90(13)* [916.769](13)*	5.74(1) [0.22](1)
5. relationship	2.94(1) [5.54](1)	0.911 [0.88]	823.10(13)* [922.09](13)*	2.94(1) [5.54](1)
6. urban	98.90(1)** [102.42](1)**	2.207 [2.244]	905.66(13)* [1004.97](13)	85.50(1)** [88.43](1)**
7. % nonowners	1.65(1) [2.34](1)	1.070 [1.084]	907.31(14)* [1007.32](14)*	1.65(1) [2.35](1)
8. mail response rates (MRR)	30.77(1) [28.37](1)	0.766 [0.774]	935.79(14)* [1032.78](14)*	30.13(1) [27.81](1)
9. % minority	10.96(1) [10.12](1)	0.838 [0.844]	916.64(14)* [1015.11](14)*	10.98(1) [10.32](1)
9. household size (HHS)	25.95(1) [21.92](1)	0.644 [0.668]	929.64(14)* [1025.35](14)*	23.98(1) [20.38](1)

[1]Test statistics in the square brackets represent the test statatistics obtained from models with HH comp 2.

Table 3 presents the results of interaction terms with their main effects and without the main effects. Column 3 shows that interactions are significant when main effects are excluded from the models. On the other hand, column 2 shows that all interaction terms are insignificant when their main effects are included in the models. This is an indication of high multicollinearity between the interaction terms and their main effects. This may also result from using DE of 20.2 to adjust for critical values for significant tests. However, this should not be a concern for developing poststratification scheme for 2000.

Table 3. Logistic Regression Results for Significant Main Effects and Their Interaction Terms

Interaction Terms	Deviance Test (includes main effects)	Deviance Test (excludes main effects)
race X age/sex	47.68 (24)	585.05 (24)
race X tenure	63.02(4)	529.624(4)*
race X HH comp 2	16.51(4)	608.14(4)*
race X urban	4.98(3)	337.83(3)*
age/sex X tenure	23.39(6)	219.40(6)**
age/sex X HHcomp2	1.45(4)	341.99(4)*
age/sex X urban	13.26(6)	167.98(6)*
tenure X HH comp 2	28.59(1)	579.269(1)*
tenure X urban	1.34(1)	321.65(1)*
HH comp 2 X urban	2.92(1)	527.30(1)*

5. Conclusions

Using the logistic regression analysis, this study identifies that household composition and urbanicity are the only two important variables in addition to race/Hispanic origin, age/sex, and tenure. Urbanicity variable contrasted all urban areas to non-urban areas. Household composition is a person level variable. A person in a housing unit is easy to enumerate if a single person resides in the unit and the person is 50 or older or the unit is occupied by a married couple of 30 years of age or over with 1-5 children of their own under 18. All other persons in the unit are considered hard to enumerate.

Due to different model inference tests and variable selection methods used in this paper, the results are a little different from those of Haines and Hill (1998). Haines and Hill find that relationship and MRR are also important variables. It may be mentioned here that the results may change if one does not force the race/Hispanic origin, age/sex, and tenure variables in the model or different DEs are being used.

References

Alho, Juha M. Mulry, Mary H. Wurdeman, Kent and Kim, Jay (1993), "Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation," *Journal of the American Statistical Association*, **88**, 1130-1136.

Haines, Dawn E and Hill, Joan M (1998), "A Method for Evaluating Alternative Raking Control Variables," *American Statistical Association Meetings.*

Hogan, Howard (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, **88**, No. 423, 1047-1060.

Mulry, Mary H., Davis, Mary C., and Hill, Joan M. (1997), "A Study in Hetergeneity of Census Coverage Error for Samll Areas," *American Statistical Association Proceedings of the Survey Research Methods Section.*

Robinson, Gregory J., Ahmed, Bashir, Gupta, Prithwis D. Woodrow, Karen A. (1993), "Estimation of Population Coverage in the 1990 United States based on Demographic Analysis," *Journal of the American Statistical Association,* **88**, No. 423, 1061-1071

Wallace, Leslie and Rust, Keith (1996), "A Comparison of Raking and Poststratification Using 1994 NAEP Data," Leslie Wallace, West Inc., 584-589.