# POST-STRATIFICATION MODELING FOR HOUSING UNITS DUAL SYSTEM ESTIMATES FOR CENSUS 2000

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

**KEY WORDS:** Logistic Regression; Dual System Estimates; Coverage Correction Factor; Mean Square Error

**ABSTRACT:** Housing coverage correction factors for the Decennial Census 2000 will be based on dual system estimation with post-stratification. The paper performs logistic regression analysis to identify a set of significant post-stratification variables. Based on the Wald test and odds ratios, type of structure, region, type of enumeration area/metropolitan statistical area, occupancy/tenure, race/Hispanic origin of household, and return rates are found important significant variables. Using these variables, the study develops a series of nested post-stratification models. The models are evaluated by comparing the standard error of dual system estimates (DSEs), relative differences of DSEs, standard errors of relative differences, and root mean square errors of DSEs.

## 1. Introduction

The Census Bureau will use Dual System Estimation with post-stratification to compute Housing Unit (HU) Dual System Estimates (DSEs) for Census 2000. It is assumed that housing units which have similar characteristics will have similar census coverage, and thus post-stratification will reduce heterogeneity bias in the model.

HU DSE coverage correction factors will be used in the long form weighting to correct for Census HU coverage and will give some indication on the quality of HU coverage on the Master Address File (MAF) at the time of the Census. Housing unit coverage also affects person coverage. The persons living in housing units will have less chance of being captured in the census if the housing unit is not included on the MAF.

The goals of this paper are twofold. One is to identify a set of significant variables to develop a series of potential post-stratification models. The other is to evaluate a set of nested post-stratification models and two other post-stratification models proposed for Census 2000 Dress Rehearsal (DR) based on relative differences of HU DSEs compared to the full model, standard errors of relative differences, and root mean square errors of HU DSEs. The full model contains all significant variables. This paper also examines HU DSEs for some sub-populations based on the variables in the full model for groups such as occupied HUs, vacant HUs, White and Other households, etc. The preliminary findings show that race/Hispanic origin, region, occupancy/tenure, return rates, and type of enumeration area/metropolitan statistical area are possible post-stratification variables for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.).

The results of logistic regression modeling and post-stratification models evaluation will only provide a guide in selecting post-stratification variables for Census 2000 A.C.E. because the modeling was based on the 1990 Housing Units Coverage Study (HUCS) data and, therefore, may not be directly applicable to define the Census 2000 A.C.E. post-stratification design.

## 2. Background

The 1990 HUCS sample was designed to produce an estimate of the net coverage of housing units within each post-stratum in the form of a dual system estimate (DSE). The dual system estimates rely on classifying each housing unit from the "true" population as being either included in the Census or not, as well as being included in the 1990 HUCS or not. The HUCS was a sample of half of the housing units (2648 block clusters and 80,000 housing units) sampled for the 1990 post-enumeration survey (PES).

The HUCS dual system estimates were computed for 180 post-strata. The post-strata were defined by: region, place type, size of structure, and occupancy/tenure status. The study found a 1.0 percent net undercount for all housing units and a 0.5 percent net undercount for occupied housing units. As expected, the net housing units' coverages in large urban areas and for occupied single housing units were extremely good. The study also found that the owner/renter status of a housing unit is not significant for coverage of housing units (Childers 1990).

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 Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Two post-stratification models, one for MAF evaluation and the other for long form estimation, were proposed for Census 2000 DR. The proposed MAF evaluation model contains type of enumeration area, type of structure, and occupancy/tenure variables. The proposed long form model contains type of enumeration area, race/Hispanic origin of householder and occupancy/tenure variables. The post-stratification variables for these proposed post-stratification models were selected based on the requirements of MAF evaluation and long form estimation.

Research was recently completed at the Census Bureau to select a post-stratification model for person level DSE for Census 2000. The research used logistic regression modeling to select the important post-stratification variables (Haines 1999, Farooque and Chen 1999, Griffin 1999).

## 3. Methodology

This section discusses data, variables, logistic regression, and post-stratification models evaluation techniques.

### 3.1 Data and Variables

The research uses HUCS data. The HUCS sample consists of overlapping E- and P-samples. The P-sample estimates the number of housing units missed by the original enumeration and E-sample estimates the number of original enumerations that are erroneous.

The HUCS samples consist of all PES small block clusters, all PES block clusters with more than 10 percent of the census persons not matching to a PES person and housing unit, all additional PES block clusters with more than 10 percent of the P-sample persons not matching to census persons and housing units, all additional block clusters with more than 10 percent of the P-sample persons matching in surrounding blocks, PES block clusters with high sampling weights, and a sample of remaining PES block clusters.

For logistic regression modeling, the study uses HUCS P-sample resolved and unresolved housing units. A housing unit is called resolved if its match status is determined and either a match probability of 0 or 1 is assigned to it. A housing unit is unresolved if its match status is unknown and an imputed probability between 0 and 1 is assigned to it.

The objective of logistic regression analysis is to estimate the probability of capture in the census. The independence assumptions of DSE imply that the P-sample match status is appropriate for use in logistic regression modeling to estimate census capture regression modeling to estimate census capture probability.

For logistic regression modeling to identify potential HU DSE post-stratification variables, the study included the following variables. These variables are considered likely to be good predictors of the probability of a HU being captured in the Census.

- Race/Hispanic Origin of Household/Vacant (6 categories): (1) Non-Hispanic White Householder, (2) Non-Black Hispanic Householder, (3) Asian & Pacific Householder, (4) Black Householder, (5) American Indian Reservations' Householder, and (6) Vacant HUs
- Type of Family/Vacant ( 4 categories): (1) Family with own children, (2) Family without own children, (3) All other non-family households, and (4) Vacant
- Household Size/Vacant (8 categories): (1) One person, (2) Two persons, (3) Three persons, (4) Four persons, (5) Five persons, (6) Six or Seven persons, (7) Eight or more people, and (8) Vacant HUs
- Region (4 categories): (1) Northeast, (2) Midwest, (3) South, and (4) West
- Region Census Center (12 categories): (1) Boston,
- (2) New York, (3) Philadelphia, (4) Detroit, (5) Chicago, (6) Kansas, (7) Seattle, (8) Charlotte, (9) Atlanta, (10) Dallas, (11) Denver, and (12) Los Angeles and San Francisco
- Type of Enumeration Area (TEA) (3 categories): (1) Tape Address Register and Prelist, (2) List/ Enumerate, and (3) Update/Leave
- Metropolitan Statistical Area (MSA) (3 categories): (1) Large MSA (population 3.5 million and above), (2) Medium MSA (population between 3.5 millions and 500,000), and (3) Small MSA (population 500,000 and below)
- TEA/MSA (4 categories): (1) Mailout/Mailback Large MSA, (2) Mailout/Mailback Medium MSA, (3) Mailout/Mailback Small MSA, and (4) Not Mailout/Mailback MSA
- Occupancy/Tenure (3 categories): (1) Occupied/Owner, (2) Occupied/Renter, and (3) Vacant
- Type of Structure ( 5 categories): (1) Single Unit, (2) Small Multi-Unit: 2-9 HUs, (3) Medium Multi-Unit: 10-49 HUs, (4) Large Multi-Unit: 50+ HUs, and (5) Other Structures
- Return Rates (2 categories): (1) Low RR (<= 25th percentile) and (2) High RR (> 25th percentile) by race/tenure category
- Percent Minority Household ( 2 categories): (1) Low (<= 24 percent) and (2) High (> 24 percent)
- Percent Vacant Housing Unit ( 2 categories): (1) Low (<= 10 percent) and (2) High (>10 percent)

- Percent Mobile Home (2 categories): (1) Low (<=12 percent) and (2) High (>12 percent)

The return rates, percent minority, percent vacant, and percent mobile housing units are tract level variables. The cut off values for return rates is based on the cut off values used in 2000 person DSE post-stratification research. The cuts off values for other three variables are based on odds ratios of weighted HU counts.

The dependent variable in the logistic regression model is a dichotomous variable for capture in the census for a given set of independent variables. For resolved cases, the dependent variable is given a value of 1 for match probability of 1 and a value of 0 for match probability of 0. Each unresolved person record is split into two records. The dependent variable is assigned a value of 1 and a weight of wp for one split record and a value of 0 and a weight of w(1-p) for another record. Here, p denotes imputed match probability for unresolved HUs and w is its corresponding final weight. We split the unresolved records into two categories because the dependent variable in our modeling requires a value of either 1 or 0 based on whether a HU is captured or not captured in the Census.

### 3.2. Logistic Regression Modeling

The Census Bureau has used the logistic regression modeling as an analytical tool to analyze the effects of geographic and demographic variables on a dichotomous dependent variable (Alho et al. 1993, Farooque and Chen 1999). For this study, logistic regression modeling is a mechanism to identify a set of geographic, housing unit, or demographic characteristics which explain census capture probability. A logistic regression model with all main effects is of the form:

$$\log it(P_i) = X_i \beta$$

where, $X_i$ = a vector of covariates of main effects for ith housing unit, $\beta$ = a vector of parameters to be estimated, and $P_i$ = the probability of capture for ith housing unit in the census.

The study uses backward elimination procedure to eliminate the insignificant variables from the model. One insignificant variable is eliminated at a time from the model. The statistical significance of a variable is determined using the Wald test and its corresponding p-value. If a variable satisfies the 10 percent significance levels, then that variable is considered as a candidate for the nested models. At each elimination step, the variable with the highest p-value of the Wald test statistic, is eliminated first (if p-value is greater than 10 percent).

SAS CALLABLE SUDAAN Software is used to determine the significant main effects and to obtain the odds ratios of significant main effects. The standard errors for parameters produced by SAS CALLABLE SUDAAN reflect complex sample design and, therefore, the calculated Wald test statistics would properly reflect the 1990 HUCS sample design.

### 3.3. Evaluation of Post-stratification Alternatives

A second objective of this paper is to evaluate the potential post-stratification models: a series of nested post-stratification models developed from the significant variables and two post-stratification models proposed for DR for census 2000. The post-stratification models will be evaluated in terms of relative differences in HU DSEs of post-stratification models with dual system estimates of the full model, standard errors (SE) of relative differences, differences in HU DSEs, and mean square errors. Similar statistics will be computed for sub-populations based on the variables in the full model.

The following steps are used to compute the above statistics.

Step 1: Compute HU DSE for post-stratification alternative model A as:

$$\hat{DSE}_A = \sum_{i=1}^{\#poststrata} CE_{Ai} * \frac{N_{pAi}}{M_{Ai}}$$

where, i = ith post-stratum, $CE$ = the weighted E-sample correctly enumerated HU estimate, $N_p$ = the weighted P-sample HU estimate, and $M$ = the weighted P-sample matched HU estimate.

Step 2: Compute the relative difference (RD) of HU DSE for model A with HU DSE for full model is defined as:

$$\hat{RD}(DSE_A) = \frac{\hat{DSE}_A - \hat{DSE}_{Full}}{\hat{DSE}_{Full}}$$

The full model is one of the post-stratification alternatives which consists of all significant variables.

Step 3: Apply the jackknife procedure to compute the standard error (SE) of RD for model A as

$$SE(\hat{RD}_A) = \sqrt{\frac{K-1}{K}\sum_{k=1}^{K}(\hat{RD}_{A(k)} - \overline{RD}_A)^2}$$

where $\hat{RD}_{A(k)}$ is the relative difference of $DSE_A$ from all clusters K in the sample except cluster k and $\overline{RD}_A$ is the average of $\hat{RD}_{A(k)}$.

Step 4: Compute the mean square error (MSE) and root mean square error (RMSE) of DSE assuming the HU DSE of the full model is an unbiased estimate of the "truth". The MSE for model A is computed as

$$MSE(DS\hat{E}_A) = (DS\hat{E}_A - DS\hat{E}_{Full})^2 - VAR\left(DS\hat{E}_A - DS\hat{E}_{Full}\right) + VAR(DS\hat{E}_A)$$

$VAR(DS\hat{E}_A - DS\hat{E}_{Full})$ and $VAR(DS\hat{E}_A)$ are computed using the jackknife procedure. Necessary pre-collapsing is done before MSEs are computed.

## 4. Results

This section is divided into two sub-sections. In Section 4.1, the results from logistic regression modeling are presented. The results of post-stratification models evaluation are presented in Section 4.2

### 4.1. Logistic Regression

The logistic regression modeling is done for occupied housing units. Since the TEA/MSA variable is created by using TEA and MSA variable, we have modeled for TEA/MSA variable only. Also, since it is suspected that region and regional census center (RCC) and type of family and household size are correlated, we have fitted five different models:(1) excluding RCC and household size, (2) excluding RCC and type of family,

Table 1: Test Statistics on Main Effect Modeling (Occupied Housing Units)

| Variable | DF | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Wald Test | Wald Test P-value | Wald Test | Wald Test P-value | Wald Test | Wald Test P-value | Wald Test | Wald Test P-value |
| Race/Hispanic Origin Household | 4 | 24.031* | <=0.00 | 22.661* | 0.0001 | 20.714* | 0.0003 | 19.515* | 0.0006 |
| Type of Structure | 4 | 80.688* | <=0.00 | 81.716* | <=0.00 | 76.412* | <=0.00 | 77.333* | <=0.00 |
| Occupancy/ Tenure | 1 | 8.762* | 0.0030 | 9.766* | 0.0017 | 8.580* | 0.0033 | 9.578* | 0.0019 |
| TEA/MSA | 3 | 8.593* | 0.0353 | 8.428* | 0.0379 | 9.823* | 0.0201 | 9.672* | 0.0215 |
| Return Rate | 1 | 8.028* | 0.0046 | 7.954* | 0.0047 | 7.127* | 0.0075 | 7.050* | 0.0079 |
| Percent Minority | 1 | 0.556 | 0.4555 | 0.601 | 0.4380 | 0.121 | 0.7270 | 0.137 | 0.7106 |
| Region | 3 | 16.177* | 0.0010 | 16.420* | 0.0009 | - | - | - | - |
| Percent Vacant | 1 | 0.303 | 0.5812 | 0.268 | 0.6044 | 1.308 | 0.2527 | 1.233 | 0.2667 |
| Percent Mobile | 1 | 3.279* | 0.0701 | 3.376* | 0.0661 | 4.370* | 0.0365 | 4.491* | 0.0340 |
| Regional Census Center | 11 | - | - | - | - | 28.376* | 0.0028 | 28.324* | 0.0028 |
| Type of Family | 2 | 14.694* | 0.0006 | - | - | 15.066* | 0.0005 | - | - |
| Household Size | 6 | - | - | 22.875* | 0.0008 | - | - | 22.116* | 0.0011 |

*indicates significant variables based on the 10 percent levels of significance.
-indicates the variable was not included in the model.

(3) excluding region and household size (4) excluding region and type of family, and (5) with all variables. The odds ratios, Wald test statistics and their p-values are obtained for these models.

The odds ratios are used to determine the hierarchy of significant variables. They are not included in the paper, are available from the authors. Odds ratios are defined as the ratio of the odds of capture for two levels of an independent variable. For a given variable, each odds ratio was computed with respect to the same reference category.

Table 1 presents results of modeling for models 1 - 4. We find that percent vacant and percent minority household are statistically insignificant based on the 10 percent significance criterion. Region, RCC, type of family, and household size are significant when they are modeled separately as shown on Table 1. The logistic regression model with all variables shows that region and type of family variables are insignificant when RCC and household size are included in the model. The results of model (5) with all variables are not included in the paper, but are available from the author.

Thus, based on the logistic regression modeling results we find that race/Hispanic origin of householder, occupancy/tenure, type of structure, TEA/MSA, region, RCC, return rates, percent mobile home, type of family, and household size are significant variables. concerns. RCC cannot be used because given the large of number categories it has, it will yield fewer observations in each post-stratum cell. Also, since the composition of type of family and household size may change due to some missclassifications between E-sample and P-sample, we will not use them in post-stratification models. Although percent mobile home and tenure have almost the same odds ratio, percent mobile home cannot be chosen over tenure because politically tenure is more important the percent mobile home.

Based on odds ratios we ranked the remaining significant variables, in the order of importance, most to the least, as type of structure (T_Struct), race/Hispanic origin (Race), TEA/MSA, region, return rates (RR), and occupancy/tenure (O_Tenure). We formed eight nested post-stratification models using these variables and forced the TEA/MSA, race/Hispanic origin, and occupancy/tenure variables into all nested models. We did not forced the most important variable, type of structure, in all models because this variable was not collected on the Census 2000 census forms and must be derived from the Decennial Mastered Address File.

Table 2. Comparison of Post-stratification Alternatives: National Level Estimates

| Alternative Post-Stratification Models | SE(DSE) | RMSE | RD(DSE) | SE(RD(DSE)) |
|---|---|---|---|---|
| 1. TEA/MSA*Race*O_Tenure*T_Struct*Region*RR | 251,320 | 251,320 | 0.00000 | 0.000000 |
| 2. TEA/MSA*Race*O_Tenure*T_Struct*RR | 248,186 | 244,823 | 0.000006 | 0.000396 |
| 3. TEA/MSA*Race*O_Tenure*T_Struct*Region | 247,038 | 243,764 | 0.000018 | 0.000391 |
| 4. TEA/MSA*Race*O_Tenure*Region*RR | 256,060 | 254,608 | -0.000573 | 0.000632 |
| 5. TEA/MSA*Race*O_Tenure*T_Struct | 247,506 | 240,526 | 0.000017 | 0.000568 |
| 6. TEA/MSA*Race*O_Tenure*RR | 256,489 | 253,981 | -0.000568 | 0.000667 |
| 7. TEA/MSA*Race*O_Tenure*Region | 255,742 | 252,869 | -0.000559 | 0.000672 |
| 8. TEA/MSA*Race*O_Tenure | 256,420 | 252,600 | -0.000569 | 0.000713 |
| 9. MAF: TEA*T_Struct*O_Tenure*Region | 262,546 | 253,589 | 0.000039 | 0.000663 |
| 10. Long Form: TEA*Race*Tenure*Region (**no vacant**) | 167,844 | 153,046 | -0.000086 | 0.000752 |
| 11. Long Form: TEA*Race*O_Tenure*Region | 270,204 | 255,999 | -0.000159 | 0.000856 |

We computed the summary statistics for total population discussed in Section 3.3 for nested models and two other proposed DR post-stratification models. The statistics are presented on Table 2. Similar statistics were computed for states and sub-populations: 5 race/Hispanic origin, 3 occupancy/tenure, 4 TEA/MSA, 4 region, and

### 4.2. Post-stratification Models Evaluation

Although a number of variables are significant, we cannot use all of them to form a series of nested post-stratification models because of some implementation 5 type of structures categories. These results are not presented in the paper, but are available from the authors.

On Table 2, Models 1-8 are the nested models, 9 is the MAF model, and 10 and 11 are the long form models without vacant and with vacant, respectively. We added the region variable to the two other proposed DR models. In order to reduce variances, all non-single housing unit of type of structure categories were collapsed into one category and we also collapsed region for everyone except the white owners. The results of table 2 show that the SE(DSE), RMSE(DSE), and SE(RD(DSE)) for all models are similar except the long form model 10, and, therefore, they are not very useful to evaluate the post-stratification models. Also, the SE(DSE) and RMSE(DSE) of model 10 are about 60 percent of SE(DSE) and RMSE(DSE) of other models, respectively. It seems that vacant housing units contribute about half the variances of other models.

Thus, based on RD(DSE), we find that, after necessary collapsing to reduce the variances, the model 1(full model) is a reasonable post-stratification model. However, the type of structure variable was not collected for Census 2000. It can only be derived by counting the number of HU of the same Basic Street Addresses of the Decennial Master Address File (DMAF). Considering the possible complexity involved with the derivation of the type of structure variable using the DMAF, we find that the Model 4 is the most reasonable post-stratification model for Census 2000 A.C.E.

### 5. Conclusions

Using the logistic regression analysis, this study identifies that race/Hispanic origin of householder, TEA/MSA, occupancy/tenure, type of structures, return rates, region, regional census center, type of family, household size, and percent mobile home are significant variables. However, considering the implementation concerns with regional census center, type of family, household size, and percent mobile home, the study developed eight nested post-stratification models using the remaining variables: race/Hispanic origin of household, TEA/MSA, occupancy/tenure, type of structures, return rates and region. The summary statistics (presented on Table 2) were computed for the nested models and for the two other proposed DR post-stratification models. It appears that, after necessary collapsing, Model 4 on Table 2 is a reasonable post-stratification model for Census 2000 A.C.E.

### 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.

Childers, Danny R. (1993), "Coverage of Housing in the 1990 Decennial Census", *1990 Decennial Census Preliminary Research and Evaluation Memorandum No. 253, Internal Census Bureau Memorandum.*

Farooque, Golam and Chen, Inez (1999), "Selecting variables for post-stratification and raking," *Proceeding of the Section on Survey Research Methods,* American Statistical Association, Alexandria, Virginia.

Griffin, Richard (September 1999), "Accuracy and Coverage Evaluation Survey: Poststratification Research Methodology", *DSSD Census 2000 Procedures and Operations Memorandum Series Q-5, Internal Census Bureau Memorandum.*

Haines, Dawn (1999), "Accuracy and Coverage Evaluation Survey: Logistic Regression Modeling for Post-stratification Variable Selection", *DSSD Census 2000 Procedures and Operations Memorandum Series Q-6, Internal Census Bureau Memorandum.*