UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

2013 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES #ACS13-RER-12-R1

DSSD 2013 AMERICAN COMMUNITY SURVEY RESEARCH MEMORANDUM SERIES #ACS13-R-02-R1

MEMORANDUM FOR ACS Research and Evaluation Advisory Group

From: Patrick J. Cantwell ***signed 12/20/13***
Chief, Decennial Statistical Studies Division

Prepared by: Don Keathley
American Community Survey Sample Design Branch
Decennial Statistical Studies Division

Subject: Revision: American Community Survey: Sample Representivity for the Nation and Puerto Rico

Attached a revision of the American Community Survey Research and Evaluation report "American Community Survey: Sample Representivity for the Nation and Puerto Rico", which was originally released in April 2013. This report gives an indication of how representative the interviewed occupied housing units in the American Community Survey (ACS) are of the ACS nonrespondents and, therefore, the sampling frame. Representivity is measured using an R-indicator statistic.

Changes from April's release of this document include:

1. A complete revision of Section II.A. – April's release contained factual errors in this section
2. The addition of Section II.D. – it contains a brief of description of the methodology we used for computing standard errors
3. An additional limitation in Section III
4. Revised -2 Log L and Adj R^2 metrics in Tables 3, 6, and 7
5. An additional limitation in Section III

United States Census Bureau

6. Revised -2 Log L and Adj R^2 metrics in Tables 3, 6, and 7
7. Corrections for minimum $\widehat{R}(\hat{\rho})$ values in the model set 2 results tables – the District of Columbia model had lower $\widehat{R}(\hat{\rho})$ values than originally reported
8. The corrections in 4 resulted in some minor corrections to some of the quartiles in Attachment E and rank correlation tables in Attachment F
9. Some clarification in the analysis in Section V.B.
10. The addition of footnote 8
11. The inclusion of standard errors in the tables in the attachments, along with some minor revisions in the results section that are associated with their inclusion in this report
12. Some editorial comments throughout the document

None of these changes altered the general results and analysis from April's release.

If you have any questions about this report, please contact Don Keathley (301-763-2225) or Steven Hefter (301-763-4082).

Attachment

cc: ACS Research and Evaluation Workgroup

D. Griffin (ACSO)	J. Chesnut	T. Tersine
M. Ikeda (CSRM)	K. Cyffka	J. Tancreto
K. Albright (DSSD)	P. Davis	R. Ramirez (POP)
M. Asiala	D. Sommers	

DECEMBER 13, 2013

American Community Survey: Sample Representivity for the Nation and Puerto Rico

Don Keathley and Steven Hefter
Decennial Statistical Studies Division

United States Census Bureau

Intentionally blank

I. Introduction

The American Community Survey (ACS) has experienced a high response rate since full implementation began in 2005. Overall weighted response rates between 2005 and 2011 range from 97.3 percent in 2005 to 98.0 percent in 2009 (U.S. Census Bureau (2012)). These rates take all three modes of data collection into account (mail, telephone, and personal visit). Vacant housing unit addresses are included in these rates as they are interviews for the ACS. See U.S. Census Bureau (2009) for details.

Although these response rates are high, two to three percent of cases still did not respond. In this evaluation we want to determine whether the nonrespondents are categorically different in any way from the respondents, i.e., are the respondents representative of the nonrespondents and, consequently, of their entire sample? Then, since we assume that each yearly ACS sample is representative of the frame from which it was sampled, we can simultaneously answer the question of whether the respondents are representative of their corresponding frame as well.

The primary statistic we use in measuring representivity is the R-indicator. It is a measure of the spread of response propensities (probabilities of a sample case responding in the survey) across both respondents and nonrespondents. We also look at sample completeness ratios for comparison purposes, which are measures of the combined levels of nonresponse and under- or overcoverage.

Our analysis in this evaluation focuses on the United States and Puerto Rico. We estimate sample representivity at the national level as a whole and by various subgroups, e.g., race categories. We anticipate that the methods and results in this evaluation will serve as a springboard for future representivity research, for both the ACS and other surveys.

II. Background

A. R-indicators

Recent years have seen the development of R-indicators. These statistics serve as "indicators" of how well or poorly the respondents of a given survey represent the nonrespondents and, consequently, the population for which the sample represents (we assume that each ACS sample is representative of the sampling frame which, in turn, is representative of the target population). The paper by Skinner, et al. (2009), describes the R-indicators; the paper by Shlomo, et al. (2009) provides a discussion of the statistical properties of the R-indicators; the paper by Schouten, et al. (2009) shows how to apply R-indicators.

Skinner, et al. (2009) and Shlomo, et al. (2009) describe two R-indicators: $R(\rho)$ and q^2, where ρ is a vector of response propensities. We focus on $R(\rho)$ in this paper, due in part to the comment in Schouten, et al. (2009), that "… both indicators lead to similar conclusions about the representativeness of response, although they stem from different objectives," and partly because $R(\rho)$ seems to be the statistic of choice in the literature, e.g., in Schouten.

The R-indicator for the population is defined as

$$R(\boldsymbol{\rho}) = 1 - 2\,S(\boldsymbol{\rho}) \qquad (1)$$

where $\boldsymbol{\rho}$ = vector of response propensities for all units in the population
$S(\boldsymbol{\rho})$ = standard deviation of $\boldsymbol{\rho}$

$$= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2} \qquad (2)$$

where N = population size
i = population unit i
ρ_i = response propensity for sample unit i
$\bar{\rho}$ = average response propensity across all sample units
$= \frac{1}{N}\sum_{i=1}^{N}\rho_i$

$S(\boldsymbol{\rho})$ is in the closed interval [0, 0.5]. This means $R(\boldsymbol{\rho})$ is in the closed interval of [0, 1]. $R(\boldsymbol{\rho}) = 1$ when $S(\boldsymbol{\rho}) = 0$, indicating all units in the population have the same propensity to respond. $R(\boldsymbol{\rho}) = 0$ when $S(\boldsymbol{\rho}) = 0.5$, indicating the maximum variation in response propensities.

Equations (1) and (2) are functions of every unit's true propensity to respond – these propensities are usually unknown in practice. When estimating R-indicators in equation (1) from a sample, the response propensities must usually be estimated as well. Equations (3) and (4) define the sample-based R-indicator and standard deviation.

$$\hat{R}(\hat{\boldsymbol{\rho}}) = 1 - 2\,\hat{S}(\hat{\boldsymbol{\rho}}) \qquad (3)$$

where $\hat{\boldsymbol{\rho}}$ = vector of estimated response propensities for the interviewed and noninterviewed sample units from a survey
$\hat{S}(\hat{\boldsymbol{\rho}})$ = standard deviation of $\hat{\boldsymbol{\rho}}$

$$= \sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i\,(\hat{\rho}_i - \hat{\bar{\rho}})^2} \qquad (4)$$

where N = population (frame) size
n = sample size
i = sample unit i
d_i = design weight for sample unit i
$\hat{\rho}_i$ = estimated response propensity for sample unit i
$\hat{\bar{\rho}}$ = average estimated response propensity across all sample units
$= \frac{1}{N}\sum_{i=1}^{n} d_i\hat{\rho}_i$

The design weight d_i we used in our computations was the ACS baseweight (BW), where each sample unit's BW is the inverse of its overall probability of selection for sample. We used $\sum_{i=1}^{n} d_i$ in place of N in equation (4).

The passage in Schouten, et al. (2009) above refers to an $\hat{R}(\hat{\boldsymbol{\rho}})$ that is adjusted for bias due to sampling. The $\hat{R}(\hat{\boldsymbol{\rho}})$ and $\hat{S}(\hat{\boldsymbol{\rho}})$ in equations (3) and (4) are unadjusted for this bias[1]. We used equations (3) and (4) due to the large sample sizes in the ACS. As a result, the $\hat{S}(\hat{\boldsymbol{\rho}})$ values we computed are in the left-open interval (0, 0.5]. This means $\hat{R}(\hat{\boldsymbol{\rho}})$ is in the right-open interval of [0, 1). This result is due to sampling variation in the estimated response propensities (Shlomo, et al. (2009)).

We estimated response propensities for ACS sample housing units in American Indian areas for the sample years 2007 through 2011 combined. We made these estimates using logistic regression models. The general form of the model is

$$\hat{\rho}_i = e^{g(\mathbf{x_i})} / \left(1 + e^{g(\mathbf{x_i})}\right) \qquad (5)$$

where $g(\mathbf{x_i})$ is a linear regression function, i.e., $\beta_0 + \beta_{1i}x_{1i} + \ldots + \beta_{ki}x_{ki}$, where k is the number of regressors in the model.

When transformed via a natural logarithm, $g(\mathbf{x_i})$ in equation (5) becomes

$$g(\mathbf{x_i}) = \ln\left[\frac{\hat{\rho}_i}{1-\hat{\rho}_i}\right] \qquad (6)$$

The regressors are variables for which all responding and nonresponding sample units have a value. These variables are referred to as sample-based auxiliary information in, e.g., Skinner, et al. (2009). We assume that this information comes from one or more sources external to the survey in question, such as administrative record data. Regressors were chosen that we found to have a strong correlation with the survey's response propensities. We chose the variables listed in Table 1 as the regressors.

Most of the regressors are unit-level variables from the 2010 Census Hundred-Percent Detailed File (HDF) for housing units, while two come from edited Master Address File (MAF) extracts and one from a Geography Division-supplied file (see Attachment A for file descriptions). Note that the variables from the edited MAF extracts are design variables – we used these instead of the geography from the HDF because we wanted to capture state and county locations of ACS sample units at the time when they were selected for sample.

We ran standard weighted stepwise logistic regressions to determine which of the regressors are significant and to help us decide which variables are worth keeping in the model[2]. Our weights were the design weights (d_i) from above. The dependent variable is a binary response indicator (RI), where $RI_i = 1$ if ACS sample unit i responded and 0 if unit i did not respond.

[1] Shlomo, et al. (2009) indicates that biases would be downward, meaning the adjusted $\hat{R}(\hat{\boldsymbol{\rho}})$ values would be higher than their unadjusted counterparts.

[2] This includes computations of standard errors for parameters, i.e., we did not use the successive difference replication method that the ACS uses for its estimates.

B. Sample Completeness Ratios (SCR)

An adjunct to $\widehat{R}(\hat{\rho})$ is the sample completeness ratio (SCR – see Albright and Starsinic (2002)). It is the ratio of the sum of the baseweights (design weights) of the responding sampled units in the survey divided by an independent count or control. These weights take personal interview subsampling into account. The general equation for an SCR is

$$SCR = (\sum_r d_r)/(N) \qquad (7)$$

where r = ACS respondent r
d_r = design weight for ACS respondent r
N = independent total

We compute SCRs at the national level and at the sub-national level for given variables (most of the regressors in Table 1). The general equation we use in this evaluation is

$$SCR_{c,HDF} = (\sum_r d_{c,HDF,r})/(N_{c,HDF}) \qquad (8)$$

where c = category/characteristic c (category variable value)
HDF = source of auxiliary data
$SCR_{c, HDF}$ = SCR for category c, where the source for category c classifications is the HDF
$d_{c, HDF, r}$ = design weight for ACS respondent r that matched to the HDF, in category c
$N_{c, HDF}$ = count of cases on the ACS frame that matched to the HDF, in category c

For example, if we computed the national SCRs for each householder's race category, then each ACS respondent's value for race will come from the HDF and the independent controls will be the counts of householders on the HDF for which a match could be made to the ACS frames for each race category – the equation is

$$SCR_{race\ c,\ HDF} = (\sum_r d_{race\ c,HDF,r})/(N_{race\ c,HDF}) \qquad (9)$$

SCRs show the proportions of the universe/frame that is represented by the respondents, before any adjustments (e.g., for nonresponse) are made to the respondents' design weights, i.e., it is not an indicator of sample respondent representivity. What they do indicate is the magnitude of nonresponse, under- or overcoverage, or both that are present in the sample. An SCR = 1 is the ideal situation – it means 100 percent coverage and, potentially 100 percent response. Any deviation from one indicates the presence of nonresponse, under- or overcoverage, or both.

C. R-Indicators and SCRs

The best-case scenario is when the R-indicator is just less than one and the SCRs are equal to one. This would show almost perfect sample representativeness combined with 100 percent coverage, response, or both. We continue to assume in the research that the sample is representative of the frame.

Lower-valued R-indicators indicate some degree of a lack of sample representativeness. Should the SCRs be close to or equal to one, however, then this lack of representativeness might not be an issue except for a small segment(s) of the population, e.g., an age group.

R-indicators close to one show good representativeness of the respondents relative to the nonrespondents and the frame. If SCRs are relatively small, however, then the frame (and therefore the sample) might not necessarily be representative of the target population.

The worst-case scenario is when both lower-valued R-indicators and relatively small or large SCRs occur. This result would indicate both a certain lack of sample representativeness combined with the possibility of the frame not being representative of the target population.

D. Standard Error Estimates

The R-indicators and SCRs are both estimates based on samples (the ACS in our case). This means they are both subject to sampling error. The ACS uses the successive difference replication (SDR) method for computing standard errors for its estimates – we do the same for the R-indicators and SCRs. The general SDR equation is

$$SE = \sqrt{\frac{4}{80}\sum_{i=1}^{80}\left(X_i - X\right)^2} \qquad (10)$$

where X_i = estimate (R-indicator, SCR) from replicate sample i, $i \in \{1, \ldots, 80\}$
X = estimate (R-indicator, SCR) from the base sample

Each ACS sample unit has a set of eighty replicate factors. We multiplied every sample unit's final baseweight by each of its replicate factors, resulting in eighty replicate samples. For the R-indicators, we ran each replicate sample through the best-fitting models below, resulting in eighty sets of R-indicators (total and by category, for the variables in Table 1). We then applied equation (10) to obtain the standard errors for the R-indicators. For the SCRs, we first computed eighty sets of numerators by summing each replicate sample's adjusted baseweights, across all units and by variable category (from Table 1). Then we computed the SCRs by dividing the replicate sample numerators by the appropriate denominators (the denominators

are from the base sample for each replicate SCR). We then applied equation (10) to get standard errors for the SCRs.

See Ash (2011) for details on the SDR method.

III. Limitations

One limitation is that our analysis was restricted to just those ACS interviews from occupied housing unit addresses and non-interviews (eligible cases) that matched to a housing unit record on the HDF. Approximately 5.2 percent of eligible cases did not match to the HDF (480,233 of 9,253,859 cases)[3]. Additionally, if the average weighted response rates are different between the matching versus nonmatching cases, then this difference could have an impact on the SCRs we actually observe had all eligible cases matched to the HDF.

Another limitation is that not all of the eligible cases that matched to an HDF record had entries for the variables on the HDF, i.e., many were vacant housing units in the 2010 Census. These records comprised approximately 4.3 percent of the eligible cases that matched to the HDF (381,101 of 8,773,626 cases).

If all of the eligible cases that did not match to an HDF record had indeed matched, and if all of the eligible cases had been occupied housing units in 2010, then our results would have been different from those observed, for both the R-indicators and the SCRs.

Another limitation is that the matching was done by MAFID only. MAFIDs might not always refer to the exact same address across time. Had the HDF contained address information, like house number and street name, then matching could have been performed using these variables. This would have potentially resulted in more accurate matching between the files.

One more limitation is that it is possible that the HDF values for the matching ACS sample cases might be different than what was reported in the ACS.

IV. Methodology

A. Input Files, Variables (Regressors)

Table 1 shows the variables we used for our regressors, along with their associated source files. It also shows the source for the dependent variable (STATUS / ACSINT). See Attachment A for descriptions of all of the files mentioned in this section.

[3] The 9,253,859 total excludes matches that were not housing units in the 2010 Census – there were 929 such cases.

We merged various files, including those shown in Table 1, to create the input file for the logistic regression modeling and R-indicator computations. This file contains all of the variables shown in Table 1. Attachment B provides a summary of the process we used to create the final input file.

The codes for each variable that we used are shown in the table in Attachment C. The last column in the table shows the code/category we used as the reference group for the regressor[4].

Table 1. Variables, Source Files

Variable	Description	Source File
BLD	Edited Building Structure Type	2010 Unit HDF *
CLUSTERNUM / SEG_GRP	Segmentation Group Code	File from Geography Division
FIPST	FIPS State Code	Edited Supplemental MAF Extracts
FCNTY	FIPS County Code	Edited Supplemental MAF Extracts
HHLDRAGE	Edited Age of Householder	2010 Unit HDF
HHSPAN	Hispanic or Latino Householder	2010 Unit HDF
HHRACE	Race of householder	2010 Unit HDF
HHT	Household Family Type	2010 Unit HDF
STATUS / ACSINT	ACS Interview Outcome Code	2007-2011 Select Files
TENSHORT	Tenure	2010 Unit HDF

* The 2010 Unit HDF is the housing-unit level data file from the 2010 Census, where the data are edited.

We copied the variable CLUSTERNUM to SEG_GRP, with a recode: CLUSTERNUM = blank became SEG_GRP = 0. This was done for programming purposes, where a blank was not an acceptable value. We recoded STATUS to ACSINT so that ACS interviews and non-interviews had codes of 1 and 0, respectively.

B. Logistic Regression Models

We ran the models shown in Table 2. All were regular stepwise regression models, and all were weighted using the sampled units' design weights (baseweights). The significance level cutoff for inclusion in the model was 0.01. We ran the models using housing unit records for which we had entries for the variables only i.e., for which the housing unit was occupied in the 2010 Census – non-vacants[5].

[4] Reference groups are the levels of the variables in a model against which the parameter estimates for the remaining levels are compared.

[5] The ACS classifies all vacant units as interviews. If we had had information on age, sex, etc. for the householders of these units, at least some of the R-indicator values we observed would have moved closer to 1.

Table 2. Models

Model / Model Set	Description
National Model 1	One national model with FIPST as the only regressor
National Model 2	One national model with main effects only, excluding FIPST and FCNTY
National Model 3	One national model with main effects only, excluding FCNTY
State Model Set 1	One model per state (FIPST) with FCNTY as the only regressor *
State Model Set 2	One model per state (FIPST) with all main effects only
State Model Set 3	One model per state with main effects and 2-way interactions, minus FCNTY

* The District of Columbia is its own county, so the D.C. model uses 2000 Census tract as the regressor in model set 1

C. R-Indicators

Once we completed the logistic regression runs, we used equations (3) and (4) to calculate the values of $\widehat{R}(\hat{\rho})$ from each logistic regression run.

D. Sample Completeness Ratios

We computed SCRs for totals and main effects. Since the logistic regression models and R-indicators are based on 2010 Census occupied housing units only, we compute SCRs for occupied housing units only as well. The numerators are weighted summations from the records in the final input file mentioned in Section III.A. The denominators are counts of matching records between the HDF and the five yearly ACS sample frames (the edited MAF extracts). Matching was on the nine-digit MAFID (the 2007 edited MAF extracts include the twelve-digit MAFIDs only, so we added their nine-digit MAFIDs from the 2010 edited MAF extracts. We matched the 2007 and 2010 extracts on the twelve-digit MAFID).

E. Model-Fit Metrics

Table 3 shows results for the three national models. Tables 6 and 7 show results for state model sets 1 and 2 (we omit results for state model set 3 – see below). The goodness-of-fit metrics are indicators of how well each model fits in comparison to the other models. -2 Log L is -2 times the log-likelihood of the model, where lower values indicate better fits[6].

Adjusted (Adj) R^2 (Nagelkerke (1991)) is the ratio of a generalization of the coefficient of determination (CD) divided by its maximum possible value:

[6] We looked at the Akaike Information Criterion (AIC) as well – we omit this statistic because the values we observed for all models was approximately the same as that for -2 Log L.

$$\text{Adj } R^2 = R^2 \,/\, \text{Max } R^2 \qquad (8)$$

where R^2 = a generalization of the CD (Cox and Snell (1989))

$$= 1 - \left(\frac{L(0)}{L(\hat{\beta})}\right)^{2/n} \qquad (9)$$

Max R^2 = maximum R^2 value

$$= 1 - (L(0))^{2/n} \qquad (10)$$

$L(0)$ = log-likelihood of the intercept-only model
$L(\hat{\beta})$ = log-likelihood of the specified model
n = weighted sample size

The reason for using Adj R^2 is that its maximum value is one, whereas it is less than one for R^2 (both statistics can take on minimum values of zero). Higher values of Adj R^2 indicate a better model fit.

The receiving operating characteristic (ROC) curve is a plot of proportions of true positive predictions (sensitivity) on the y-axis versus proportions of false positive predictions (1 – specificity) on the x-axis, at various sensitivity levels. The sensitivity levels range from zero to one, inclusive. Each level indicates the proportion of true positives that are classified as positives by the model, given a probability cut-off point; in our case, positives are interviews and negatives are noninterviews. For example, if the cut-off point is 80 percent, then any case with a predicted probability greater than or equal to 80 percent is classified as a positive (interview, in our case); those with a predicted probability less than 80 percent are classified as a negative (noninterview). The sensitivity level is then the proportion of true positives (interviews) that are classified as positives (interviews) given the 80 percent cut-off point. The false positive (interview) rate associated with a given sensitivity level indicates the proportion of true negatives (noninterviews) that are classified as positives (interviews) given the cut-off point. Thus, the ROC curve for each of our models is a plot of proportions of true interview classifications versus false interview classifications.

The area under the ROC curve indicates how well a model differentiates between true positives (interviews) and true negatives (noninterviews). An area of one shows perfect predictions, or discrimination, in the model – all of the cases that are predicted to be positive at any given sensitivity level are true positives. An area of 0.5 indicates zero discrimination – half of the cases that are predicted to be positive at any sensitivity level are true positives and half are true negatives. As areas increase from 0.5 to 1, the ability of the model to discriminate between true positives and negatives increases. Areas less than 0.5 indicate a negative discrimination, where more than half of cases predicted to be positive are actually true negatives. See Kleinbaum and Klein (2010) for more information on ROC curves.

V. Results

A. National Models

Table 3 shows the goodness-of-fit metrics for each of the three national models; Table 4 shows the $\widehat{R}(\hat{\rho})$ values for each model.

The results in Table 3 show that including all of the variables in the model except FCNTY (national model 3) results in the best fit, with the smallest -2 Log L value (22,848,888) and largest Adj. R^2 value (0.124). Model 3 also has the best ability to predict whether an ACS housing unit address will be an interview, with an area under the ROC curve of 0.688.

Table 3. Summary of Logistic Regression Runs for National Models 1, 2, and 3

			Goodness-of-Fit Metrics		
Model	**Steps**	**Variables in Model**	**-2 Log L**	**Adj R^2**	**Area under ROC Curve**
1	-	FIPST	23,484,126	0.051	0.607
2	7	All	23,168,040	0.088	0.675
3	8	All	22,848,888	0.124	0.688

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

The $\widehat{R}(\hat{\rho})$ values in Table 4 are all fairly close to the maximum of one, indicating good sample representivity. National model 3, which was the best fitting model, had an $\widehat{R}(\hat{\rho})$ value of 0.965. Note there is an inverse relationship between the number of auxiliary variables in a model and the resulting $\widehat{R}(\hat{\rho})$ values (Schouten, et al. (2009)) – this accounts for the results in Table 4.

Table 4. $\widehat{R}(\hat{\rho})$ Values for National Models 1, 2, and 3

Model	**$\widehat{R}(\hat{\rho})$ (S.E.)^**
1	0.980 (0.00017)
2	0.971 (0.00019)
3	0.965 (0.00021)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

^ S.E. = standard error

Because of the results in Tables 3 and 4, we favored national model 3 over national models 1 and 2. This led us to provide $\widehat{R}(\hat{\rho})$ values by main effect category for model 3 only, in Attachment D. The smallest $\widehat{R}(\hat{\rho})$ values for model 3 are 0.941 for segmentation group 4 (economically disadvantaged II, renter skewed) in Table D-7 and 0.948 (Black alone) in Table D-5. All of the other $\widehat{R}(\hat{\rho})$ values are greater than

0.950. These results indicate generally good representivity among the main effect categories.

Table 5 shows SCRs for totals; Attachment D shows SCRs by main effect category. All SCRs are at the national level, i.e., across all levels of geography. All SCRs in the attachment omit vacant housing units.

Table 5. Sample Completeness Ratios for Totals

Regressor	SCR (S.E.)^
Total	0.820 (0.00025)
Total, minus vacants	0.874 (0.00027)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

^ S.E. = standard error

The national SCR, minus vacant units, in Table 5 is 0.874. The majority of SCRs in Attachment D are between 0.8 and 0.9, with two less than 0.8 (SCR = 0.543 for Other building structure types in Table D-1; SCR = 0.777 for 0-24 year old householders in Table D-2) and five being over 0.9 (maximum SCR = 0.927 for no segmentation group code areas in Table D-7). These values indicate some undercoverage, nonresponse, or both, at the national level, overall and across the main effect categories.

B. State Model Sets

Tables 6 and 7 give distributions of the goodness-of-fit metrics for state model sets 1 and 2, with the exception of dividing -2 Log L by the sample size for each state (resulting in per-sample unit -2 Log L averages). We omitted state model set 3 due to the lack of validity of many of the individual models (29 of the 52 models had questionable validity[7]). The headings in the tables refer to the minimum (Min), 25th percent quartile (P25), median (Median), 75th percent quartile (P75), maximum (Max), and average (Average) values of the metrics across all state models.

The distributions, with two exceptions, favor model set 2 as it has consistently lower -2 Log L / n values and larger Adj. R^2 and ROC curve area values each for each quartile (the exceptions are the maximum -2 Log L / n and Adj. R^2). These results are the same for each state as well, except for the District of Columbia (D.C.) models (not shown).

[7] The models with questionable validity had enough of a lack of fit that their predictive abilities were not usable.

Table 6. Goodness-of-Fit Metric Distributions for State Model Set 1

Good-of-Fit Metrics	Min	P25	Median	P75	Max	Average
-2 Log L / n	0.9	1.8	2.4	3.4	6.4	2.7
Adj R^2	0.007	0.034	0.057	0.090	0.476	0.071
Area under ROC Curve	0.522	0.583	0.608	0.645	0.725	0.613

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

Table 7. Goodness-of-Fit Metric Distributions for State Model Set 2

Good-of-Fit Metrics	Min	P25	Median	P75	Max	Average
-2 Log L / n	0.9	1.8	2.4	3.4	6.3	2.6
Adj R^2	0.066	0.100	0.126	0.162	0.505	0.143
Area under ROC Curve	0.646	0.673	0.687	0.707	0.753	0.692

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

None of the models in state model set 1 were of questionable fit, while the Montana and Ohio models were of questionable fit in state model set 2.

All variables were retained in every model in state model set 1, while thirty-seven of the fifty-two models in state model set 2 contained all variables. The fifteen exceptions in model set 2 were Arkansas, Florida, Illinois, Maine, Mississippi, Missouri, New York, North Carolina, and Oklahoma (HHSPAN not included); Wyoming (BLD); Puerto Rico (SEG_GRP); Vermont (TENSHORT); Nevada (BLD, HHSPAN); Washington, D.C. (HHSPAN, SEG_GRP); and West Virginia (HHSPAN, TENSHORT).

Metrics for the District of Columbia model in state model set 1, which used 2000 Census tract in lieu of county as the regressor, are: -2 Log L / n = 6.4, Adj R^2 = 0.426, and ROC curve area = 0.725. The values for all three metrics are the maximums in Table 7 for their ranges – the next largest -2 Log L / n, Adj R^2 and ROC curve areas are 4.9, 0.182, and 0.700, respectively. The result is similar for model set 2, as the -2 Log L / n, Adj R^2, and ROC curve area values for the District of Columbia are the maximum. The next largest -2 Log L / n, Adj R^2 and ROC curve areas in model set 2 are 4.7, 0.322, and 0.745, respectively. These values are also the reason the average is higher than the median for all three statistics in both sets.

The data above show that, in spite of the two questionable fits in state model set 2, the models in this set generally have better predictive ability than their counterparts in state model set 1.

Table 8 shows distributions of $\widehat{R}(\hat{\rho})$ values across the two state model sets. The values, with one exception (District of Columbia in both sets) are all greater than 0.9, indicating good sample representivity for the individual state models. As in the previous section, the $\widehat{R}(\hat{\rho})$ values are generally lower in state model set 2, which has the larger number of auxiliary variables in its models (Schouten, et al. (2009)).

Table 8. $\widehat{R}(\hat{\rho})$ Value Distributions for State Model Sets 1 and 2

Model Set	Min	P25	Median	P75	Max	Average
1	0.890	0.976	0.982	0.985	0.992	0.977
2	0.875	0.956	0.967	0.973	0.986	0.963

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

Because of the foregoing results, we preferred the models in state model set 2 (the two questionable fit models notwithstanding). This led us to include distributions of $\widehat{R}(\hat{\rho})$ values by main effect category for set 2 only, in Attachment E. The $\widehat{R}(\hat{\rho})$ values across the models for state model set 2 are generally above 0.900, and the minimum P25 (25^{th} percentile) value was 0.939. These results indicate good representivity for the majority of the state model, total and by main effect.

Table 9 shows distributions of SCRs for totals, by state. Attachment E shows distributions of SCRs across states, by main effect category. All SCRs in Attachment E omit vacant units.

Table 9. Distribution of Sample Completeness Ratios for Totals, Across States (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
Total	0.718	0.788	0.820	0.845	0.873	0.815
Total, minus vacants	0.819	0.852	0.879	0.891	0.936	0.874

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

The SCRs in Table 9 and those in the tables in Attachment E indicate varying levels of nonresponse and undercoverage. The majority of the state-level SCRs in Table 9 (Total, minus vacants) are in the 0.8 and 0.9 range. Ranges of SCRs in Attachment E vary, depending on the main effect. Some, like the SCRs for owner-occupied units in Table E-4B, with a range of 0.849 to 0.965, exhibit relatively good completeness (response and coverage). Others, like 0-24 year old householders in Table E-2B, with a range of 0.654 to 0.886, indicate completeness levels that are not quite so good. These results show that undercoverage, nonresponse, or both, exist to varying degrees across both states and main effect categories.

Table 10 and the tables in Attachment F show Spearman rank correlations for state model set 2 $\widehat{R}(\hat{\rho})$ values versus SCR values. Our motivation for this analysis was to give us an indication if any efforts to improve one statistic would have the effect of improving the other as well.

All but two of the correlations are positive, and neither of the negative correlations were significant at the ten-percent level of significance (both are in Table F-1). Twenty-four of the positively correlated ranks, including that for the total in Table 10, are statistically significant at the ten-percent level of significance. This indicates improvements made in the SCRs (coverage, response, or both) could improve the sample representivity that we have observed in this paper.

Table 10. Spearman Rank Correlations of States for State Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – Total, minus Vacant Units

Regressor	Rank Correlation	P-Value
Total, minus vacants	0.302	0.029

VI. Conclusions

With one exception, $\widehat{R}(\hat{\rho})$ values for national model 3 (Table 4 and the those in Attachment D) are all greater than 0.940 (segmentation group 4 in Table D-7 has $\widehat{R}(\hat{\rho}) = 0.941$, but it was not statistically significantly greater than 0.940). Most of the $\widehat{R}(\hat{\rho})$ values from state model set 2 are greater than 0.900 (Table 7 and those in Attachment E; 1,844 of the 1,979 $\widehat{R}(\hat{\rho})$ values were significantly statistically greater than 0.900). These results indicate that, for the most part, the ACS respondents that were input into the logistic regression models are representative of the non-respondents. In turn, this indicates that these respondents are representative of the parts of the frame from which they were selected (not all ACS interviews and non-interviews in the 2007 to 2011 period were input into the models; see limitations section).

$\widehat{R}(\hat{\rho})$ values for the main effects indicate differences in representivity between main effect groups in Attachments D and E. Some of the differences are expected, e.g., $\widehat{R}(\hat{\rho})$ values for owner-occupied units are mostly higher than those for renter-occupied units[8] (Tables D-4 and E-4A). Even so, the differences generally do not appear to be large enough to be alarming. An example of an exception to this is the $\widehat{R}(\hat{\rho})$ value of 0.795 for some other race alone in Table E-5A (Vermont). This value might be of some concern when compared to the other $\widehat{R}(\hat{\rho})$ values.

SCR values in Tables 5 and 9 and in Attachments D and E indicate general undercoverage, nonresponse, or both at the national level (total and by main effect category) and at the

[8] The national and state-level differences of owner minus renter $\widehat{R}(\hat{\rho})$ values were all positive, but nine of the state-level differences were not statistically significant.

state level relative to the 2010 Census. Since overall ACS response rates are high, the SCR values in Tables 8 and 9 are probably influenced more by undercoverage than by nonresponse. This means there are parts of the target population for which the levels of sample representivity might not apply. An extreme case is the minimum SCR = 0.249, for Native Hawaiian / Pacific Islanders (NHPI) in West Virginia, in Table E5A; the $\widehat{R}(\hat{\rho})$ for this state × race category is one. The NHPI respondents from the model perfectly represent the NHPI nonrespondents (if any) and the frame from which they were selected. But, the SCR is so low that a large proportion of the NHPI target population in West Virginia is not being represented by the respondents nor the frame and could be systematically different from the ACS respondents.

The positive correlations in state-level $\widehat{R}(\hat{\rho})$ vs SCR ranks show evidence of higher $\widehat{R}(\hat{\rho})$ values being associated with larger SCRs. Improvements in response rates, coverage, or both would have the effect of improving the already good sample representivity.

VII. Future Research

Future research could include the use of ACS data, auxiliary information from other sources, or both. Examples of other auxiliary information sources are the Census Bureau's planning database and Internal Revenue Service records. We would potentially have a higher proportion of ACS sample cases with complete auxiliary information from alternate sources than we did for this analysis.

We could conduct this research for subsets of the ACS samples, e.g., ACS data collection mode and by ACS sampling stratum. It is possible that representivity could fluctuate between modes or strata, or both.

Other research could explore the use of the bias-adjusted $\widehat{R}(\hat{\rho})$ and the q^2 R-indicators mentioned in Schouten, et al. (2009), as comparisons to the results in this report. We could also compare the standard errors we computed with those from a Taylor linearization method from the literature, for comparison purposes.

Some additional research could include an examination of why we observed outliers in the results, e.g., why West Virginia had such a low SCR for Native Hawaiian / Pacific Islanders. Matching the ACS and Census records on address information, while more involved, would allow us to compare the results of this matching with the matching we did for this evaluation (by MAFID). We could compute R-indicators across time, e.g., on a yearly basis, as a monitoring device.

VIII. References

Albright, K. and Starsinic, M. (2002), "Coverage and Completeness in the Census 2000 Supplementary Survey," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 3,345-3,349

Ash, S. (2011), "Using Successive Difference Replication for Estimating Variances," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 3,534-3,548

Boone, T. (2008), "Segmenting the Population for the 2010 Census Integrated Communications Campaign," C2PO 2010 Census Integrated Communications Research Memoranda Series Number 1, available at http://www.census.gov/2010census/partners/pdf/C2POMemoNo_1_10-24-08.pdf, last accessed in December 2013

Cox, D.R. and Snell, E.J. (1989), *The Analysis of Binary Data (2nd Edition),* London: Chapman and Hall

Keathley, D. and Hefter, S. (2013), "Revision: American Community Survey: R Indicators for American Indian Areas," 2013 American Community Survey Research and Evaluation Report, Series Number ACS13-RER-13-R1

Kleinbaum, D. and Klein, M. (2010), *Logistic Regression (3rd Edition),* New York: Springer

Nagelkerke, N.J.D. (1991) "A Note on a General Definition of the Coefficient of Determination," *Biometrika,* 78, 691-692

Schouten, B., Morren, M., Bethlehem, J., Shlomo, N., Skinner, C. (2009), "How to use R-indicators?", Work Package 4, Deliverable 3, RISQ Project, 7th Framework Programme (FP7) of the European Union, available at http://www.risq-project.eu/papers/RISQ-Deliverable-3.pdf, last accessed in December 2013

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J., Zhang, L. (2009), "Statistical Properties of R-indicators," Work Package 3, Deliverable 2.1, RISQ Project, 7th Framework Programme (FP7) of the European Union, available at http://www.risq-project.eu/papers/RISQ-Deliverable-2-1-V2.pdf, last accessed in December 2013

Skinner, C., Shlomo, N., Schouten, B., Zhang, L., Bethlehem, J. (2009), "Measuring Survey Quality through Representativeness Indicators using Sample and Population Based Information", *Paper presented at the NTTS Conference, 18-20 February 2009, Brussels, Belgium,* available at http://www.risq-project.eu/papers/skinner-shlomo-schouten-zhang-bethlehem-2009-a.pdf, last accessed in December 2013

U.S. Census Bureau (2009), "(ACS) Design and Methodology," available at http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf, last accessed in December 2013

U.S. Census Bureau (2012), "Response Rates – Data": http://www.census.gov/acs/www/methodology/response_rates_data/, initially accessed in August 2012

Table A. Input, Output Files

File	Description
ACS Sample File	File used as input to the logistic regression models and for the numerators in the SCR equations
Edited MAF Extracts (EDMAF)	Edited MAF extracts that have been through ACS edits and code assignments; used as inputs for ACS sampling.
EDMAF-HDF Match File	Sample-year files containing matching records between the edited MAF extracts for the given year and the HDF. Used to compute the denominators in the SCR equations.
Geographic Reference File – Codes	Files that contain block-level geographic codes, e.g., codes for legal/statistical area descriptions*
Geographic Reference File – Names	Files that contain names for the geographic entities in the codes files, except block and "filler" codes*
Hundred-Percent Detail File (HDF)	A file containing edited characteristics and records for all households in the 2010 Census. The data have also been through a disclosure avoidance and tabulation geography application.
Sample Delivery File	Final sample files sent to the American Community Survey Office, as inputs to their sample control system. They are subsets of the second-stage sample files, containing valid records only.
Segmentation Group File	A tract-level file containing segmentation group (CLUSTERNUM) codes for each applicable tract.
Select File	Files that contain the final interview status code for ACS sample housing unit addresses.
Second-Stage Sample File	Output files from the housing unit address sample selection process. They include invalid records.

* used only in Keathley and Hefter (2013).

Summary of the Input File Creation Process

Notes: - *All of the files in this summary are shown in the table in Attachment A.*

- *The variables we captured from the GRFC and GRFN files were needed for the American Indian version of this report only (Keathley and Hefter, 2013)*

We started by creating five files that contained one record per ACS sample housing unit address (ACSSAMP). Each file contained the sampled addresses for one of the five sample years in which we were interested, i.e., 2007 through 2011. Each file was a concatenation of the corresponding year's sample delivery files. There are eight sample delivery files per year, four for the United States and four for Puerto Rico.

We added interview status for each sampled address by matching each ACSSAMP file to its corresponding year's select file, on CMID (nine-digit continuous measurement id). Then we matched the ACSSAMP files to their corresponding second-stage sample files, also on CMID, to pick up each sampled address' baseweight, second-stage sampling stratum, CAPI sub-sampling stratum, reduction measure-of-size, and some geography variables. We matched ACSSAMP to the edited supplemental MAF extracts to pick up FIPST and FCNTY codes.

We then merged the ACSSAMP files to each corresponding year's geographic reference files–codes (GRFC), to pick up the Alaska Native Regional Corporation (ANRC) code (ANRCCE in 2007 and 2008, ANRCFP in 2009, 2010, and 2011) for each sampled address in Alaska that was in an ANRC. We did this matching only for those areas where the American Indian Area code (AINDN; is referred to as AIANHH in the GRFC documentation) was blank (ANRC-only areas), as those with filled AINDN codes were also in Alaska Native Village Statistical Areas, and we wanted to code them as such. Matching was at the block level.

We picked up legal/statistical designation codes (LSADC) for ANRC-only areas by matching the ACSSAMPs to their corresponding year's geographic reference file-names (GRFN). Matching was done on a state -by- ANRCCE/ANRCFP level. We picked up LSADCs for the remaining sampled addresses from the GRFNs as well. Matching for these cases was on state -by- American Indian area -by- tribal subdivision level.

The foregoing process of matching to the GRFC and GRFN files was necessary due to the ANRC values for the LSADC not having been present on any ACS sample files.

Summary of the Input File Creation Process (continued)

The MAF (Master Address File) Tiger Feature Class Code (MTFCC) variable that we needed was already present on the GRFN files for 2009, 2010, and 2011. They did not exist on the 2007 and 2008 GRFNs, however – we used the variables record type (RT) and American Indian Area code (AINDN) from the GRFNs to create MTFCCs for sampled records in these two years.

We obtained the variable CLUSTERNUM by matching the ACSSAMP files to a segmentation cluster file that was created by geography division. This file contained one record per tract. Matching between the two files was at the tract level. Not all tracts are represented on the cluster file, so some records in ACSSAMP did not have a segmentation group code.

Finally, we merged each ACSSAMP to the 2010 Census unit-level hundred-percent detail file (HDF in Table 1). The matching was done on the nine-digit MAFID code. Since MAFIDs in 2007 were the old twelve-digit versions, we needed to match the 2007 ACSSAMP to the 2010 supplemental edited master address files to pick up the 2007 sample's nine-digit MAFIDs prior to matching to the HDF.

The final ACSSAMP files contain only those sample records that matched to the HDF This is because non-matching records from the ACSSSAMPs would not have any data for the majority of the independent variables in the logistic regression models

The actual input file to the logistic regression modeling and R-indicator computations is a concatenation of the final individual year ACSSAMP files.

Table C. Variable Values for the Regressors

Variable	Regressor	Values	Reference Group
BLD	BLD	S = one-family house M = multi-family house T = trailer/mobile home O = other (boat/RV/van, etc)	S
CLUSTERNUM	SEG_GRP	See Attachment G	0
HHLDRAGE	AGE	1 = 0 to 24 2 = 25 to 34 3 = 35 to 44 4 = 45 to 54 5 = 55 to 64 6 = 65 to 74 7 = 75+	2
HHSPAN	HHSPAN	1 = not Hispanic or latino 2 = Hispanic or latino	1
FIPST	FIPST	Two-digit FIPS state codes	01
FCNTY	FCNTY	Three-digit FIPS county codes	001
HHRACE	RACE	1 = White alone 2 = Black alone 3 = Amerind/Alaskan Native alone 4 = Asian alone 5 = Native Hawaiian/pacific islander alone 6 = Some other race alone 7 = Multi-race	1
HHT	HHT	1 = Husband/wife family household 2 = Other family household: male householder 3 = Other family household: female householder 4 = Nonfamily household: male householder, living alone 5 = Nonfamily household: male householder, not living alone 6 = Nonfamily household: female householder, living alone 7 = Nonfamily household: female householder, not living alone	1
STATUS	ACSINT	1 = Interview (ACSINT = 1) 4 = Non-Interview (ACSINT = 0) All other codes were out-of-scope for this evaluation	-
TENSHORT	TENSHORT	1 = Owner-occupied unit 2 = Renter-occupied unit	1

Table D-1. Model 3 $\widehat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by BLD (Edited Building Structure Type) Category

Regressor	$\widehat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
One-family house	0.976 (0.00020)	0.895 (0.00035)
Multi-family house	0.954 (0.00036)	0.823 (0.00072)
Trailer/mobile home	0.979 (0.00037)	0.807 (0.00138)
Other (boat/RV/van, etc.)	0.975 (0.00436)	0.543 (0.01576)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table D-2. Model 3 $\widehat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by AGE (Edited Age of Householder) Category

Regressor	$\widehat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
0 to 24	0.961 (0.00063)	0.777 (0.00180)
25 to 34	0.960 (0.00036)	0.832 (0.00116)
35 to 44	0.963 (0.00036)	0.863 (0.00076)
45 to 54	0.966 (0.00032)	0.884 (0.00078)
55 to 64	0.971 (0.00029)	0.896 (0.00073)
65 to 74	0.976 (0.00030)	0.904 (0.00093)
75+	0.981 (0.00023)	0.905 (0.00102)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table D-3. Model 3 $\widehat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by HHSPAN (Hispanic or Latino Householder) Category

Regressor	$\widehat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
Not Hispanic or latino	0.965 (0.00021)	0.876 (0.00028)
Hispanic or latino	0.967 (0.00047)	0.859 (0.00096)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table D-4. Model 3 $\hat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by TENSHORT (Tenure) Category

Regressor	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
Owner-occupied unit	0.978 (0.00019)	0.904 (0.00035)
Renter-occupied unit	0.957 (0.00032)	0.818 (0.00060)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

^ S.E. = standard error

Table D-5. Model 3 $\hat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by RACE (Race of Householder) Category

Regressor	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
White alone	0.975 (0.00019)	0.884 (0.00031)
Black alone	0.948 (0.00047)	0.825 (0.00094)
American Indian / Alaska Native alone	0.963 (0.00102)	0.870 (0.00400)
Asian alone	0.972 (0.00048)	0.881 (0.00179)
Native Hawaiian / Pacific Islander alone	0.975 (0.00266)	0.888 (0.01300)
Some other race alone	0.967 (0.00064)	0.846 (0.00168)
Multi-race	0.963 (0.00074)	0.846 (0.00293)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

^ S.E. = standard error

Table D-6. Model 3 $\hat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by HHT (Household Family Type) Category

Regressor	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
Husband/wife family household	0.981 (0.00017)	0.902 (0.00044)
Other family/household: male householder	0.967 (0.00044)	0.849 (0.00163)
Other family household: female householder	0.962 (0.00038)	0.854 (0.00090)
Nonfamily household: male householder, living alone	0.954 (0.00043)	0.825 (0.00103)
Nonfamily household: male householder, not living alone	0.968 (0.00052)	0.821 (0.00181)
Nonfamily household: female householder, living alone	0.960 (0.00038)	0.869 (0.00083)
Nonfamily household: female householder, not living alone	0.969 (0.00058)	0.834 (0.00203)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

^ S.E. = standard error

Table D-7. Model 3 $\hat{R}(\hat{\rho})$ Values and Sample Completeness Ratios, by SEG_GRP (Segmentation Group)

Regressor	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
No segmentation group code	0.971 (0.00042)	0.927 (0.00147)
All around average I (homeowner skewed)	0.974 (0.00023)	0.870 (0.00052)
All around average II (renter skewed)	0.964 (0.00039)	0.865 (0.00083)
Economically disadvantaged I (homeowner skewed)	0.960 (0.00046)	0.841 (0.00123)
Economically disadvantaged II (renter skewed)	0.941 (0.00082)	0.806 (0.00167)
Ethnic enclave I (homeowner skewed)	0.979 (0.00040)	0.871 (0.00168)
Ethnic enclave II (renter skewed)	0.966 (0.00069)	0.862 (0.00200)
Young/mobile/singles	0.957 (0.00045)	0.835 (0.00107)
Advantaged homeowners	0.977 (0.00023)	0.898 (0.00048)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

^ S.E. = standard error

Table E-1A. Distribution of $\widehat{R}(\hat{\rho})$ Values for BLD (Edited Building Structure Type), across FIPST (includes Puerto Rico) – Model Set 2

Regressor	Min	P25	Median	P75	Max	Average
One-family house	0.899	0.967	0.974	0.978	0.988	0.970
Multi-family house	0.864	0.946	0.960	0.968	0.981	0.954
Trailer/mobile home *	0.933	0.967	0.973	0.981	1.000	0.972
Other (boat/RV/van, etc.) *	0.874	0.975	1.000	1.000	1.000	0.980

* Omits the District of Columbia – there were zero housing unit interviews in these two BLD categories in the District.

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-1B. Distribution of Sample Completeness Rates for BLD (Edited Building Structure Type), across FIPST (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
One-family house	0.848	0.873	0.896	0.915	0.942	0.893
Multi-family house	0.721	0.796	0.819	0.847	0.885	0.820
Trailer/mobile home	0 *	0.779	0.818	0.856	1.062	0.797
Other (boat/RV/van, etc.)	0 *	0.434	0.528	0.661	1.261	0.563

* The zero minimums are for the District of Columbia – there were zero housing unit interviews in these two BLD categories in the District.

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-2A. Distribution of $\widehat{R}(\hat{\rho})$ Values for AGE (Edited Age of Householder), across FIPST (includes Puerto Rico) – Model Set 2

Regressor	Min	P25	Median	P75	Max	Average
0 to 24	0.881	0.949	0.961	0.971	0.987	0.955
25 to 34	0.870	0.951	0.964	0.971	0.986	0.959
35 to 44	0.846	0.953	0.966	0.973	0.986	0.961
45 to 54	0.869	0.958	0.967	0.974	0.987	0.963
55 to 64	0.900	0.966	0.972	0.977	0.988	0.969
65 to 74	0.919	0.973	0.977	0.982	0.989	0.975
75+	0.943	0.979	0.981	0.985	0.990	0.979

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-2B. Distribution of Sample Completeness Rates for AGE (Edited Age of Householder, across FIPST (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
0 to 24	0.654	0.745	0.772	0.814	0.886	0.777
25 to 34	0.775	0.808	0.830	0.858	0.917	0.835
35 to 44	0.795	0.839	0.869	0.885	0.919	0.863
45 to 54	0.809	0.859	0.886	0.902	0.973	0.882
55 to 64	0.841	0.873	0.897	0.913	0.954	0.894
65 to 74	0.850	0.885	0.906	0.922	0.958	0.904
75+	0.848	0.890	0.912	0.921	0.976	0.906

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-3A. Distribution of $\hat{R}(\hat{\rho})$ Values for HHSPAN (Hispanic or Latino Householder, across FIPST (includes Puerto Rico) – Model Set 2

Regressor	Min	P25	Median	P75	Max	Average
Not Hispanic or Latino	0.873	0.957	0.967	0.974	0.986	0.962
Hispanic or Latino	0.904	0.946	0.966	0.975	0.986	0.959

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-3B. Distribution of Sample Completeness Ratios for HHSPAN (Hispanic or Latino Householder), across FIPST (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
Not Hispanic or Latino	0.708	0.853	0.879	0.894	0.934	0.873
Hispanic or Latino	0.755	0.831	0.848	0.871	0.986	0.850

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-4A. Distribution of $\widehat{R}(\hat{\rho})$ Values for TENSHORT (Tenure), across FIPST (includes Puerto Rico) – Model Set 2

Regressor	Min	P25	Median	P75	Max	Average
Owner-occupied unit	0.905	0.968	0.977	0.981	0.990	0.974
Renter-occupied unit	0.863	0.946	0.959	0.969	0.982	0.955

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-4B. Distribution of Sample Completeness Ratios for TENSHORT (Tenure), across FIPST (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
Owner-occupied unit	0.849	0.879	0.908	0.922	0.965	0.903
Renter-occupied unit	0.741	0.793	0.808	0.836	0.886	0.815

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-5A. Distribution of $\hat{R}(\hat{\rho})$ Values for RACE (Race of Householder), across FIPST (includes Puerto Rico) – Model Set 2

Regressor	Min	P25	Median	P75	Max	Average
White alone	0.927	0.964	0.971	0.978	0.987	0.970
Black alone	0.860	0.944	0.962	0.969	1.000	0.953
American Indian / Alaska Native alone	0.861	0.945	0.961	0.972	1.000	0.954
Asian alone	0.897	0.958	0.972	0.979	1.000	0.965
Native Hawaiian / Pacific Islander alone	0.861	0.973	1.000	1.000	1.000	0.983
Some other race alone	0.795	0.950	0.965	0.975	0.989	0.958
Multi-race	0.860	0.949	0.966	0.973	0.988	0.959

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-5B. Distribution of Sample Completeness Ratios for RACE (Race of Householder), across FIPST (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
White alone	0.835	0.859	0.887	0.901	0.925	0.880
Black alone	0.773	0.812	0.836	0.858	0.909	0.837
American Indian / Alaska Native alone	0.668	0.822	0.856	0.881	1.166	0.853
Asian alone	0.760	0.835	0.867	0.894	0.987	0.866
Native Hawaiian / Pacific Islander alone	0.249	0.808	0.882	1.003	1.564	0.899
Some other race alone	0.689	0.803	0.828	0.852	0.995	0.830
Multi-race	0.752	0.811	0.845	0.864	0.942	0.842

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-6A. Distribution of $\widehat{R}(\hat{\rho})$ Values for HHT (Household Family Type), across FIPST (includes Puerto Rico) – Model Set 2

Regressor	Min	P25	Median	P75	Max	Average
Husband/wife family household	0.916	0.973	0.981	0.985	0.992	0.978
Other family/household: male householder	0.863	0.958	0.968	0.976	0.991	0.963
Other family household: female householder	0.853	0.954	0.967	0.975	0.988	0.960
Nonfamily household: male householder, living alone	0.857	0.944	0.957	0.965	0.981	0.952
Nonfamily household: male householder, not living alone	0.914	0.955	0.971	0.979	0.989	0.964
Nonfamily household: female householder, living alone	0.885	0.953	0.964	0.971	0.985	0.959
Nonfamily household: female householder, not living alone	0.892	0.957	0.973	0.979	0.989	0.964

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-6B. Distribution of Sample Completeness Ratios for HHT (Household Family Type), across FIPST (includes Puerto Rico)

Regressor	Min	P25	Median	P75	Max	Average
Husband/wife family household	0.844	0.881	0.905	0.922	0.957	0.902
Other family/household: male householder	0.791	0.830	0.844	0.869	1.026	0.851
Other family household: female householder	0.781	0.836	0.858	0.878	0.980	0.858
Nonfamily household: male householder, living alone	0.760	0.794	0.822	0.849	0.880	0.821
Nonfamily household: male householder, not living alone	0.711	0.786	0.815	0.848	0.920	0.816
Nonfamily household: female householder, living alone	0.817	0.851	0.868	0.888	0.917	0.868
Nonfamily household: female householder, not living alone	0.742	0.807	0.834	0.858	0.929	0.835

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-7A. Distribution of $\hat{R}(\hat{\rho})$ Values for SEG_GRP (Segmentation Group), across FIPST (excludes states that did not contain data for a particular segmentation group) – Model Set 2

Regressor	n*	Min	P25	Median	P75	Max	Average
No segmentation group code	52	0.853	0.955	0.969	0.977	0.992	0.964
All around average I (homeowner skewed)	51	0.918	0.963	0.969	0.975	0.986	0.967
All around average II (renter skewed)	51	0.868	0.954	0.968	0.974	0.983	0.961
Economically disadvantaged I (homeowner skewed)	48	0.882	0.955	0.965	0.976	0.987	0.959
Economically disadvantaged II (renter skewed)	47	0.858	0.939	0.956	0.971	0.994	0.950
Ethnic enclave I (homeowner skewed)	41	0.906	0.958	0.973	0.980	1.000	0.968
Ethnic enclave II (renter skewed)	29	0.897	0.945	0.968	0.983	1.000	0.965
Young/mobile/singles	51	0.893	0.952	0.967	0.979	1.000	0.963
Advantaged homeowners	51	0.919	0.973	0.978	0.981	0.987	0.975

* n = number of states containing the segmentation group. Note that Puerto Rico did not have segmentation groups.

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table E-7B. Distribution of Sample Completeness Ratios for SEG_GRP (Segmentation Group), across FIPST (excludes states that did not have data for a particular segmentation group)

Regressor	n*	Min	P25	Median	P75	Max	Average
No segmentation group code	52	0.753	0.861	0.934	1.035	1.191	0.946
All around average I (homeowner skewed)	51	0.812	0.851	0.880	0.894	0.922	0.872
All around average II (renter skewed)	51	0.801	0.843	0.869	0.886	0.910	0.864
Economically disadvantaged I (homeowner skewed)	48	0.766	0.834	0.854	0.884	1.704	0.874
Economically disadvantaged II (renter skewed)	47	0.683	0.777	0.808	0.838	1.183	0.816
Ethnic enclave I (homeowner skewed)	41	0.762	0.810	0.855	0.893	1.721	0.878
Ethnic enclave II (renter skewed)	29	0.582	0.804	0.842	0.873	1.060	0.833
Young/mobile/singles	51	0.755	0.817	0.839	0.860	0.916	0.837
Advantaged homeowners	51	0.832	0.884	0.896	0.913	0.926	0.894

* n = number of states containing the segmentation group. Note that Puerto Rico did not have segmentation groups.

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-1. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – BLD (Edited Building Structure Type)

Regressor	Rank Correlation	P-Value
One-family house	0.272	0.051
Multi-family house	0.353	0.010
Trailer/mobile home	-0.026	0.858
Other (boat/RV/van, etc.)	-0.104	0.467

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-2. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – AGE (Edited Age of Householder)

Regressor	Rank Correlation	P-Value
0 to 24	0.468	< 0.001
25 to 34	0.418	0.002
35 to 44	0.273	0.050
45 to 54	0.227	0.106
55 to 64	0.176	0.213
65 to 74	0.091	0.520
75+	0.145	0.306

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-3. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – HHSPAN (Hispanic or Latino Householder)

Regressor	Rank Correlation	P-Value
Not Hispanic or latino	0.310	0.025
Hispanic or latino	0.198	0.160

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-4. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – TENSHORT (Tenure)

Regressor	Rank Correlation	P-Value
Owner-occupied unit	0.342	0.013
Renter-occupied unit	0.300	0.031

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-5. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – RACE (Race of Householder)

Regressor	Rank Correlation	P-Value
White alone	0.307	0.027
Black alone	0.424	0.002
American Indian / Alaska Native alone	0.079	0.580
Asian alone	0.413	0.002
Native Hawaiian / Pacific Islander alone	0.005	0.975
Some other race alone	0.230	0.101
Multi-race	0.215	0.126

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-6. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – HHT (Household Family Type)

Regressor	Rank Correlation	P-Value
Husband/wife family household	0.317	0.022
Other family/household: male householder	0.022	0.876
Other family household: female householder	0.241	0.085
Nonfamily household: male householder, living alone	0.358	0.009
Nonfamily household: male householder, not living alone	0.372	0.007
Nonfamily household: female householder, living alone	0.369	0.007
Nonfamily household: female householder, not living alone	0.389	0.004

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Table F-7. Spearman Rank Correlations of States for Model Set 2 $\widehat{R}(\hat{\rho})$ Values vs Sample Completeness Ratio Values – SEG_GRP (Segmentation Group)

Regressor	Rank Correlation	P-Value
No segmentation group code	0.398	0.004
All around average I (homeowner skewed)	0.196	0.167
All around average II (renter skewed)	0.442	0.001
Economically disadvantaged I (homeowner skewed)	0.503	< 0.001
Economically disadvantaged II (renter skewed)	0.384	0.008
Ethnic enclave I (homeowner skewed)	0.385	0.013
Ethnic enclave II (renter skewed)	0.010	0.960
Young/mobile/singles	0.435	0.001
Advantaged homeowners	0.126	0.379

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011

Attachment G
Page 1 of 1

Table G. Segmentation Group Codes

Segmentation Group (SG)	Percent Occupied Housing Units	Census 2000 Mail Return Rate	Characteristics
0 – CLUSTERNUM is blank	-	-	-
1 – All around average I (homeowner skewed)	35%	77.3%	- 75% owners - 80% non-Hispanic white - largest % of rural tracts - unemployment, poverty, education and mobility levels are close to national averages - skewed towards older persons
2 – All around average II (renter skewed)	16%	74.2%	- more urban and densely populated than SG 1 - above average % of renters and multi-units - skewed towards younger persons
3- Economically Disadvantaged I (homeowner skewed)	6%	66.5%	- 92% of tracts - 49% black - above average % of children - skewed towards older homeowners - higher percentage unemployment, poverty, receiving public assistance, without high school education
4 – Economically Disadvantaged II (renter skewed)	3%	58.0%	- 99.9% of tract are urban - 54% black and 21% hispanic - 81% renters - 1/3 of households speak a language other than english - highest poverty, public assistance, unemployment of all SGs
5 – Ethnic Enclave I (homeowner skewed)	3%	69.8%	- 61% Hispanic - above-average percentage of children - like SG 6 except less linguistic isolation, lower mobility, higher homeownership, fewer asians, less urban, less densely populated - 43% foreign born, 58% of households speak spanish at home
6 – Ethnic Enclave II (renter skewed)	2%	63.6%	- 59% hispanic, 11% Asian - above average % of children - 75% are renters - 34% linguistically isolated - exclusively urban, most densely populated SG, crowded housing - 50% without high school degree
7 – Young/mobile/singles	8%	67.1%	- densely populated and almost exclusively urban - overwhelming majority of households are non-spousal renters in multi-units - skewed to a more educated population - racial and ethnic diversity
8 – Advantaged homeowners	26%	83.2%	- least racially diverse with 85% non-hispanic white - least densely populated - very high percentage of owners, few multi-unit structures, high education, very low levels of poverty and unemployment, low mobility, few non-spousal households

Source: Boone (2008)