# Sample Representivity in the American Community Survey

Don Keathley, Steven P. Hefter
U.S. Census Bureau, Washington, D.C., 20233

**Abstract**
The American Community Survey (ACS) selects national housing unit address samples on a yearly basis. Each sample is selected systematically, using geography and estimated occupied housing unit counts within specific geographies as sort variables. Every housing unit address on the ACS frame is eligible for sample once every five years, with approximately one-fifth of the addresses being eligible in a given year. Weighted response rates for the yearly  samples average above ninety-seven percent, so the ACS has respondent information on the vast majority of the sampled units. But, there is still a two-plus percent nonresponse rate, and it is uncertain as to whether these cases are systematically different from the respondents, for one or more estimation categories of interest, e.g., race. Sample representivity statistics attempt to quantify the representativeness of the responding units to the nonresponding units and, by extension, to the entire frame for these categories

**Key Words:** American Community Survey, housing unit address sample, sample representivity

## 1. Background

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 Internet page). 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 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 looked at sample completeness ratios (SCR) for comparison purposes, which are measures of the combined levels of

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.*

nonresponse and under- or overcoverage. Due to space limitations, however, we omit the discussion of SCRs from this paper (see Keathley (2013a) and (2013b)).

Our analysis in this evaluation focuses on both the United States and Puerto Rico as well as American Indian areas only. The American Indian areas include regions such as reservations and tribal statistical areas. We estimate sample representivity at the national level as a whole for both groups as well as 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.

### 1.1 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 $\rho$ 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

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

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

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

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

S($\rho$) is in the closed interval [0, 0.5]. This means R($\rho$) is in the closed interval of [0, 1]. R($\rho$) = 1 when S($\rho$) = 0, indicating all units in the population have the same propensity to respond, regardless of the characteristics of the population units. R($\rho$) values less than one indicate response propensities that vary across the population units – these propensities could be correlated with one or more population characteristics, e.g., age.

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.

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

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

$\hat{\mathrm{S}}(\hat{\boldsymbol{\rho}})$ = standard deviation of $\hat{\boldsymbol{\rho}}$

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

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

The design weight $\mathrm{d_i}$ 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 $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{d_i}$ in place of N in equation (4).

A difference between the standard deviations in equations (2) and (4) is that $\hat{\mathrm{S}}(\hat{\rho})$ values are in the left-open interval (0, 0.5]. This means $\widehat{\mathrm{R}}(\hat{\rho})$ values are in the right-open interval of [0, 1). This downward bias in $\widehat{\mathrm{R}}(\hat{\rho})$ is due to sampling variation in the estimated response propensities (Shlomo, et al (2009)) [1].

We estimated response propensities for ACS sample housing units for the nation as a whole, for each state (including Washington, D.C. and Puerto Rico), and for American Indian only areas as a whole, for the sample years 2007 through 2011 combined. We made these estimates using logistic regression models. The general form of these models is

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

where $\mathrm{g}(\mathbf{x_i})$ 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, $\mathrm{g}(\mathbf{x_i})$ in equation (5) becomes

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

[1] Shlomo, et al (2009) presents a (bias) adjustment that takes this variation into account; Schouten, et al (2009) mentions that the bias-adjusted $\widehat{\mathrm{R}}(\hat{\rho})$ reduces to the $\widehat{\mathrm{R}}(\hat{\rho})$ in equation (3) as n→∞, for simple random and stratified simple random samples.

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 sample units' response propensities. We chose the variables listed in Table 1 as the regressors.

Six of the regressors are unit-level 2010 Census variables from the 2010 Census Hundred-Percent Detailed File (HDF) for housing units, two come from edited Master Address File (MAF) extracts, two come from Geography Division files, and one (CLUSTERNUM) was produced as part of the analysis in U.S. Census Bureau (2008). Note that the variables from the edited MAF extracts are sample 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 for our multiple regression models. 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. All runs used reference group parameterization[2]. The significance level for adding and retaining a variable to the model was 0.01.

## 2. 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 the national eligible cases (including Puerto Rico) did not match to the HDF (480,233 of 9,253,859 cases)[3]; this figure was about 9.5 percent of the eligible cases in American Indian areas (24,450 of 258,423 cases)[4].

A second limitation is that not all of the eligible cases that matched to an HDF record had entries for the variables of interest (regressors) on the HDF, i.e., many were vacant housing units in Census 2010. These records comprised approximately 4.3 percent of the national eligible cases that matched to the HDF (381,101 of 8,773,626 cases) and about 5.5 percent of the matching eligible American Indian area cases (12,921 of 233,883).

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.

A third 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

[2] Reference groups are the levels of the variables in a model against which the parameter estimates for the remaining levels are compared.
[3] The 9,253,859 total excludes matches that were not housing units in Census 2010 – there were 929 such cases.
[4] The 258,423 total excludes matches that were not housing units in Census 2010 – there were 17 such cases.

performed using these variables. This would have potentially resulted in more accurate matching between the files.

A potential limitation concerns ACS interviews that matched to the HDF and where the matching HDF record had non-blank entries for the regressors. Because of the time lag between the Census and the periods of ACS data collection, some of the these cases may have differing ACS and Census values for the regressors, e.g., the age of the householder could have changed between the ACS interview and the Census. Using ACS response data instead of what was in the HDF might have altered the modeling results.

## 3. Methodology

This section reviews the variables and models we used and the statistics we generated.

### 3.1 Input Files, Variables

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 Attachments A and B for descriptions of all of the files mentioned in this section.

**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 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 |
| LSADC | Legal/Statistical Area Definition Code | 2007-2011 GRFC#, GRFN& |
| MTFCC | MAF Tiger Feature Code | 2007-2011 GRFN# |
| 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.
# Geographic Reference Files, with geography Codes
* Geographic Reference Files, with geography Names

We included all variables *except* LSADC and MTFCC in the national models, whereas we included all variables *except* FIPST and FCNTY in the American Indian area models only.

We merged various files, including those shown in Table 1, to create the input files for the logistic regression modeling and R-indicator computations. These files contain all of the variables shown in Table 1. See Keathley (2013a) or (2013b) for a description on how we merged the files.

The codes for each variable that we used are shown in the tables in the attachments. The last column in the table shows the code/category we used as the reference group for the regressor.

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.

### 3.2 Logistic Regression Models

We ran the national models in Table 2 and the American Indian area models in Table 3. 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 Census 2010 – non-vacants[5].

**Table 2. National 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

**Table 3. American Indian Area Models**

| Model / Model Set | Description |
|---|---|
| 1L | All main effects only from Table 1, except FIPST, FCNTY, and MTFCC, using cases with acceptable LSADC values only |
| 1M | All main effects only from Table 1, except FIPST, FCNTY, and LSADC, using cases with acceptable MTFCC values only |
| 2L | Same as 1L, except with two-way interactions |
| 2M | Same as 1M, except with two-way interactions |
| 3 | Same as 1L/1M – all main effects except FIPST, FCNTY, MTFCC, and LSADC |
| 4 | Same as 2L/2M – all main effects, except FIPST, FCNTY, MTFCC, and LSADC, and all interactions except those involving MTFCC and LSADC |
| 5L | Same as 1L, except collapsed only those LSADC categories that had non-significant parameters |

We ran two sets of logistic regression models for the American Indian areas, one where we omitted LSADC and one where we omitted MTFCC (there are only nine American Indian are MTFCCs, compared to over thirty LSADCs in these areas). We ran these models using only those ACS interviews and noninterviews with an MTFCC that was

[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 one.

equal to one of the codes in Table 4 (eligible cases). Each eligible case had an LSADC equal to one of the codes shown in Attachment D in Keathley (2013b).

**Table 4. MAF (Master Address File) Tiger Feature Class Code (MTFCC) for American Indian Areas**

| MTFCC | MTFCC Description |
|---|---|
| G2100 | Legal American Indian Area |
| G2120 | Hawaiian Homeland |
| G2130 | Alaska Native Village Statistical Area (ANVSA) |
| G2140 | Oklahoma Tribal Statistical Area (OTSA) |
| G2150 | State Designated Tribal Statistical Area (SDTSA) |
| G2160 | Tribal Designated Statistical Area (TDSA) |
| G2170 | Joint-use Area |
| G2200 | Alaska Native Regional Corporation (ANRC) |
| G2300 | Tribal Subdivision |

All of the models except 5L were exploratory models, where we compared the models to each other, primarily with respect to model fit. Model 5L took the best fitting of the six models (1L – see Table 10) and collapsed the American Indian areas with parameters from model 1L that were not significant into one parameter.

### 3.3 R-indicators

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

### 3.4 Model Goodness-of-Fit Metrics

Model goodness-of-fit metrics are indicators of how well each model fits in comparison to the other similar 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:

$$\text{Adj } \mathrm{R}^2 = \mathrm{R}^2 \,/\, \text{Max } \mathrm{R}^2 \tag{7}$$

where $\mathrm{R}^2$ = a generalization of the CD (Cox (1989))

$$= 1 - \left(\frac{\mathrm{L}(0)}{\mathrm{L}(\hat{\beta}\,)}\right)^{2/n} \tag{8}$$

[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.

$$\text{Max } R^2 = \text{maximum } R^2 \text{ value} = 1 - (L(0))^{2/n} \qquad (9)$$

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. In our analysis, positives are interviews, with a negative being a noninterview. 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 (2010) for more information on ROC curves.

## 4. Results/Analysis

This section presents results and analyses for the models presented in section 3.2 and the R-indicators that are based on these models.

### 4.1 National Models

Table 5 shows the goodness-of-fit metrics for each of the three national models; Table 6 shows the $\hat{R}(\hat{\boldsymbol{\rho}})$ values for each model.

The results in Table 5 show that national model 3 is the best fitting national model, with the smallest -2 Log L value (22,848,888) and largest Adj. $R^2$ value (0.124). It had the best ability to predict whether a housing unit address will be an interview in the ACS as well, with an area under the ROC curve of 0.688. All of the main effects in the stepwise regression made it into the model. Because of these results, we favored model 3 over models 1 and 2. The $\hat{R}(\hat{\boldsymbol{\rho}})$ value for model 3 is 0.965.

**Table 5. 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 | 0.089 | 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

## 4.2 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 -2 Log L / n distributions are nearly the same, but the Adj. $R^2$ and ROC curve area distributions favor model set 2.

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.

**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 | 0.426 | 0.071 |
| Area under ROC Curve | 0.522 | 0.583 | 0.608 | 0.645 | 0.725 | 0.612 |

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

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

**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

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 metrics above show that, in spite of the two questionable fits in state model set 2, the models in this set generally have better fits and predictive ability than their counterparts in state model set 1. Because of these results, we preferred the models in model set 2. Table 8 shows the distribution of the $\widehat{R}(\widehat{\rho})$ values across the states for model set 2.

**Table 8. $\widehat{R}(\widehat{\rho})$ Value Distribution for State Model Set 2**

| Min | P25 | Median | P75 | Max | Average |
|---|---|---|---|---|---|
| 0.875 | 0.956 | 0.967 | 0.973 | 0.986 | 0.962 |

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

### 4.3 American Indian Area Models

Table 9 shows the American Indian models we ran. The first six models (1L through 4) were exploratory models. The model with the best fit (smallest -2 Log L and largest Adj. $R^2$) and best predictive ability (largest area under the ROC curve) is model 1L, where we included LSADC in the stepwise regression but omitted MTFCC. We then tweaked model 1L by collapsing the LSADC categories with non-significant parameters into one category – the result was model 5L. The $\widehat{R}(\hat{\rho})$ values for the two models are shown in Table 10.

**Table 9. Summary of Logistic Regression Runs – American Indian Areas**

| | | | Goodness-of-Fit Metrics | | |
|---|---|---|---|---|---|
| **Model** | **Steps** | **Variables in Model** | **-2 Log L** | **Adj. $R^2$** | **Area under ROC Curve** |
| 1L | 8 | All | 419,201 | 0.087 | 0.693 |
| 1M | 8 | All | 421,627 | 0.075 | 0.677 |
| 2L | 36 | All, but the model fit was questionable after step 3 | 435,491 | 0.005 | 0.725 |
| 2M | 36 | All, but the model fit was questionable after step 4 | 433,109 | 0.017 | 0.711 |
| 3 | 7 | All | 426,223 | 0.052 | 0.663 |
| 4 | 28 | All, but the model fit was questionable after step 15 | 425,535 | 0.056 | 0.667 |
| 5L | 8 | All | 419,281 | 0.087 | 0.693 |

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

**Table 10. $\widehat{R}(\hat{\rho})$ Values – American Indian Areas**

| Model | $\widehat{R}(\hat{\rho})$ |
|---|---|
| 1L | 0.965 |
| 5L | 0.965 |

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

Due to the results in Tables 9 and 10, we preferred model 5L over 1L.

## 5. Conclusions

The $\widehat{R}(\hat{\rho})$ values for the preferred models in section 4 are evidence that, at the national, state, and American Indian levels, the ACS respondents on the logistic regression input files are representative of the non-respondents that are present on these files. In turn, this would indicate 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).

## 6. Future Research

Future research could include the use of ACS data, auxiliary information from other external 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.

The $\widehat{\mathrm{R}}(\widehat{\boldsymbol{\rho}})$ values we calculated are estimates and, as such, have standard errors (Shlomo, et al (2009)). The large sample sizes and $\widehat{\mathrm{R}}(\widehat{\boldsymbol{\rho}})$ values themselves should lead to fairly small coefficients of variation (CV; standard error ($\widehat{\mathrm{R}}(\widehat{\boldsymbol{\rho}})$) / $\widehat{\mathrm{R}}(\widehat{\boldsymbol{\rho}})$), but computing variance estimates of these values would give an indication of the magnitude of the CVs. Standard errors would also allow us to make statistical comparisons between $\widehat{\mathrm{R}}(\widehat{\boldsymbol{\rho}})$ values.

The use of the bias-adjusted $\widehat{\mathrm{R}}(\widehat{\boldsymbol{\rho}})$ or the $q^2$ R-indicators mentioned in Schouten, et al (2009), could be explored, as comparisons to the results in this report.

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.

## 7. References

Cox, 1989 – Cox, D.R. and Snell, E.J. 1989, “The Analysis of Binary Data,” Second Edition, London: Chapman and Hall

Keathley, 2013a -- Keathley, D. and Hefter, S., “American Community Survey: R Indicators for the Nation and Puerto Rico,” – to appear

Keathley, 2013b -- Keathley, D. and Hefter, S., “American Community Survey: R Indicators for American Indian Areas”

Kleinbaum, 2010 – Kleinbaum, D. and Klein, M., “Logistic Regression,” Third Edition, Springer 2010

Nagelkerke, N.J.D. 1991 “A Note on a General Definition of the Coefficient of Determination,” Biometrika, 78, 691-692

Schouten, et al, 2009 – Schouten, B., Morren, M., Bethlehem, J., Shlomo, N., Skinner, “How to use R-indicators”, http://www.risq-project.eu/papers/RISQ-Deliverable-3.pdf

Shlomo, et al, 2009 – Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J., Zhang, L., “Statistical Properties of R-indicators,” http://www.risq-project.eu/papers/RISQ-Deliverable-2-1-V2.pdf

Skinner, et al, 2009 – Skinner, C., Shlomo, N., Schouten, B., Zhang, L., Bethlehem, J., “Measuring Survey Quality through Representativeness Indicators using Sample and Population Based Information”, http://www.risq-project.eu/papers/skinner-shlomo-schouten-zhang-bethlehem-2009-a.pdf

Starsinic, 2002 -- Albright, K., Starsinic, M., “Coverage and Completeness in the Census 2000 Supplementary Survey,” http://www.census.gov/acs/www/Downloads/library/2002/2002_Starsinic_01.pdf

U.S. Census Bureau, 2008**.** Segmenting the Population for the 2010 Census Integrated Communications Campaign. http://www.census.gov/2010census/partners/pdf/C2POMemoNo_1_10-24-08.pdf

U.S. Census Bureau, 2009, “Design and Methodology,” http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf

U.S. Census Bureau internet page: http://www.census.gov/acs/www/methodology/response_rates_data/

Attachment A

## Table A. Variable Values for the Regressors

| Variable | Regressor | Values | Reference Group |
|---|---|---|---|
| BLD | BLD | S = one-family house<br>M = multi-family house<br>T = trailer/mobile home<br>O = other (boat/RV/van, etc) | S |
| CLUSTERNUM | SEG_GRP | See Keathley (2013a) or (2013b) | 0 |
| HHLDRAGE | AGE | 1 = 0 to 24<br>2 = 25 to 34<br>3 = 35 to 44<br>4 = 45 to 54<br>5 = 55 to 64<br>6 = 65 to 74<br>7 = 75+ | 2 |
| HHSPAN | HHSPAN | 1 = not Hispanic or latino<br>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<br>2 = Black alone<br>3 = Amerind/Alaskan Native alone<br>4 = Asian alone<br>5 = Native Hawaiian/pacific islander alone<br>6 = Some other race alone<br>7 = Multi-race | 1 |
| HHT | HHT | 1 = Husband/wife family household<br>2 = Other family household: male householder<br>3 = Other family household: female householder<br>4 = Nonfamily hhld: male hhldr, living alone<br>5 = Nonfamily hhld: male hhldr, not living alone<br>6 = Nonfamily hhld: female hhldr, living alone<br>7 = Nonfamily hhld: female hhde, not living alone | 1 |
| STATUS | ACSINT | 1 = Interview (ACSINT = 1)<br>4 = Non-Interview (ACSINT = 0)<br>All other codes were out-of-scope for this evaluation | - |
| TENSHORT | TENSHORT | 1 = Owner-occupied unit<br>2 = Renter-occupied unit | 1 |

Attachment B

## Table B. Segmentation Group Codes

| Segmentation Group (SEG_GRP) | 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<br>- 80% non-Hispanic white<br>- largest % of rural tracts<br>- unemployment, poverty, education and mobility levels are close to national averages<br>- skewed towards older persons |
| 2 – All around average II (renter skewed) | 16% | 74.2% | - more urban and densely populated than SG 1<br>- above average % of renters and multi-units<br>- skewed towards younger persons |
| 3- Economically Disadvantaged I (homeowner skewed) | 6% | 66.5% | - 92% of tracts<br>- 49% black<br>- above average % of children<br>- skewed towards older homeowners<br>- 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<br>- 54% black and 21% hispanic<br>- 81% renters<br>- 1/3 of households speak a language other than english<br>- highest poverty, public assistance, unemployment of all SGs |
| 5 – Ethnic Enclave I (homeowner skewed) | 3% | 69.8% | - 61% Hispanic<br>- above-average percentage of children<br>- like SG 6 except less linguistic isolation, lower mobility, higher homeownership, fewer asians, less urban, less densely populated<br>- 43% foreign born, 58% of households speak spanish at home |
| 6 – Ethnic Enclave II (renter skewed) | 2% | 63.6% | - 59% hispanic, 11% Asian<br>- above average % of children<br>- 75% are renters<br>- 34% linguistically isolated<br>- exclusively urban, most densely populated SG, crowded housing<br>- 50% without high school degree |
| 7- Young/ mobile/ singles | 8% | 67.1% | - densely populated and almost exclusively urban<br>- overwhelming majority of households are non-spousal renters in multi-units<br>- skewed to a more educated population<br>- racial and ethnic diversity |
| 8 – Advantaged homeowners | 26% | 83.2% | - least racially diverse with 85% non-hispanic white<br>- least densely populated<br>- 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: U. S. Census Bureau 2008