Issues Concerning Imputation of Hispanic Origin due to Administrative Record Enumeration for the 2020 Census[1]

Richard A. Griffin, U.S. Census Bureau

1. Introduction

For the 2020 Decennial Census, we are looking into reducing the cost of Non-response Follow-up (NRFU) by using administrative records as a substitute for field follow-up for housing units that do not respond for themselves by, for example, mail or the internet. Enumeration using administrative records is referred to as ADREC Enumeration. In many cases, with the sources we currently have available to us, administrative records have no information on Hispanic Origin or the available information may not be accurate. With these existing sources, the imputation rate for Hispanic Origin for ADREC Enumeration would be very high.

Missing data on Hispanic Origin (binary variable, Hispanic or Non-Hispanic) is ignorable if true origin status is independent of whether the origin status of a person could be resolved without imputation. If we were to use administrative records, one approach to filling in missing Hispanic origin for these cases would be to use Title 13 sources, such as previous censuses or the American Community Survey (ACS). Thus, for this paper, resolved is defined as origin self-response or origin is available from Title 13 data. Ignorable missing origin implies that the expected proportion of resolved origin persons that are Hispanic equals the expected proportion of unresolved persons that are Hispanic. This is unlikely for persons on the administrative records. It may be that those for whom previous Title 13 data is not available are more likely to be Hispanic.

The simulated census presented in this paper uses methodology that we will call the IRS/UAA approach. **This approach is only one of many currently being investigated for use in the 2020 Census**. IRS denotes the Internal Revenue Service and UAA stands for Undeliverable as Addressed. In the IRS/UAA approach, we assume census questionnaires are delivered to all housing units on the mailing list. Data for this paper comes from IRS returns for 2009 and filed before April 30, 2010 and from 2010 Census internal detail files. The IRS sends data to the Census Bureau Center for Administrative Records Research & Applications (CARRA). CARRA has merged address data from tax forms with Social Security Numbers (SSN). For confidentiality a unique identification number, called a protection identification key (PIK) was created for each SSN. The tax form address is matched to the 2010 Census address and all persons on the tax form with a PIK can be used for ADREC Enumeration.

For non-responding housing units, if an IRS return is not filed before April 30 and a vacant UAA code has been assigned by the United States Postal Service, the non-responding housing unit is classified as vacant and there are no further attempts at enumeration. This is called the vacant rule.

We will allow all non-responding housing units that are not classified as vacant by the vacant rule one NRFU visit. We will subject those housing units for which we do not get an interview at this first visit to the Occupancy Rule: If the housing unit has an IRS return

[1] *Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.*

filed before April 30 and has twelve or fewer persons on the IRS return, we will classify the housing unit as occupied with a population count from the IRS return. This is called ADREC enumeration. Enumeration is thus complete for these housing units classified as occupied by the Occupancy Rule.

After the first NRFU attempts have been completed, the proportion of housing units in each tract, including ADREC enumeration, that remain unresolved is computed. If this proportion is greater than 20%, we will allow up to two additional NRFU attempts. Any housing units in these tracts that remain unresolved after these two additional NRFU attempts we will send to count imputation. All housing units remaining unresolved after the first enumeration attempt in tracts with an unresolved rate less than or equal to 20% are sent to count imputation. All demographic data including Hispanic Origin status must be imputed for housing units sent to count imputation.

Table 1 (all numbers in thousands) provides information on Hispanic Origin for this simulation using ADREC Enumeration from sources outside the Census Bureau in place of NRFU. For this paper, administrative record data available on Hispanic Origin is ignored (its accuracy may be poor). If data on Hispanic Origin is available from a previous census response or other Title 13 data such as ACS, that response will be treated as observed. If no such Title 13 data is available, it will be necessary to impute using some other approach. For housing units enumerated by mail return , NRFU 1, NRFU 2, or NRFU 3 (NRFU j indicates enumeration during the jth NRFU visit to the housing unit), the 2010 Census Hispanic Origin response is used and treated as observed even though it may have been imputed.

Table 1*: Census 2010 and Title 13 Hispanic Origin Comparisons for Simulations

Description	IRS/UAA Approach Simulation[2]	% Hispanic
(1) Total HU Population (2)+(4)+(5)+(7)+(8)	297,266	
(2) 2010 Census Persons Enumerated by Mail	219,341	14.8%
(3) 2010 Census persons in HUs enumeration as vacant due to the Vacant Rule	1,959	13.7%
(4) 2010 Census persons enumerated in NRFU 1	32,084	21.1%
(5) Persons on IRS record for HUS in ADREC Enumeration due to the Occupancy Rule	24,670	
(6) Title 13 Data Available	13,919	15.9%
(7) 2010 Census persons enumerated in NRFU 2 or 3	2,579	23.1%
(8) 2010 Census persons enumerated by Count Imputation	18,593	16.1%
Persons with Resolved Hispanic Origin Status (2)+(4)+(6)+(7)	267,923	
Persons with Unresolved Hispanic Origin Status (5) - (6)+(8)	29,344	

*Numbers in Thousands

Using data from Table 1, 21.2% of the persons enumerated by NRFU 1, NRFU 2, or NRFU 3 were Hispanic. Persons counted by ADREC enumeration for whom previous Title 13 data was available were 15.9% Hispanic. Thus, there is concern that the overall proportion enumerated Hispanic after imputation for persons with no previous Title 13 data may be too low. Missing data on Hispanic Origin is ignorable if true origin status is independent of whether the origin status of a person could be resolved. Ignorable missing origin implies that the expected proportion of resolved origin persons that are Hispanic equals the expected proportion of unresolved persons that are Hispanic. This is unlikely for persons on the administrative records. It may be that those for whom previous Title 13 data is not available are more likely to be Hispanic.

Section 2 describes a potential non-ignorable missing data procedure and an ignorable missing data procedure using methodology adapted from Little and Rubin (1987). Using a missing data estimation procedure that is not appropriate (i.e., use of an ignorable model when a non-ignorable model is appropriate) creates a bias in the resulting proportion of persons who are Hispanic. Section 3 applies both the ignorable and non-ignorable procedures for the IRS/UAA simulation. The purpose is to demonstrate there may be statistically defensible missing data estimation methodology that will increase the imputation of an origin status of Hispanic for ADREC Enumeration persons with no previous Title13 data available on Origin. Section 4 provides a summary.

2. A Non-ignorable as well as Ignorable Missing Data Methodology

Little and Rubin (1997) provide non-ignorable missing data models for categorical data (section 11.6, pages 235-241). They describe a method using hierarchical loglinear models for the joint distribution of the categorical variables and indicator variables for non-response.

Adapting their example to our problem, let Y_1 denote a 2-category variable defined as follows:

On the 2012 Planning Database, an internal Census Bureau research file, there is an Hispanic concentration variable at the block-group level that can be used to calculate the percent Hispanic. It is based on 2006-2010 ACS data. Using this data, percentiles are used to classify the block-groups into high and low Hispanic groups. Low Hispanic block-groups will be defined as all block-groups with 0% Hispanic plus selection of from one up to nine deciles of all other block groups ranked by percent Hispanic.

Level 1: a person who lives in one of the High Hispanic block-groups.

Level 2: a person who lives in one of the Low Hispanic block-groups.

Let Y_2 denote a 2-category response variable defined as follows:

Level 1: an Hispanic person

Level 2: a non-Hispanic person

Define R to take the value:

1 if Y_2 is observed (value of Hispanic Origin from 2010 Census for mail or NRFU respondents or from previous Title 13 data if available for ADREC Enumeration persons).

0 if Y_2 is missing (unresolved).

The data are as shown in Table 2 with m_{++} persons classified on both Y_1 and Y_2 and r_{++} classified by Y_1 but not Y_2 that form an observed supplemental margin for which only Y_1 is observed. r_{1+} and r_{2+} are known. A jk subscript indicates level j of Y_1 and level k of Y_2. Y_1 is completely observed.

Table 2: 2×2 Contingency Table with One Partially Classified Margin

Live in a High Hispanic block-group (Y_1)	Resolved (R = 1)			Unresolved (R = 0)		
	Hispanic Origin (Y_2)			Hispanic Origin (Y_2)		
	Yes (1)	No (2)	Total	Yes (1)	No (2)	Total
Yes (1)	m_{11}	m_{12}	m_{1+}	$r_{11=?}$	$r_{12=?}$	r_{1+}
No (2)	m_{21}	m_{22}	m_{2+}	$r_{21=?}$	$r_{22+?}$	r_{2+}
Total	m_{+1}	m_{+2}	m_{++}			r_{++}

Little and Rubin display all the hierarchical models that include the main effects of Y_1, Y_2, and R. Here we assume that Y_1 and Y_2 are dependent (correlated). We would not use Y_1 to predict Y_2 if Y_1 and Y_2 were not related.

Within the { } notation below for three variable (X, Y, and Z) log linear models, if two variables are conditionally dependent, they are written together; {XY} indicates X and Y are conditionally dependent while if X and Y are not written together they are conditionally independent. X and Y are conditionally independent if at each level of Z they are marginally independent (i.e., ordinary two-way independence). Conditional independence does not imply marginal independence. It is possible that Y_1 and R are marginally dependent even though they are conditionally independent.

- If both Y_1 and Y_2 are conditionally independent of R, we have the ignorable model denoted by $\{Y_1Y_2, R\}$.
- If Y_2 is conditionally independent of R but Y_1 is conditionally dependent of R, we have the ignorable model denoted as $\{Y_1Y_2, Y_1R\}$.
- If Y_2 is not conditionally independent of R but Y_1 and R are conditionally independent, we have the non-ignorable conditional independence model denoted by $\{Y_1Y_2, Y_2R\}$.
- If Y_2 is not conditionally independent of R, and if Y_1 is not conditionally independent of R, we have the non-ignorable model denoted by $\{Y_1Y_2, Y_1R, Y_2R\}$.

Model $\{Y_1Y_2, Y_1R, Y_2R\}$ has inestimable parameters and additional information is needed to estimate the cell probabilities. Rubin and Little do not provide estimates for this model. Unfortunately, there is no data available to determine which of these models is correct. Here we will assume that if the missing data is non-ignorable, model $\{Y_1Y_2, Y_2R\}$ is appropriate. Thus, in this case we are assuming that Y_1 and R are conditionally independent.

For ignorable models $\{Y_1Y_2, R\}$ and $\{Y_1Y_2, Y_1R\}$, the maximum likelihood estimates arise from distributing the observed marginal counts, r_{1+} and r_{2+}, into the table to match the row distributions of the fully observed data. Thus, for these models we have

$$\hat{r}_{jk} = \frac{m_{jk}}{m_{j+}} r_{j+} \qquad (1)$$

For non-ignorable model $\{Y_1Y_2,\ Y_2\ R\}$, the maximum likelihood estimates, $\hat{r}^*_{jk}$ satisfy the following condition:

$$\hat{r}^*_{jk} = \frac{m_{jk}}{m_{+k}}\hat{r}^*_{+k} \tag{2}$$

Thus, the unresolved data match the column distributions of the resolved data.

To solve for the rates, first note that

$$\hat{r}^*_{12} = r_{1+} - \hat{r}^*_{11} \tag{3}$$

and

$$\hat{r}^*_{21} = r_{2+} - \hat{r}^*_{22} \tag{4}$$

In addition, model $\{Y_1Y_2,\ Y_2\,R\}$ has Y_1 and R independent at each level of Y_2. Under this assumption:

$$\frac{m_{11}\hat{r}^*_{21}}{m_{21}\hat{r}^*_{11}} = \frac{m_{12}\hat{r}^*_{22}}{m_{22}\hat{r}^*_{12}} = 1 \tag{5}$$

Equation (5) states that under this model at each level of Y_2 the 2 × 2 table crossing Y_1 and R has odds ratio equal to one.

Next substituting (3) and (4) in (5), we solve the following system of equations for $\hat{r}^*_{11}$ and $\hat{r}^*_{22}$:

$$\hat{r}^*_{22} = (r_{1+} - \hat{r}^*_{11})\frac{m_{22}}{m_{12}} \tag{6}$$

$$r_{2+} - \hat{r}^*_{22} = \hat{r}^*_{11}\frac{m_{21}}{m_{11}} \tag{7}$$

$$\hat{r}^*_{11} = (r_{2+} - r_{1+}\frac{m_{22}}{m_{12}})(\frac{m_{21}}{m_{11}} - \frac{m_{22}}{m_{12}})^{-1} \tag{8}$$

$$\hat{r}^*_{22} = (r_{1+}\frac{m_{21}}{m_{11}} - r_{2+})(\frac{m_{21}m_{12}}{m_{11}m_{22}} - 1)^{-1} \tag{9}$$

The results from using (8) and (9) are used in (3) and (4) to obtain the remaining two estimates.

To yield nonnegative estimates, $\hat{r}_{jk}^{*}$, the marginal column odds, $\frac{r_{1+}}{r_{2+}}$, must lie between the smallest and largest of the column odds, $\frac{m_{11}}{m_{21}}$ and $\frac{m_{12}}{m_{22}}$.

3. Application of Non-Ignorable and Ignorable Models to the IRS/UAA Simulation

In order to fill in the Table 2 counts, initially define Low Hispanic block-groups to be block groups with 0% Hispanic plus the first five deciles of the other block-groups.

The marginal (summed over Y_2) odds ratio for Y_1 and R is about 0.67, indicating that resolved persons are less likely to live in one of the High Hispanic block-groups than unresolved persons.

Table 3 provides the counts described in Table 2 using the UAA/IRS simulation.

Table 3: 2 × 2 Contingency Table for UAA/IRS Simulation: Person Counts*

Live in a High Hispanic block-group (Y_1)	Resolved (R = 1)			Unresolved (R = 0)		
	Hispanic Origin (Y_2)			Hispanic Origin (Y_2)		
	Yes (1)	No (2)	Total	Yes (1)	No (2)	Total
Yes (1)	35,683	73,280	108,961	$r_{11 = ?}$	$r_{12 = ?}$	14,831
No (2)	6,314	152,648	158,961	$r_{21 = ?}$	$r_{22 + ?}$	14,513
Total	41,995	225,927	267,923			29,344

*Numbers in Thousands

Using this table and the formulas from section 2, the resulting estimates for the non-ignorable and ignorable models are as shown in Table 4.

Table 4: IRS/UAA Simulation Estimates

Non-ignorable Model		Ignorable Model	
$\hat{r}_{11}^{*} = 8{,}594$	$\hat{r}_{12}^{*} = 6{,}237$	$\hat{r}_{11} = 4{,}857$	$\hat{r}_{12} = 9{,}974$
$\hat{r}_{21}^{*} = 1{,}521$	$\hat{r}_{22}^{*} = 12{,}993$	$\hat{r}_{21} = 576$	$\hat{r}_{22} = 13{,}937$
Overall Hispanic Proportion	0.175		0.160
Proportion Imputed Hispanic	0.345		0.185

These calculations were repeated for eight other definitions of High Hispanic block-groups. The overall Hispanic proportion and the proportion imputed Hispanic for all nine definitions are shown in Table 5.

Table 5: Hispanic Proportions for Alternative High Hispanic Block-Group Definitions

High Hispanic Block-Group Definition	Non-Ignorable Model		Ignorable Model	
	Overall Hispanic Proportion	Proportion Imputed Hispanic	Overall Hispanic Proportion	Proportion Imputed Hispanic
Deciles 2-10	0.176	0.352	0.158	0.166
Deciles 3-10	0.177	0.359	0.158	0.171
Deciles 4-10	0.177	0.358	0.159	0.175
Deciles 5-10	0.176	0.354	0.159	0.180
Deciles 6-10	0.175	0.345	0.160	0.185
Deciles 7-10	0.174	0.332	0.160	0.190
Deciles 8-10	0.172	0.316	0.160	0.193
Deciles 9-10	0.170	0.293	0.160	0.193
Decile 10	0.167	0.261	0.159	0.182

4. Summary

Results did not vary much by definition of High Hispanic block-group. Thus, results will be discussed for the High Hispanic block-group definition of deciles 6-10 that was used for Tables 3 and 4.

Missing Hispanic Origin after ADREC Enumeration may not be missing at random. It is quite possible that persons with unresolved Hispanic Origin are more likely to be Hispanic Origin than resolved origin persons. If so, a non-ignorable missing data model is more appropriate than an ignorable missing data model. For the UAA/IRS Simulation defining High Hispanic block-groups by the top 5 deciles using 2006-2010 ACS data, the non-ignorable model imputes 34.5 % Hispanics and the ignorable model imputes 18.5% Hispanics. For the housing unit population the final overall Hispanic proportion is 17.5% using the non-ignorable model and 16.0% using the ignorable model. This analysis merely serves to show that there may be statistically defensible ways to impute a higher proportion of Hispanics. Additional work will be done on the non-ignorable approaches. The methodology used in this paper only used High Hispanic blocks as the one covariate to form a 2x2 table. The use of additional covariates in an EM algorithm approach could result in different results than those shown in this paper.

5. Reference

Little, R. and Rubin, D. (1987), “Statistical Analysis with Missing Data”, Wiley Series in Probability and Mathematical Statistics