USING SURNAMES TO OVERSAMPLE HISPANICS FROM A LIST FRAME

David Judkins, James Massey and Valerija Smith
Westat, Inc.1650 Research Blvd. Rockville, Md

Keywords: Survey design, Rare Populations

Abstract

There is a keen interest in augmenting the Hispanic sample size for the National Health Interview Survey (NHIS). Oversampling geographic areas such as blocks that have strong concentrations of Hispanics can be an effective procedure but does not result in as strong a boost in the effective sample size for elderly Hispanics as desired. We have been conducting research on the use of Social Security lists for supplementing the elderly Hispanic sample. Unfortunately, these lists do not indicate ethnic origin for those who are currently elderly. As a proxy, we are examining the usefulness of a list of likely Hispanic surnames developed at the Census Bureau by Passel and Word. We have matched this list to the 1988 NHIS and compared the results with the self-classification of survey respondents to measure both the specificity and sensitivity of the surname as an indicator of ethnic origin. False positive and false negative rates are broken down by various demographic characteristics. Implications for sample design are given.

Introduction

The National Health Interview Survey uses a mixed area/permit sampling frame (onsite listing of blocks for old construction with list sampling from building permit registries for new construction, see Massey, Moore, Parsons and Tadros, 1989). There is strong interest in improving the reliability of age- and sex-specific statistics about minorities from this survey. A joint research program has established that these objectives can be met for the most part by a combination of oversampling blocks with high concentrations of minorities and household screening. However, elderly blacks and Hispanics are too rare (particularly males) for this procedure to work well. Since the Social Security Agency (SSA) maintains files with excellent coverage of the elderly population, we have been researching dual-frame sampling that would combine the traditional area/permit sample with a supplemental list sample. The task of oversampling elderly blacks from SSA files is fairly straightforward since race is indicated for about 97 percent of the file. Unfortunately, SSA files do not have an indicator for Hispanic origin. In order to use the SSA files for oversampling of elderly Hispanics despite the lack of an indicator for Hispanic origin, we considered the strategy that persons likely to be Hispanic be identified on the basis of surname, using the Hispanic surname file developed by J. Passel and D. Word at the Census Bureau (Passel and Word, 1980) for the purpose of classifying surnames by ethnic origin. The Passel-Word file contains 12,497 surnames that tend to belong to Hispanics.

Of course, the precision and cost of a dual-frame sample based upon surnames depends strongly on the sensitivity and specificity of the Passel-Word file. Every false positive costs money to interview or screen out and every false negative increases the sampling weights of the portion of the area/permit sample that is not covered by the list, thereby increasing design effects due to unequal weights. A past study (Passel and Word, 1980) indicated that false positives (also called errors of commission) run at around 15% and that false negatives (also called errors of omission) run at around 20%. Given the time lapse since that original study, we thought it prudent to repeat the study. To that end, we undertook the matching of the surname of every member of the 1988 NHIS against the Passel-Word file. We have done that and can now compare the self-reported ethnicity with the ethnicity that would be imputed on the basis of surname using this file. At the same time, we analyzed some of the characteristics of persons with nonconforming names with the idea that this information should be useful in decisions about sample allocation.

Methodology

From the 1988 NHIS file we formed an extract containing a list of variables which might help explain the relationship between surname and self-declared ethnic origin. The variables we considered were education, income, poverty status, metropolitan status, urbanicity, marital status, size of metropolitan area, Census Division, Census Region, sex, age, family size, and detailed self-reported Hispanic Origin. For those person in the NHIS who did not declare ethnicity, we made the decision to consider them as nonHispanic. The extracted 1988 file was then merged by survey ID to the surnames of the individuals (ordinarily kept apart from the rest of the information). We removed obvious embedded titles such as "Jr.," "Sr.," "III," etc. We also removed embedded blanks and converted all lower case to upper case so that "De Jesus" became "DEJESUS." We did not remove hyphens or match each component of a hyphenated name separately against the Passel-Word list. We did not exclude persons with such obvious non-names as "DOE" and "REFUSED." The merged NHIS file was then merged by surname with the Passel-Word file. As a result of the merge, we created a Hispanic surname indicator for each NHIS respondent. Subsequently, we ran weighted frequency counts of the Hispanic origin indicator for the individuals with Hispanic surnames and of the Hispanic surname indicator for the individuals who have declared themselves of Hispanic origin.

Results

The attached tables contain the observed false positive and false negative rates crossed by the variables mentioned above. The overall false positive rate is

12.6 percent while the overall false negative rate is 31.6 percent. The former rate is lower than we had expected based upon previously published research; while the later is higher than expected. Passel and Word originally reported error rates with the 1976 March Current Population Survey (CPS) of 15.0% false positive and 20.7% false negative. Part of the discrepancy involves persons who did not classify their ethnicity, who had a hyphenated name or who refused to provide their name, all of which Passel and Word treated differently than we did. An inspection of nonconforming surnames of Hispanics indicates, however, that even allowing for these variations in matching procedure, there is a substantial and unexplained difference in the false negative rate. The reason for the poorer performance could either be lower quality recording, transcription, and keying of names on the 1988 NHIS than on the 1976 March CPS (interviewers, not respondents record name spellings) or real change in the population in the relationship between surname and self-reported ethnicity. Both rates are higher for the elderly: 16.6% false positives and 33.6% false negatives.

Errors are far more common among women than among men. Intermarriage evidently plays a strong role in both error rates. Widowed and divorced Hispanics are quite likely to have surnames that aren't on the Passel-Word file. At the other extreme, married Hispanics that are neither living with their spouses nor formally separated from them can be covered quite well via their surnames. False positive rates are sharply higher among the ever married than among the never married, with the exception of the group that are neither together nor separated. Perhaps that group contains a disproportionate number of new immigrants and migrant workers.

Socio-economic status also plays a strong role in both error rates. The general trend seems to be that higher socio-economic status means a weaker association between surname and self-reported ethnic origin. For example, the false negative rate climbs monotonically by level of education completed from elementary through post graduate. False positive rates follow pretty much the same trend with a slight dip from college graduate to post-graduate achiever. (The "None" category on education is very different.) Both error rates bounce around a bit across the low income classes. However, both rates are substantially higher for the middle and upper income classes than for the low income classes, and there are monotonic upward patterns in the error rates among the middle income classes. If information about family size is available, the combination of family size and surname is a very powerful indicator. (Of course this last point would be more useful in an imputation project than in a sampling project.)

The sensitivity and specificity of the surname indicator is better in central cities than in the suburbs. The surname indicator works better in the very large metropolitan areas than in the smaller metropolitan areas. Sampling error may be a factor, but we found amazingly bad performance of the indicator in MSAs with under 100,000 population. The indicator does not work well in nonmetropolitan areas. Performance is slightly worse in rural nonmetropolitan areas than in urban nonmetropolitan areas, except for rural farm areas, where it works very well.)

Across the divisions, the surname indicator works best in the West South Central (TX, OK, AR, and LA) Division and worst in the West North Central (the northern plains) and East South Central (KY, TN, MS, and AL) and New England Divisions.

Among the detailed categories of Hispanic origin, the sensitivity of the indicator is best for those from Mexico. It is a little worse for those from Puerto Rico. It is not good for those from Cuba or other Latin American countries. It is particularly poor for those with multiple Hispanic backgrounds, Hispanics from outside Latin America and Hispanics who do not identify with a more detailed origin.

Implications for Sample Design

Although the estimated sensitivity of the Passel-Word Hispanic surname list was not as good as we had hoped, the idea of using it to create a list sample of elderly Hispanics still has good potential. The technique is still far cheaper than area sampling with screening while maintaining about the same level of biases. In this section, we explain these implications more fully. Table 7 contrasts some of the numbers that appear in the following text.

Under current plans for 1995 and beyond, the area/permit NHIS sample will yield nominal elderly Hispanic sample sizes of about 500 males and 700 females. After accounting for the design effect due to disproportionate sampling of heavily Hispanic blocks, the effective sample sizes (compared to a similarly clustered sample with equal probabilities) will only be about 350 and 500. By adding 1000 males with Hispanic surnames from SSA lists and another 1000 females in the same manner, we can boost the nominal sample sizes to around 1240 and 1400 and the effective sample sizes to around 770 and 920. (The effective sample sizes don't increase as much as the nominal sample sizes because of the design effect due to the large weights that Hispanics without Hispanic surnames will bear.) To get a comparable boost from the area/permit sample alone would require the screening of an additional 100,000 households!

As another contrast, suppose that one used SSA lists as a supplemental list sampling frame without paying attention to surname. In that case, more than 20,000 persons would have to be located and screened on ethnicity in order to get comparable boosts for elderly Hispanics. Thus, although screening a list of elderly persons is far more efficient than screening a sample of households, even if those households were heavily skewed toward heavily Hispanic blocks, it is nowhere near as efficient as using a list in combination with the surname list.

Nonetheless, we had hoped to boost the effective sample sizes even more sharply for elderly Hispanics, to as high as 1000 males and 1000 females. Here it turns out that the false negative rate is too high to make that economical. About 5000 list persons with Hispanic surnames would have to be added to the sample.

We are thus extremely interested in ways to decrease the false negative rate, even if it means some increase in the false positive rate, as it likely would. If for example, the 1976 findings of Passel and Word still held, the supplemental sample size required for the desired effective sample sizes by sex would be just 2400 instead of 5000. (Even with a perfect indicator of Hispanic origin on the list, a supplemental sample of 1350 persons would be required to achieve the desired effective sample sizes of 1000 by sex.) We suspect that adjustments in interviewer training and in keypunching could reduce the frequencyof false negatives. (Even though we would be sampling from an SSA list that would probably have higher quality name spelling, we would still need to classify everyone we would still need to classify everyone in the area/permit smaple in order to work out appropriate sampling weights for the dual-frame estimator.)

It is possible, however, that there has been a sea-change in the relationship between Hispanic origin and surname and that improvements will be difficult.

David Word (in a personal communication) doubts that such a change has occurred. He is conducting research similar to ours on the much larger sample in the Census Post Enumeration Survey. It will be very interesting to compare results.

References

Massey, J.T., Moore, T.F., Parsons, V.L. and Tadros, W. (1989). Design and Estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, Vital Health Stat 2(110).

Passel, J.S. and Word, D.L. (1980). "Constructing the List of Spanish Surnames for the 1980 Census: An Application of Bayes' Theorem." Presented at the Annual Meeting of the Population Association of America.

Table 1. Sensitivity and specificity of Hispanic surmane as a surrogate for self-reported Hispanic origin by education, income, and poverty status

	Sensitivity of Hispanic surname		Specificity of Hispanic surname	
Characteristic	Estimated persons of Hispanic origin	False negative rate	Estimated persons with Hispanic surnames	False positive rate
	(in thousands)		(in thousands)	
Education				
< 5 years (NA)	2,147	32.1	1,651	11.8
None	1,255	27.2	997	8.4
Elementary	6,065	24.6	4,839	5.5
Some high school	2,718	28.5	2,159	10.0
High school graduate	3,773	36.1	2,976	19.0
Some college	2,045	39.9	1,550	20.7
College graduate	685	47.0	505	28.0
Post college	498	50.2	322	23.0
Unknown	208	35.4	196	31.7
Income (thousands)				
< 5	1,089	29.0	844	8.5
5-6	768	29.8	570	5.5
7-9	1,329	22.3	1,114	7.3
10-14	2,186	29.4	1,669	7.5
15-19	2,345	25.2	1,970	11.0
20-24	1,615	28.1	1,321	2.2
25-34	2,738	32.3	2,118	12.5
35-50	2,521	37.5	1,951	19.3
50+	1,636	45.8	1,163	23.8
Unknown	3,167	31.9	2,476	12.9
Poverty Status				
Above	13,245	34.2	10,253	15.1
Below	3,992	22.9	3,257	5.5
Unknown	2,157	31.0	1,686	11.7

Source: 1988 National Health Interview Survey (NHIS). 9600 NHIS sample persons reported Hispanic origin and 7500 NHIS sample persons gave a surname listed on the Passel-Word Hispanic surname file. Estimates shown in table are weighted to U.S. population.

Table 2. Sensitivity and specificity of Hispanic surname as a surrogate for self-reported Hispanic origin by metro and urban-rural status

	Sensitivity of Hispanic surname		Specificity of Hispanic surname	
Characteristic	Estimated persons of Hispanic origin	False negative rate	Estimated persons with Hispanic surnames	False positive rate
	(in thousands)		(in thousands)	
Metro-status				
Central city	10,081	26.0	8096	7.8
Suburb	7,688	36.1	5930	17.1
Non metro	1,624	44.9	1170	23.5
Urbanicity				
Urban	17,440	30.5	13783	12.0
Rural/non-farm	1,746	44.1	1239	21.3
Farm	205	19.1	170	2.2
MSA				
1,000,000 or more	11,722	29.2	9353	11.2
250,000 - 999,999	5,283	31.5	4119	12.1
100,000 - 249,999	668	35.1	517	16.0
Under 100,000	96	77.5	37	40.9
Non-MSA				
Other urban	835	44.7	562	17.9
Rural	789	45.1	607	28.6

Source: 1988 National Health Interview Survey (NHIS). 9600 NHIS sample persons reported Hispanic origin and 7500 NHIS sample persons gave a surname listed on the Passel-Word Hispanic surname file. Estimates shown in table are weighted to U.S. population.

Table 3. Sensitivity and specificity of Hispanic surname as a surrogate for self-reported Hispanic origin by census region and division

	Sensitivity of Hispanic surname		Specificity of Hispanic surname	
Characteristic	Estimated persons of Hispanic origin	False negative rate	Estimated persons with Hispanic surnames	False positive rate
	(in thousands)		(in thousands)	
United States	19,393	31.6	15195	12.6
Region				
Northeast	3,285	34.0	2,553	15.1
Midwest	1,691	37.8	1,318	20.2
South	6,397	27.3	5,133	9.4
West	8,021	32.6	6,191	12.7
Division				
New England	527	41.1	409	24.0
Mid Atlantic	2,758	32.7	2,144	13.4
East North Central	1,443	33.5	1,190	19.4
West North Central	248	63.1	128	28.4
South Atlantic	1,931	39.0	1,389	15.3
East South Central	140	57.9	111	47.0
West South Central	4,326	21.1	3,632	6.0
Mountain	1,282	39.3	926	15.9
Pacific	6,738	31.3	5,266	12.1

Source: 1988 National Health Interview Survey (NHIS). 9600 NHIS sample persons reported Hispanic origin and 7500 NHIS sample persons gave a surname listed on the Passel-Word Hispanic surname file. Estimates shown in table are weighted to U.S. population.

Table 4. Sensitivity and specificity of Hispanic surname as a surrogate for self-reported Hispanic origin by sex, age, and marital status

	Sensitivity of Hispanic surname		Specificity of Hispanic surname	
Characteristic	Estimated persons of Hispanic Origin	False negative rate	Estimated persons with Hispanic Surnames	False positive rate
	(in thousands)		(in thousands)	
Sex				
Male	9,452	27.7	7,597	10.1
Female	9,941	35.2	7,598	15.2
Age				
0-4	2,147	32.1	1,651	11.8
5-17	5,132	31.4	3,832	8.2
18-24	2,600	31.7	2,019	12.0
25-44	6,180	31.7	4,990	15.4
45-64	2,464	30.2	2,010	14.4
65+	871	33.6	694	16.6
Marital Status				
Under 14 yrs.	5,793	31.6	4,364	9.1
Married, spouse in household	7,302	31.1	5,973	15.8
Married, spouse not in household	280	22.7	242	10.5
Widowed	467	38.2	349	17.3
Divorced	773	38.8	568	16.6
Separated	414	31.6	332	14.8
Never Married	4,282	30.9	3,271	9.5
Unknown	83	30.8	96	40.2

Source: 1988 National Health Interview Survey (NHIS). 9600 NHIS sample persons reported Hispanic origin and 7500 NHIS sample persons gave a surname listed on the Passel-Word Hispanic surname file. Estimates shown in table are weighted to U.S. population.

Table 5. Sensitivity and specificity of Hispanic surname as a surrogate for self-reported Hispanic origin by family size

	Sensitivity of Hispanic surname		Specificity of Hispanic surname	
Family size	Estimated persons of Hispanic origin	False negative rate	Estimated persons with Hispanic surnames	False positive rate
	(in thousands)		(in thousands)	
1	1,285	44.5	873	18.3
2	2,447	39.3	1,800	17.5
3	3,326	36.5	2,420	12.7
4	4,567	32.1	3,689	15.9
5	3,635	29.5	2,881	11.0
6	2,044	21.4	1,754	8.5
7	962	27.6	737	5.3
8	466	8.7	465	8.4
9+	660	13.6	577	1.2

Source: 1988 National Health Interview Survey (NHIS). 9600 NHIS sample persons reported Hispanic origin and 7500 NHIS sample persons gave a surname listed on the Passel-Word Hispanic surname file. Estimates shown in table are weighted to U.S. population.

Table 6. Sensitivity of Hispanic surname as a surrogate for self-reported Hispanic origin by Hispanic subgroup

Hispanic subgroup	Sensitivity of Hispanic Surname	
	Estimated persons of Hispanic origin	False negative rate
	(in thousands)	
Multiple Hispanic	321	52.2
Puerto Rican	2,417	30.3
Cuban	1,119	35.0
Mexican-Mexican	3,558	19.4
Mexican-American	6,927	23.9
Chicano	139	24.4
Other Latin American	1,977	37.1
Other Spanish	2,381	57.8
Spanish, DK type	552	60.9

Source: 1988 National Health Interview Survey (NHIS). 9600 NHIS sample persons reported Hispanic origin and 7500 NHIS sample persons gave a surname listed on the Passel-Word Hispanic surname file. Estimates shown in table are weighted to U.S. population.

Table 7. Required sample sizes by sampling method and precision target

Method	Sex	Effective Sample Size	Screener Interviews	Effective Sample Size	Screener Interviews
SSA list with match to Passelword file	M	770	1,000	1,000	3,050
	F	920	1,000	1,000	1,950
			2,000		5,000
SSA list without match to Passelword file	M	700	9,750	1,000	15,100
	F	920	10,350	1,000	12,500
			20,100		27,600
Area sample	M	770		1,000	
	F	920		1,000	
			100,000*		180,000*
SSA list with match to Passelword file under 1976 error rates	M			1,000	1,280
	F			1,000	1,120
					2,400
SSA list with match to Passelword file under 0 error rates	M			1,000	760
	F			1,000	590
					1,350

*Households, on top of planned 99,000.