Propensity Stratification with Auxiliary Data for Address-Based Sampling Frames

Jamie L. Ridenhour[1], Joseph P. McMichael[1]
[1]3040 Cornwallis Road, Research Triangle Park, NC 27709

Abstract
Auxiliary information appended to Address-Based Sampling (ABS) frames can be incorporated at the sample design stage to increase the likelihood of sampling addresses with target groups of interest; however, this information needs to be accurate and nonmissing. For much of the auxiliary data currently available the drawbacks are that its accuracy is unknown and it is missing for many addresses on the sample frame making it difficult to use for stratification. In this paper we discuss how we solve this problem by using data collected from a nationally representative household survey of youths ages 11-16 to create a propensity model for stratification. This propensity model is then applied to the address frame for a subsequent survey targeting the same age group in rural areas and stratifies those addresses by how likely they were to have members of the eligible population. We discuss the impact this approach had on survey efficiency, both in terms of field cost and variance.

Key Words: Address-Based Sampling, ABS, propensity stratification

1. Background

The Computerized Delivery Sequence file (CDS) from the United States Postal Service (USPS) is the frame for Address-Based Sampling surveys. The CDS, when combined with the No-Stat file, contains all postal delivery points serviced by USPS (Shook-Sa et al.). RTI receives monthly updates of the CDS from a qualified vendor.

1.1 Components of CDS

The CDS contains the components of a mailing address: the first and second line of the address, city, state, ZIP, and ZIP+4 (see the left portion of Figure 1). Though the CDS contains all postal delivery points it lacks additional information potentially useful for sample design. Ways to make use of any additional information include stratifying the addresses for sample selection and oversampling members of the target population or key subgroups of interest.

1.2 Auxiliary Data

Additional data from other sources can be merged onto the CDS forming RTI's Enhanced ABS Frame. While there are numerous fields on the Enhanced Frame, examples include flags for children in specified age groups, adult age groups, and whether the surname of a person at the address is likely Hispanic (see the right portion of Figure 1).

The utility of the auxiliary data is limited by its completeness, or lack thereof. Flags are not available for all addresses and for most variables there is no way to discern between the address lacking a particular attribute and the information just not being available.

Components from the CDS					Auxliary information from commercial databases					
Address	City	State	ZIP	ZIP+4	Child 0-5	Child 6-11	Child 12-17	Adult 18-25	Adult 65+	Hispanic surname
101 Main Street	Raleigh	NC	12345	6789	Y	Y				
103 Main Street	Raleigh	NC	12345	6789						Y
…										
200 First Lane Apt 1	New Orleans	LA	67890	1234				Y		
200 First Lane Apt 2	New Orleans	LA	67890	1234				Y		
200 First Lane Apt 3	New Orleans	LA	67890	1234					Y	
200 First Lane Apt 4	New Orleans	LA	67890	1234			Y			
200 First Lane Apt 5	New Orleans	LA	67890	1234						
…										

Figure 1: Example CDS and auxiliary information.

We sought to reduce the missingness of the auxiliary data on the Enhanced ABS Frame to increase its utility for oversampling our target population.

1.3 Research Application

For a longitudinal evaluation of an anti-tobacco media campaign (funded by the US Food and Drug Administration) our target population was males ages 11-16 living in "rural" areas across the country. Nationwide, the proportion of households with a male 11-16 is 9%. We had not previously fielded a study for this target population but we had fielded a study for a similar target population: *all youths* ages 11-16 nationwide. This prior study was also an in-person ABS survey and for all screened households we knew whether the household contained at least one member of our target population or not.

2. Methods

2.1 Regression Model

The prior survey consisted of a national sample of roughly 45,000 addresses. The same auxiliary data on the Enhanced Frame that we intended to use for the new study was available for the prior study from the time at which the sample was drawn. Additionally, we had the ground truth for the portion of addresses which were screened: whether or not the household had at least one youth ages 11-16.

For the new survey we had the address frame in the selected rural areas along with the auxiliary data from the Enhanced Frame and we wanted to calculate a propensity of the address being eligible for the survey (i.e., the address has at least one male ages 11-16).

We took the auxiliary data available on the Enhanced Frame for the survey which had already been fielded and developed a logistic regression model where the Y variable was whether or not the address had someone in the target population. We then took the

estimated model and used it on the frame for the new survey to calculate a propensity of being eligible for each address on the frame.

2.1.1 Prior Study Data

Of the 45,000 sampled addresses from the old survey there were approximately 23,000 with a known eligibility status. In developing the regression model we tested out many variables from the Enhanced Frame including income categories, adults in age groups (e.g., 25-34, 35-44, 60+), youths in eligible age ranges, and address information (e.g., vacant, seasonal, multi-family, high-rise). Neither the income variables nor the address variables proved to be useful predictors.

While the variables in the auxiliary data typically only have an indication of whether the attribute is true and we cannot distinguish between not true and missing information, sometimes this information is contained in other variables. For example, a variable that ended up being a good predictor for this regression model was the variable indicating that the youngest adult at the address was 60 or older. While other variables were useful for picking out addresses likely to be eligible this variable was useful for identifying the low propensity addresses (those unlikely to have at least one male ages 11-16).

2.1.2 Applying the model

The parameter estimates from the model developed on the prior study data were then applied to the 537,000 addresses on the Enhanced Frame for the new study and a predicted probability of having at least one eligible male 11-16 was calculated for each address on the frame.

We then used a clustering algorithm (SAS PROC FASTCLUS) to group the addresses into six propensity strata.

Table 1: Propensity Strata, with Data Collection Results

Stratum	Predicted Eligibility	Frame		Sample		Screening Rate	Observed Eligibility
1	2.9%	32,850	6.1%	3,014	4.8%	43.9%	1.2%
2	4.4%	150,625	28.1%	4,197	6.7%	42.0%	2.5%
3	9.3%	224,887	41.9%	13,595	21.6%	36.5%	7.5%
4	13.3%	111,866	20.8%	34,073	54.1%	35.2%	16.1%
5	25.6%	13,679	2.5%	6,782	10.8%	34.7%	35.8%
6	30.0%	2,781	0.5%	1,339	2.1%	34.8%	42.9%

3. Results

The six propensity strata and their attributes are shown in Table 1. Stratum 1 was the lowest propensity stratum with a predicted eligibility of 2.9% whereas stratum 6 was the highest propensity stratum with a predicted eligibility of 30.0%.

Comparing the frame and sample distributions we were oversampling the three high propensity strata (strata 4-6) and undersampling the low propensity strata (strata 1-3). During data collection for the new survey we were able to screen a greater proportion of the low propensity strata compared with the high propensity strata. The observed eligibility rates panned out nicely: our low propensity strata had low eligibility rates and our high propensity strata had high propensity rates.

We sampled 63,000 addresses to find 2,000 males ages 11-16 and our data collection costs would have been significantly higher if not for our ability to accurately oversample likely eligible households. While the household eligibility rate is 9% nationally, in our sample 15.4% of households were eligible (had at least one male 11-16). We also saved on data collection costs because we utilized a mail screener to identify eligible households prior to in-person data collection. The screener made no mention of the target population (males 11-16) or that it was a survey about tobacco use. Our oversample was not free – the disproportionate sampling rates across propensity strata resulted in an overall unequal weighting effect of 3.1.

We found that auxiliary data appended to an ABS frame can be effectively used to find less common populations of interest when prior field data can be used to ameliorate the missingness of the auxiliary data. As we amass more field data we think we will be able to refine our regression models. Currently we are only utilizing main effects but future research would look at the inclusion of interaction terms. As we also gain more information on true eligibility and cost we will be able to come up with a better optimal allocation where we either need less sample for the sample overall unequal weighting effect or we are able to achieve a lower unequal weighting effect for the same sample size.

References

Shook-Sa, Bonnie E, et al. "Extending the Coverage of Address-Based Sampling Frames Beyond the Usps Computerized Delivery Sequence File." *Public Opinion Quarterly* 77.4 (2013): 994-1005. Print.