Evaluation of Combining Consumer Marketing Data Used for Address-Based Sampling

Joseph McMichael
Rachel Harter
Bonnie Shook-Sa
Jamie Ridenhour
Joey Morris

August 5, 2014
Joint Statistical Meetings
Boston, MA

ABS & Auxiliary Data

ABS Frame

- USPS Computerized Delivery Sequence (CDS) file
- elements from No-Stat file
- contains all US mail delivery points
- addresses serve as a proxy for US households

Consumer Marketing Databases

- used for direct mail, phone and email marketing
- demographic, socio-economic, spending habits, hobbies, etc...
- information about areas, households and persons
- public records, DMV, retailers, surveys, data sharing.

Acxiom, Epsilon, Experian, InfoUSA, KBM, Targus

Acxiom InfoBase

- Adult person level file
- > 1000 fields
- > 400 million records
- > 500 GB (uncompressed flat file)
- High rates of missing for many fields

- Address
- Child Age – 1 year increment
- Surname

Enhanced ABS Frame (ABS + Acxiom)

Combine ABS Frame with Acxiom data

Rollup Acxiom to *address-level* [created indicators]

- Number of children age X [**Child Age X**]
- Number surnames [**Surname**]
- Number of Hispanic surnames [**Hispanic Surname**]

85% of address-level Acxiom merged to CDS

90% of these have at least one surname

77% of Enhanced ABS Frame have at least one surname

Goal: Increase Sample Efficiency

Scenario: Sampling a rare population.

Disproportionate Stratification (Kalton 2009, Waksberg 1973)

- Create higher density stratum (i.e. – higher prevalence eligible)
- Oversample higher density stratum
- Decrease cost *or* Increase precision

Two Stratum Design (High Density and Low Density)

Address-level stratification by appended eligibility indicators

- Addresses with eligibility indicator placed in high density stratum

Goal: Increase Sample Efficiency

Evaluating Efficiency

- Compare Allocation: Optimal vs. Proportional
- Cost Reduction for equal precision

Key Factors Influencing Efficiency

- Prevalence of the eligible population
- Eligibility Indicator Append Rate to ABS Frame
- **Eligibility Indicator Accuracy**
- Cost relationship between screening and primary data collection.

Study: a community health survey

- Large community-based survey
- Target population contains > 5 million HH
- ABS Sampling Frame
- Two modes of data collection: Outbound CATI, PAPI
- 30,000 successfully screened
- 20,000 responding households

Evaluated auxiliary data:

- Households w/ Child age 3-17
- Hispanic Households w/ Hispanic Surname Flags

Study: a community health survey

	Confirmed NO child 3-17	Confirmed child 3-17	
No Predicted Child 3-17	16,760	6,701	23,461
Predicted Child 3-17	2,348	3,587 **(True Positive)**	5,935 **(Total Predicted)**
	19,108	10,288	29,396

Eligibility Indicator Accuracy

True Positive / Total Predicted = **0.60**

Study: a community health survey

Eligibility Indicator Accuracy

- HH w/ Child 3 - 17 = 60.0%
- HH w/ Hispanic = 78.2%
- HH w/ Hispanic Child 3 - 17 = 59.9%

Households w/ Children 3 - 17

Prevalence Elig Pop	Append Rate Elig Indicator	Accuracy Elig Indicator	Cost Relationship (Scrn/Main)	Optimal Allocation UWE	Cost Comparison (Optimal/Prop)
28%	7.3%	60%	0.10	1.00	1.00
28%	7.3%	60%	0.50	1.02	0.99
28%	**7.3%**	**60%**	**1.00**	**1.04**	**0.98**
28%	7.3%	60%	2.00	1.06	0.97
28%	7.3%	60%	10.00	1.08	0.96

Households w/ Hispanics

Prevalence Elig Pop	Append Rate Elig Indicator	Accuracy Elig Indicator	Cost Relationship (Scrn/Main)	Optimal Allocation UWE	Cost Comparison (Optimal/Prop)
14%	3.7%	78%	0.10	1.08	0.94
14%	3.7%	78%	0.50	1.29	0.75
14%	**3.7%**	**78%**	**1.00**	**1.39**	**0.68**
14%	3.7%	78%	2.00	1.47	0.63

Households w/ Hispanic Children Age 3 - 17

Prevalence Elig Pop	Append Rate Elig Indicator	Accuracy Elig Indicator	Cost Relationship (Scrn/Main)	Optimal Allocation UWE	Cost Comparison (Optimal/Prop)
6.40%	0.39%	60%	0.10	1.23	0.77
6.40%	0.39%	60%	0.50	1.49	0.52
6.40%	**0.39%**	**60%**	**1.00**	**1.59**	**0.46**
6.40%	0.39%	60%	2.00	1.67	0.42

Contact

Joseph McMichael
Statistician
919.485.5519
mcmichael@rti.org