LOCATION AND RESPONSE PROPENSITY MODELING FOR THE 1995 NATIONAL SURVEY OF FAMILY GROWTH

Vincent G. Iannacchione, Research Triangle Institute
P.O.Box 12194, Research Triangle Park, NC 27709

Key Words: Linked Sample Design, Survey Nonresponse, Logistic Regression, ROC Curves

Abstract

Distinct patterns of location and response propensity were found among women selected for Cycle 5 of the National Survey of Family Growth (NSFG-5). For example, minority women were harder to locate than were other sample women. Once found however, these same women were more likely to participate in the survey than other sample women. Two logistic regression models were developed so that predictors related to the locating process could be distinguished from those related to the cooperation process. Receiver Operating Characteristics (ROC) curves were used to assess the overall predictive ability of the models. The linkage of the NSFG-5 to the 1993 National Health Interview Survey (NHIS) enabled a large number of candidate predictors to be considered for each model. As expected, predictors indicating the presence or absence of NHIS contact data (e.g., a telephone number) were significant factors affecting location propensity. The absence of contact data also affected response propensity (especially when accompanied by other item nonresponse such as income, a traditionally sensitive item), and was interpreted as an indication of resistance to participate in both surveys. The predicted location and response propensities obtained from the models were used to compute non-response adjustment factors for the sampling weights.

Introduction

The sample for the NSFG-5 (Potter et al 1998) consisted of a subsample of 14,000 women between the ages of 15 and 44 from households that participated in the 1993 NHIS (Massey et al 1989). The objectives of linking the two surveys were to reduce the cost of the NSFG-5 while maintaining the statistical accuracy of the survey estimates, and to expand analytic opportunities by linking data from the NHIS to the NSFG-5.

The linkage between the NHIS and the NSFG-5 provided a large amount of data about NSFG-5 sample members including those who could not be located and those who were located but refused to participate in the NSFG. Between the 1993 NHIS and the 1995 NSFG, many women in the sample moved, and substantial effort was made to identify their new addresses. Tracing activities were successful in locating 94.6% of all sample women. The participation rate among sample women who were located and eligible for the survey was 83.2% The overall response rate among eligible sample women was 78.7%.

When available, NHIS information about the sample woman was used for tracing as were contact data for the NHIS reference person (typically a spouse) and the NHIS contact person (usually a relative, neighbor, or friend). About 38% of all sample women had one or more pieces of tracing information missing: 31% had missing Social Security Numbers (SSNs), 25% had no contact person listed, and 11% either did not have a telephone or refused to give their telephone numbers.

Some of the variables from the NHIS indicated varying degrees of ability to locate while others indicated varying degrees of resistance or hostility to surveys. For example, failure to obtain a telephone number was related to inability to locate, whereas refusal to give an SSN or the name of a contact person indicated resistance to participate. These distinctive patterns suggested that the location process be treated as a different outcome variable than the cooperation process among those who were located. As Groves and Couper (1998) point out, the location and cooperation processes are different in most household interview surveys. As a result, the bias caused by non-contact usually is not the same as the bias caused by refusal to participate.

Development of the Models

Response propensity weight adjustments (Folsom 1991) uses logistic regression to model the functional relationship between a set of response predictors and a (dichotomous) response outcome. If the relationship is significant, and if the response propensities are non-zero over conceptual repetitions of the study, the model-based adjustment factors applied to the sampling weights greatly reduce the potential for nonresponse bias. In addition, response propensity modeling provides a formal statistical setting for evaluating variables believed to be related to response. This was particularly useful for evaluating the large number of potential predictors available from the NHIS database.

Two logistic regression models were developed for NSFG-5 sample women so that predictors related to the locating process could be distinguished from those related to the cooperation process. The two models enabled separate adjustment factors to reflect the distinct patterns of availability, including change of address, lack of some or all contact information, and resistance to participation. Mobility was an expected artifact of the NSFG-5 sampling design because of the linkage to the 1993 NHIS and the long time period between the two surveys. The lack of contact information was generally considered as indirect resistance to participation but also represented, in some cases, a failure to collect accurate and complete contact information during the NHIS interview

Segmentation analysis using the CHAID software (Magidson 1993) was used to detect interactions among the set of potential predictors. This allowed for a parsimonious model-building process that focused on segments that showed distinct location and response propensities, and avoided the dilemma of examining "all possible interactions" that is inherent to regression modeling. For design consistency, a weighted segmentation of the sample using CHAID was performed and then included with main-effect predictors in the logistic regression procedure in SUDAAN (Shah et al 1997).

The overall response propensity for each sample woman *i* was subdivided into the following components.

$$L_i = \begin{cases} 1 & \text{if sample woman i was located, and} \\ 0 & \text{Otherwise} \end{cases}$$

and:

$$R_i = \begin{cases} 1 & \text{if sample woman i responded, and} \\ 0 & \text{Otherwise} \end{cases}$$

Then, the overall probability that sample woman *i* responded was written as:

$$\begin{aligned} P[R_i=1] &= P[L_i=1] \cdot P[R_i=1 \mid L_i=1] \\ &= \lambda_i \quad \cdot \quad \rho_i \end{aligned}$$

The overall probability of response was estimated with two logistic regression models. The first model for location propensity was applied to the entire sample of 14,000 sample women. The second model for response propensity was applied to the 13,038 sample women who were located and eligible for the study.

Location Propensity Model

The following logistic model was developed to estimate the probability that sample woman *i* was located:

$$\begin{aligned} \hat{\lambda}_i &= P[\, L_i=1 \mid X_i, \hat{\beta}\,] \\ &= [\, 1 + \exp(-X_i\, \hat{\beta}\,)\,]^{-1} \end{aligned}$$

where $X_i \equiv$ a vector of NHIS location predictors

The logistic regression coefficients $\hat{\beta}$ were estimated iteratively to satisfy the following estimation equations:

$$\sum_{i\in S}(W_i \div \hat{\lambda}_i)\, X_i^T \hat{\lambda}_i = \sum_{i\in S}(W_i \div \hat{\lambda}_i)\, X_i^T L_i,$$

where S = Sample of 14,000 women, and
W_i = Initial NSFG-5 sampling weight.

Then, the location adjusted weight was computed as

$$W_i^L = \begin{cases} W_i \div \hat{\lambda}_i & \text{if } L_i = 1 \\ 0 & \text{if } L_i = 0 \end{cases}$$

This location adjusted weight was used to develop the following response propensity model.

Response Propensity Model

The probability of participation given that sample woman *i* was located and eligible was estimated as:

$$\begin{aligned} \hat{\rho}_i &= P[R_i=1 \mid L_i=1, Z_i\hat{\theta}\,] \\ &= [1 + \exp(-Z_i\hat{\theta}\,)]^{-1} \end{aligned}$$

where $Z_i \equiv$ a vector of NHIS response predictors.

Analogous to the location propensity model, the logistic regression coefficients $\hat{\theta}$ were estimated iteratively to satisfy the following estimation equations:

$$\sum_{i\in\xi} (\, W_i^L \div \hat{\rho}_i\,)\, Z_i^T \hat{\rho}_i = \sum_{i\in\xi} (\, W_i^L \div \hat{\rho}_i\,)\, Z_i^T R_i$$

where ξ = Sample of 13,038 located women.

Then, the response-adjusted weight was computed as:

$$W_i^R = \begin{cases} W_i \div (\hat{\lambda}_i \hat{\rho}_i) & \text{if } R_i = 1 \\ 0 & \text{if } R_i = 0 \end{cases}$$

The most influential components of Z_i, the vector of predictors for the response propensity model, and X_i, the vector of predictors for the location propensity model, are described in the next sections.

Factors Affecting Location Propensity

As expected, predictors indicating the presence or absence of NHIS contact data were significant factors in the final location propensity model. The segmentation of the sample shown in **Figure 1** suggests that these predictors interacted with a number of demographic factors, especially family income. For example, among sample women with low or unknown family incomes, the presence or absence of a telephone number resulted in a 10% difference in the location rate. In fact, the lowest segment-level location rate (63.4%) occurred among sample women with low or unknown family incomes who either refused or did not have telephone numbers and who completed the NHIS but did not provide their names. In contrast, the lack of segmentation among sample women with (known) annual family incomes of $20,000 or more suggests that sample women who were willing to provide income (a traditionally sensitive item) were also likely to provide contact information.

In addition to the interactive effects identified by the segmentation analysis, several demographic variables were significant "main effect" factors in the location propensity model. For example, after adjusting for other covariates in the model, unmarried women were significantly harder to locate than their married counterparts. Similarly, women with no college education had lower location propensities than those with some college. Region of the country was also a significant factor with women in the Midwest located more easily than those in other regions of the U.S.

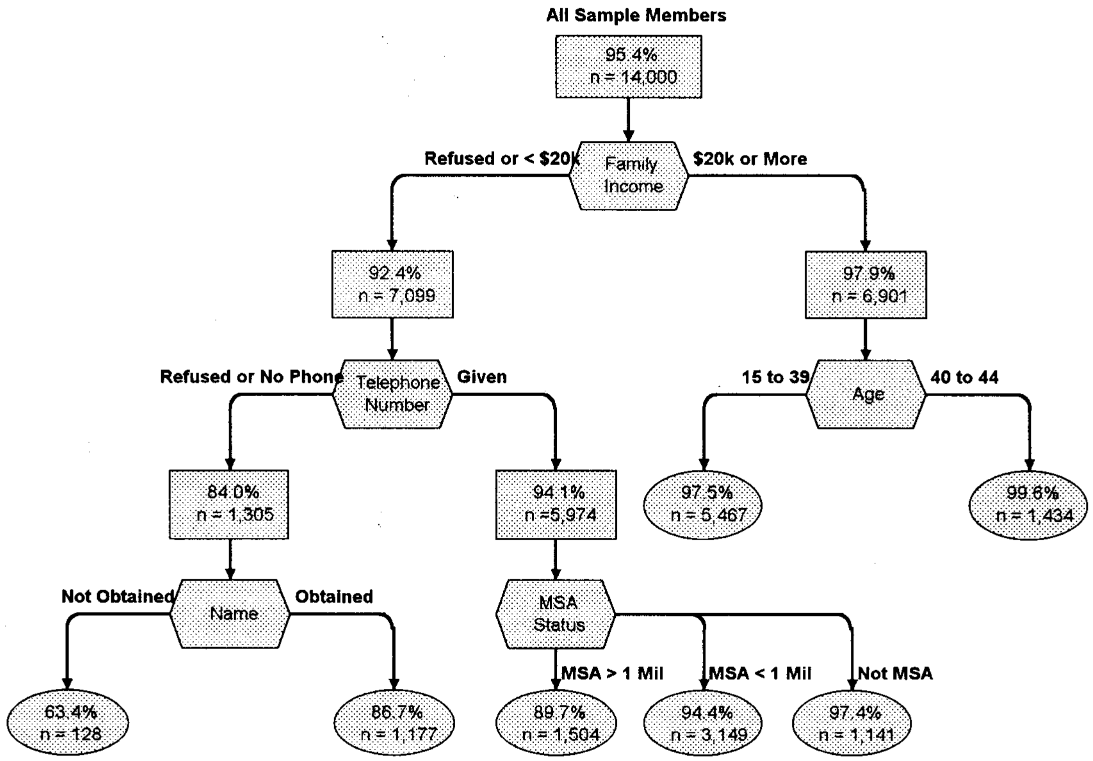

Figure 1. Segmentation of NSFG-5 sample members by weighted location rate

Factors Affecting Response Propensity

As in the location propensity model, several of the predictors related to the presence or absence of contact data also were significant in the response propensity model. For example, more than 2,000 sample women refused to provide a SSN during the NHIS but were subsequently located and found eligible for the NSFG-5. Once located, however, these sample women were significantly less likely to participate than those who provided a SSN. Similar patterns can be seen for those who refused to provide a telephone number or the name of a contact person in the segmentation modeling of located eligibles shown in **Figure 2**. Unlike the location model, the absence of contact data in the response model was thought to be an indication of hostility to the interview process.

Although not significant in the location propensity model, both race and Hispanic background were important response predictors. Among sample women who provided a SSN to the NHIS, Asians and Pacific Islanders were less likely to participate than other racial groups. Among those who refused to provide a SSN, the participation rate among Hispanic women was 15 percent higher than among non-Hispanic women. In fact, the 92 percent participation rate among non-working Hispanic women who refused to provide (or did not have) SSNs to the NHIS is a notable exception to the general trend of being able to predict hostility to the NSFG-5 by hostility to the NHIS.

The significant non-interactive (i.e., main effect) factors in the response propensity model included age, number of children, and number of health conditions. Sample women between 15 and 24 were more likely to participate than older women while those with one or no children were less likely to participate than those with two or more children. Also, sample women with no reported health conditions were less likely to participate than those who reported one or more health conditions.

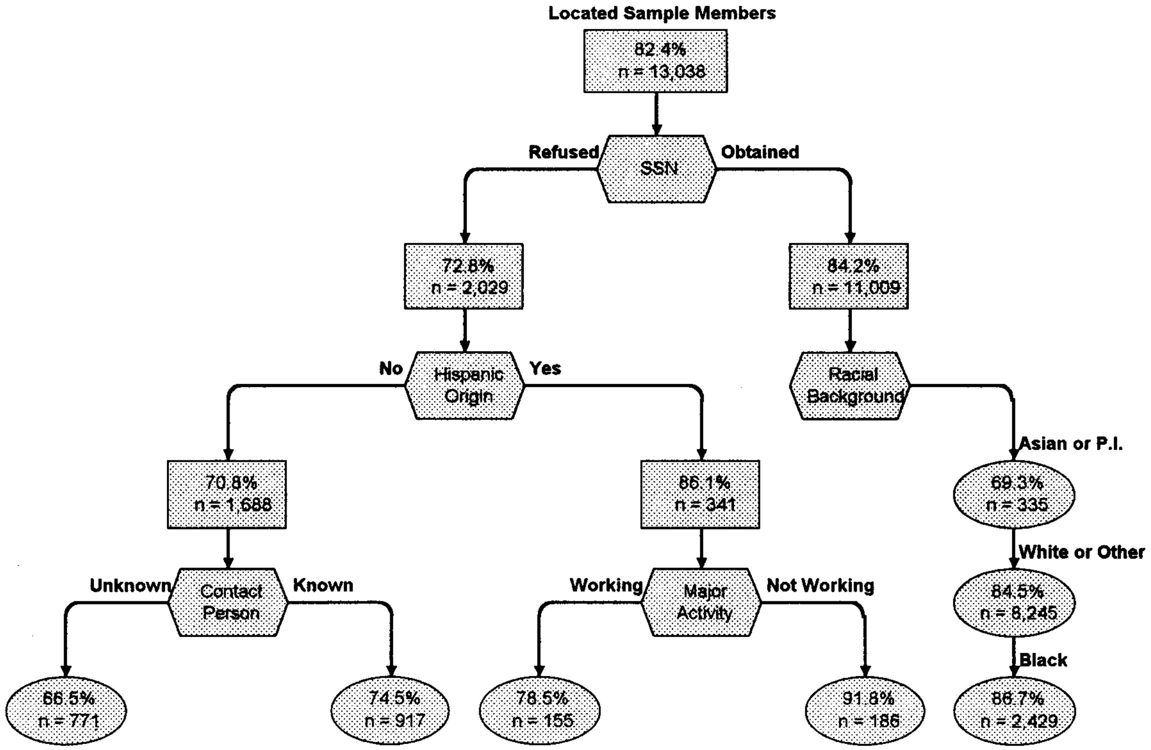

Figure 2. Segmentation of located NSFG sample members by weighted response rate.

Model Adjusted Location and Response Propensities

Predictive margins (Korn and Graubard 1997) were computed to estimate the overall effect of three sources of NHIS contact data (SSN, telephone number, and name of contact person) on NSFG-5 location and response propensities. Each of these variables were significant predictors in both models and each interacted with a number of demographic characteristics.

The predictive margins shown in **Table 1** may be viewed as the expected location and response rates if everyone in the sample provided (or didn't provide) a piece of contact information. For example, the observed location rate for sample members who provided the name of a contact person was 6% higher than those who refused. However, if everyone in the sample had provided the name of a contact person (while retaining their demographics) the location rate would only increase by 2% compared to a sample where no one provided the contact name. A similar result can be seen for refusing to provide an SSN. In contrast, if a telephone number was not obtained or there was no telephone number, the observed location rate was almost 11% lower than when the number was obtained. Even after adjusting for other location predictors however, the 5% difference in predictive margins implies that the presence of a telephone number was an important factor affecting location propensity.

As shown in the segmentation analysis, sample members who refused to provide their SSN during the NHIS were significantly less likely to participate in the NSFG-5 than those who did provide their SSN. Even after adjusting for other predictors in the response model, the absence of an SSN was found to adversely affect response rates by about 26%. The observed response rate among sample members who provided the name of a contact person during the NHIS was about 10% percent more than those who refused. This difference was largely unaffected by other predictors in the response model. In contrast, after adjusting for other response predictors, the predictive margins for the presence or absence of a telephone number indicate virtually no effect on response propensity.

Table 1. Influence of NHIS Contact Data on NSFG-5 Location and Response Propensity

	Location Propensity[1]		Response Propensity[2]	
NHIS Contact Data	**Observed Rate**	**Predictive Margin**	**Observed Rate**	**Predictive Margin**
	% ± 95% CI	% ± 95% CI	% ± 95% CI	% ± 95% CI
Name of Contact Person:				
Obtained for NHIS	96.5 ± 0.5	95.9 ± 0.5	84.2 ± 0.9	83.4 ± 0.9
Refused	89.8 ± 1.4	93.7 ± 0.9	74.0 ± 2.4	75.3 ± 4.0
Social Security Number:				
Given for NHIS	96.3 ± 0.5	95.8 ± 0.5	84.1 ± 0.9	82.5 ± 0.9
Refused	89.8 ± 1.5	93.7 ± 1.1	72.7 ± 2.4	56.1 ± 18.4
Telephone Number:				
Given for NHIS	96.4 ± 0.4	96.9 ± 2.2	82.7 ± 0.9	82.4 ± 0.9
Refused or no phone	85.6 ± 2.2	91.9 ± 1.4	79.2 ± 3.0	82.3 ± 2.8

[1] Among 14,000 NSFG-5 sample members. The observed location rates and predictive margins were computed using the NSFG-5 sampling weights.

[2] Among 13,038 NSFG-5 sample members who were located and eligible for the survey. The observed response rates and predictive margins were computed using the location adjusted weights.

Evaluating the Overall Response Propensity

Generalized Wald statistics, adjusted for design effects, were used to test the goodness-of-fit of the location and response propensity models. Although each model was found to be significant, the predicted overall response propensity ($\hat{\lambda}_i \bullet \hat{\rho}_i$) was not amenable to conventional regression analysis because of the lack of independence between the models. Therefore, a Receiver Operating Characteristics (ROC) curve was used to assess the overall predictive ability of the combined models.

As shown in **Figure 3**, the area under the ROC curve developed for the overall predicted response propensity was 0.65 which corresponds to a highly significant Wilcoxon test statistic (Hanley and McNeil 1982). The curve indicates that in two of every three randomly chosen pairs of sample women, one responding and the other nonresponding, the predicted overall response propensity of the respondent will be greater than that of the nonrespondent. This level of discrimination implies that the NHIS variables used in the two models are informative but not definitive predictors of a sample woman's overall response propensity.

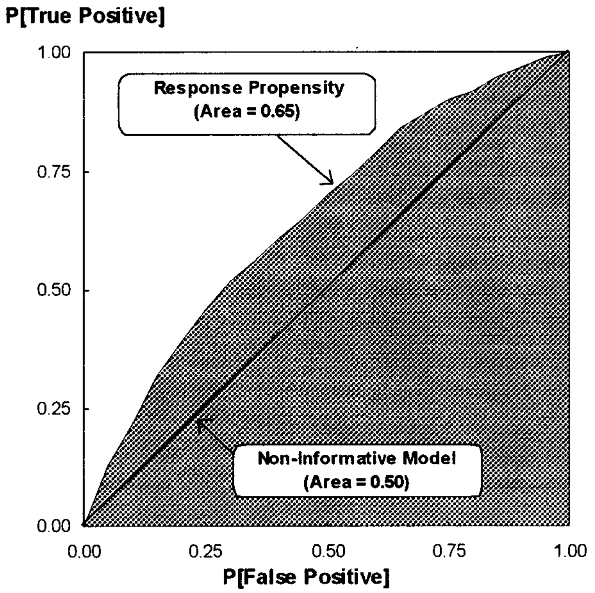

Figure 3. ROC Curve for Overall Response Propensity

References

Folsom RE (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the American Statistical Association, Social Statistics Section,* 197-201.

Groves RM Couper MP (1998). *Nonresponse in Household Interview Surveys.* John Wiley & Sons, New York, NY.

Hanley JA, McNeil BJ (1982). The meaning and use of the area under a receiver-operating characteristic (ROC) curve. *Diagnostic Radiology*, 143:29-36.

Korn EL, Graubard BI, (1997). Predictive margins with survey data. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 651-656.

Magidson J (1993). *SPSS for Windows: CHAID, Release 6.0.* Statistical Innovations, Inc., Belmont, MA.

Massey JT, Moore TF, Parsons VL, Tadros W (1989). *Design and estimation for the National Health Interview Survey, 1985-94.* National Center for Health Statistics. Vital Health Statistics 2(101).

Potter FJ, Iannacchione VG, Mosher, WD, Mason RE, Kavee, JD (1998). *Sample Design, Sampling Weights, Imputation, and Variance Estimation in the 1995 National Survey of Family Growth.* National Center for Health Statistics. Vital Health Statistics 2(124).

Shah BV, Barnwell BG, Bieler GS (1997). *SUDAAN User's Manual, Release 7.5.* Research Triangle Institute, Research Triangle Park, NC.