The Effect of Intensity of Effort to Reach Survey Respondents: A Toronto Smoking Survey

LOUIS T. MARIANO and JOSEPH B. KADANE*

Abstract

The number of calls in a telephone survey is used as an indicator of how difficult an intended respondent is to reach. This permits a probabilistic division of the non-respondents into non-susceptibles (those who will always refuse to respond), and the susceptible non-respondents (those who were not available to respond) in a model of the non-response. Further, it permits stochastic estimation of the views of the latter group and an evaluation of whether the non-response is ignorable for inference about the dependent variable. These ideas are implemented on the data from a survey in Metropolitan Toronto of attitudes toward smoking in the workplace. Using a Bayesian model, the posterior distribution of the model parameters is sampled by Markov Chain Monte Carlo methods. The results reveal that the non-response is not ignorable and those who do not respond are twice as likely to favor unrestricted smoking in the workplace as are those who do.

Keywords: call-backs, number of; Bayesian analysis; Markov Chain Monte Carlo method; informative non-response; ignorable non-response.

*Louis T. Mariano is a Ph.D. candidate, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Joseph B Kadane is Leonard J. Savage University Professor of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

1 Introduction

Given the reality of non-response in every survey, it is of interest to determine how to account for this non-response in the interpretation of the collected data. Rubin (1976) gives necessary and sufficient conditions for such an analysis to be identical from, respectively, a frequentist, likelihood, and Bayesian perspectives, to an analysis based on a model incorporating a missingness mechanism. Building on this, Little and Rubin (1987) led to an extensive literature modeling non-response in an informative, non-ignorable way.

Information about the interaction between the survey and the surveyed can sharpen the analysis of the import of missing data in a survey. The example in this paper concerns the attitudes of Toronto citizens about smoking in the workplace. Random telephone numbers were chosen; at least twelve calls were made to try to reach the intended respondents. Our data for the respondents includes only the number of calls until the survey was completed, not the timing of the unsuccessful calls. With even this attenuated data on how difficult the respondent was to reach, we find our view of the results of the survey to be importantly informed by the number of unsuccessful calls.

The use of information on the number of calls to a subject chosen to participate in a survey is not unique. Pottoff, Manton and Woodbury (1993) present a method for correcting for survey bias due to non-availability by weighting based on the number of call-backs. While our analysis also focuses on the bias due to non-availability, there are major differences. Instead of assuming that refusals do not exist, we allow for and utilize their potential existence in modeling the mechanism which causes non-response. In the analysis that follows, the relationship of non-response to the response variable of interest in the survey is evaluated along with other explanatory variables, after weighting for both household size and the appropriate population demographics. In doing so we address not only whether error exists due to non-availability, but also whether stratification of the respondents by household size and the then current age/sex distribution may eliminate the necessity for accounting for the error by the introduction of a mechanism which describes the non-response. Note that here we match the groupings of Pederson, Bull and Ashley (1996) used in the original published analyses of the dataset; more complex cell adjustment procedures are possible (e.g. Little, 1996; Eltinge and Yansaneh, 1997, and references cited therein).

The remainder of this article is organized as follows: Section 2 gives more detail on the survey; Section 3 introduces the methodology employed; Sections 4 and 5 respectively explore missing-at-random and non-ignorably-missing models; Section 6 discusses the priors distributions chosen for the main analysis, whose results are explained in Section 7. Finally, Section 8 gives our conclusions.

2 The Survey

A bylaw regulating smoking in the workplace in the City of Toronto took effect on March 1, 1988. From January 1988 to the present, a series of six surveys have been conducted to assess attitudes of the public toward smoking, awareness of health risks related to smoking, and the impact of the law on the residents of Metropolitan Toronto. The data being utilized in this analysis comprises the third phase of this series. Northrup (1993) provides the technical documentation for this survey. For clarity, when necessary, the data being analyzed here is referred to as the Phase III data, and information from the first two surveys is referred to as the Phase I & II data.

Northrup (1993) indicates that the data of interest, which were made available by the Institute for Social Research (ISR) at York University, were collected from 1,429 residents of the Metropolitan Toronto area in December 1992 and March 1993. A two-stage probability selection process was utilized to select survey respondents. The first stage employed random digit dialing. The second stage used the most recent birthday method to select one adult individual once an eligible residence was reached. The responses were then weighted by the number of adults in the household. In the analysis that follows, post-stratification weighting was also applied to the census age-sex distribution to adjust for the underrepresentation of some population subgroups. The number of distinct phone lines in the household was not taken into consideration during the data collection.

The number of calls it took to reach each respondent is included as a variable in the dataset, and there are no missing values for this variable. Northrup (1993) explains that the 1,429 responses came from a sample of 5,702 telephone numbers generated by the random digit dialing method. Of these numbers, 2,286 were verified to be eligible households, and 3,150 of the numbers in the sample were not eligible. The status of the remaining 266 numbers was not able to be determined. It has been assumed by ISR that the household eligibility rate of these 266 numbers was equal to the rate for the rest of the sample. This eligibility rate implies an estimated total of 2,398 households in the sample and a response rate of 60%. Thus, an estimated 969 subjects chosen to participate in the survey did not respond. Each subject received a minimum of 12 calls, including day, night, and weekend calls, before being classified as non-respondent.

The dependent variable, for the purpose of this analysis, is an individual's opinion on the regulation of smoking in the workplace, in one of three categories. Category "0" indicates smoking should be permitted in restricted areas only, category "1" indicates smoking should not be permitted at all, and category "2" indicates smoking should not be restricted at all. For each subject chosen to participate in the survey, let $Y_i \in \{0, 1, 2\}$ represent the opinion of subject i.

The data comprises of the answers to 50 survey questions as well as 18 other variables identifying characteristics of the subject. Included in these are:

- "K-risk" is an integer score from 0 to 12 which indicates knowledge of the risks and effects of second-hand smoke.
- "Smoker" indicates the smoking status of the subject:
  "Current smoker" (S), "Former smoker" (SQ), or "Never smoked" (NS).
- "Bother" indicates if second-hand smoke bothers the subject:
  "Always bothers" (b.A), "Usually bothers" (b.USUL), or "Does not bother" (b.NO).
- "Age": (Age in years - 50) / 10.

Pederson, Bull, Ashley and Lefcoe (1989) created a "Knowledge of health effects score" on passive smoking out of the answers to six survey questions, which measured a subject's knowledge of the effects of second-hand smoke. Pederson, et al.'s questions were used in Phase III to create their score, here renamed "K-risk". A higher K-risk score indicates a greater knowledge of the risks of second-hand smoke. The variable "Age" was shifted and rescaled to match how age was treated by Bull (1994) in the Phase I & II analysis.

3 Overview of Methodology

The fundamental question of interest is: "May we ignore the unit non-response and treat the observed data as a random subsample of the population?" Mapping to the terminology of Little and Rubin (1987) and Rubin (1976): If we may treat the observed data for the dependent variable of interest as a random subsample, we call the missing data "missing completely at random" (MCAR). If we may treat the observed data for the dependent variable of interest as a random subsample, after conditioning on the explanatory variables, we call the missing data "missing at random" (MAR). Let θ represent the parameters of the data and let π represent the parameters describing the missing data process. Rubin (1976) calls the parameters π and θ distinct "if there are no *a priori* ties, via parameter space restrictions or prior distributions, between π and θ." If either the MCAR or MAR cases apply and if π and θ are distinct, the mechanism which causes the missing data is said to be "ignorable" for inference about the distribution of the variable of interest. If the missing data for the dependent variable of interest is dependent on the values of that data, then the mechanism which causes the missing data is said to be "non-ignorable" (NI). Groves and Couper (1998) note that when the likelihood of participation is a function of the desired response variable, the non-response bias can be relatively high, even with a good response rate.

Let R_i be an indicator of response. $R_i = I_{\{responded\}}(\text{subject } i)$ and $R = (R_1, \ldots, R_n)^T$. Little and Rubin (1987) suggest that one possible method for accounting for the non-response mechanism is to include this response indicator variable in the model. We may call the mechanism which causes the missing data ignorable if π and θ are distinct and:

$$f(R|Y_{obs}, Y_{mis}, \pi) = f(R|Y_{obs}, \pi) \tag{1}$$

where Y_{obs} and Y_{mis} represent the observed and missing portions of the dependent variable of interest.

The terms "MAR assumption" and "NI assumption" will be used throughout this analysis. For clarity, the term "MAR assumption" is defined as the assumption that the missing data mechanism is ignorable for inference with respect to the dependent variable identified in Section 2. That is, the observed values of that variable are a random subsample of the population, possibly within poststrata, and it is not necessary to account for the missing data mechanism. The term "NI assumption" is defined as the assumption that the missing data mechanism is non-ignorable and the data collected for the dependent variable of interest cannot be treated as a random subsample. Specifically, inference for the population must involve the missing data mechanism.

The approach to assessing the MAR assumption is comprised of three steps. The first step is the examination of what one might do under the MAR assumption. Since the dependent variable of interest has three categories and some of the explanatory variables are quantitative, polytomous logistic regression is employed. Both frequentist and Bayesian forms of the logistic regression model are examined.

In the second step, an NI model is constructed. The non-response mechanism is modeled utilizing the information available about the number of calls made to each subject. Here, the idea of a surviving fraction in the sample is examined to model whether it is actually possible to reach all the intended respondents. Then, the non-response mechanism is related to the dependent variable by including the number of calls in the logistic regression model.

In the development of the NI model, we employ a Bayesian approach to allow for an examination of the values the missing data are likely to take, given the observed data and the model parameters. This is accomplished by utilizing a data augmentation approach, where the missing data are imputed in each iteration of a Markov Chain Monte Carlo (MCMC) simulation. A possible alternative would be to utilize the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) to compute the maximum likelihood estimates (MLE's) of the missing values.

In the third step, an evaluation of the MAR assumption is made. Non-zero coefficients for the number of calls in the logistic regression portion of the NI model will imply that the number of calls does make a difference; i.e. the opinions of those who did not respond in the first 12 calls are likely to differ from those who responded in just a small number of calls. In this case, the missing data mechanism is not independent of the values of the missing data and an MAR assumption would be inappropriate. Next, the log odds of response among the three models are examined. Differences here identify the magnitude of the error that a faulty MAR assumption causes. So, in the evaluation of the MAR assumption, the questions "is there a difference?" and "how large is the difference?" are both addressed.

4 MAR Models

4.1 Logistic Regression

Using the data collected from the (m = 1429) subjects that did respond to the survey, weighted logistic regression was employed to model the public's opinion on smoking in the workplace. The collection of candidate predictors found in the survey questions and the background information was narrowed utilizing a series of Wald tests. Then likelihood ratio tests, AIC, and BIC were used to compare the possible models. The model with the best fit was found to be the one which included additive terms for the variables "K-risk", "Smoker", "Bother", and "Age", as defined in Section 2.

As each of the models examined in this analysis employs a logistic regression component, it is useful here to illustrate the notation being used. Category "0", "smoking allowed in restricted areas only" was chosen to be the reference category. Recall $Y_i \in \{0, 1, 2\}$. For the MAR model, we use only the observed values of the subject's opinion on workplace smoking, $Y_{obs} = (Y_1, \ldots, Y_m)$. Let $Y_{ij} = I_{\{j\}}(Y_i)$ be an indicator of subject i responding in category j, and let W_i represent the weight each subject received. As in the original published analyses of this dataset (Pederson, et al., 1996) both household (see Northrup, 1993) and post-stratification (see Appendix A) weighting were used in the consideration of all models here.

The two categorical explanatory variables, "Smoker" and "Bother", were included in the model by utilizing indicator variables for two of the three categories, with the effect of the third category being absorbed in the intercept term. For "Smoker", "S_i" and "SQ_i" were included as indicators that subject i was either a current smoker or a smoker who had quit. For "Bother", "$b.USUL_i$" and "$b.NO_i$" were included as indicators that second hand smoke usually bothered or did not bother subject i.

Let $X_i =$ represent the vector of explanatory variables for subject i. Then,

$$X_i = (\ K\text{-}risk_i\ ,\ S_i\ ,\ SQ_i\ ,\ b.USUL_i\ ,\ b.NO_i\ ,\ Age_i\ )$$

Here we use an unordered multinomial logit model to consider $p_j(x_i) = P(Y_{ij} = 1|X_i = x_i)$, the probability that subject i responds in category $j \in \{0, 1, 2\}$, given the observed explanatory variables for subject i. This model, of course, utilizes linear equations η_{ij} describing the log odds of subject i responding in category j versus the reference category $j = 0$. So, for $j = 1, 2$ we wish to examine:

$$ln \frac{p_j(x_i)}{p_0(x_i)} = \eta_{ij} = \beta_{0j} + X_i\beta_j \tag{2}$$

with $\eta_{i0} = 0$. The two resultant linear equations, η_{i1} and η_{i2}, each have seven coefficients, including an intercept term β_{0j} and those displayed below:

$$\beta_j = (\,\beta_{K\text{-}risk_j}\ ,\ \beta_{S_j}\ ,\ \beta_{SQ_j}\ ,\ \beta_{b.USUL_j}\ ,\ \beta_{b.NO_j}\ ,\ \beta_{Age_j}\,)$$

The MAR logistic regression model has 14 parameters. The vector of these 14 parameters, represented by $\beta = (\beta_{01}, \beta_1, \beta_{02}, \beta_2)$ has the likelihood (or, more appropriately, pseudo-likelihood, since the weights are incorporated through the variable W_i):

$$L(\beta) \propto \prod_{i=1}^{m} \prod_{j=0}^{2} \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} \tag{3}$$

4.2 Bayesian Logistic Regression

The likelihood in equation (3) and the data collected from the survey respondents are utilized in the Bayesian analysis. The same four explanatory variables selected in the frequentist analysis above are used as the explanatory variables here. Prior distributions, discussed in Section 6, were assigned to the logistic regression parameters. An MCMC simulation is utilized in order to draw from the posterior distribution of the parameters.

5 NI Model

5.1 Modeling the Non-Response Mechanism

Since the missing values are not necessarily missing at random, the mechanism which caused them to be missing must be addressed. Northrup (1993) indicates that non-respondent subjects chosen to participate in the survey were called a minimum of 12 times, including a minimum of three day, four evening and four weekend calls. Unfortunately, other useful information regarding the number of calls was not retained. We do not know which of the non-respondents were called more than twelve times or whether an individual call was placed during the day, evening, or weekend. We also are unaware of the details of the non-response, such as whether the subject was contacted but refused to participate, whether the calls were ever answered by a machine, or whether they were answered at all. Thus, stratification of the non-respondents was not possible, and they were all treated as exchangeable in this analysis.

Each subject was called a number of times until the survey was successfully completed or they were classified as non-respondent. For the respondents, the number of calls variable (C_i) describes the number of trials until the first success for subject i. Thus, one might expect the number of calls to follow a Geometric distribution with truncated observations for the non-respondents. Specifically, let $\pi = P(\text{a call is successful})$; then, consider $C_i \sim Geometric(\pi)$ and $P(C_i = c_i) = \pi(1-\pi)^{c_i - 1}$. Note that if auxiliary information about the number of calls to the non-respondents were available (e.g. Grover and Couper, 1998), we could have also considered conditional response probabilities here.

The histograms in Figure 1, compare the data (through the first twelve calls) to a Geometric distribution with parameter $\pi = .225$, which appears to match fairly well. The sample order statistics suggest $\pi \in (.2, .25)$. The histogram of the actual survey data reveals that the number of subjects reached on the first call are fewer than the number reached on the second call. It is possible that more of the second calls were placed at a time which had a higher success rate.

[Figure 1 about here.]

Suppose $\pi = .225$; by the memoryless property of the Geometric distribution, we would expect 218 of the 969 non-respondents to reply on the 13th call. This would make the data through the first 13 calls appear as in Figure 2. Clearly, Figure 2 does not display the behavior of a Geometric random variable. Consider the following question: "If all subjects were called an unlimited amount of times, would they all have been reached?" Answering "yes" to that question for this dataset results in the problem illustrated in Figure 2.

[Figure 2 about here.]

Given the information outlined above, the assertion that "not all subjects chosen for the survey are reachable" is a viable one. Maller and Zhou (1996) discuss immune subjects – individuals who are not subject to the event of interest. Following their terminology, if it is not possible to procure a response from a subject chosen for the survey given an unlimited amount of calls, that subject is categorized as immune. Subjects who are not immune are categorized as "susceptible". The set of immune (i.e. non-susceptible) subjects comprise the "surviving fraction" of the sample. Mapping to more familiar terminology, the immune subjects include those who were reached and refused, those who would have refused if they had been reached, and those cases of a physical or mental inability to ever participate. Northrup (1993) indicates that those who initially refused to participate were subsequently contacted by the most senior interviewers, so, we make the assumption here that all remaining refusals would not ever participate. The susceptible group includes the respondents, those who would have responded if successfully contacted, and those who were physically or mentally unable to participate during the data collection period but were willing and able at some other time.

Let the variable $Z_i = I_{\{susceptible\}}(\text{subject } i)$ be an indicator of the the susceptibility of subject i, and $\rho = P(\text{subject } i \text{ is susceptible})$, i.e. $Z_i \sim Bernoulli(\rho)$. Now suppose that the number of calls to the susceptible subjects follows a Geometric distribution, i.e. $C_i|Z_i = 1 \sim Geometric(\pi)$. Does this eliminate the problem illustrated in Figure 2?

Let R_i be an indicator of response of subject i. The non-response mechanism can be accounted for by including these response indicators in the model. However, the introduction of the susceptibility variable implies two distinct classes of non-response. So, it is possible to be more detailed and use both the susceptibility $Z = (Z_1, \ldots, Z_n)^T$ and the response R indicators in a mixture model describing the non-response. Updating Equation (1), the missing data mechanism is ignorable if and only if (π, ρ) is distinct from θ and

$$f(R, Z|Y_{obs}, Y_{mis}, \pi, \rho) = f(R, Z|Y_{obs}, \pi, \rho) \tag{4}$$

Let $C_{obs} = (C_1, \ldots, C_m)$ and $Z_{obs} = (Z_1, \ldots, Z_m)$ be the vectors of the number of calls and the observed susceptibility for each respondent. Also, let $R = (R_1, \ldots, R_n) =$ be the vector of response for each intended respondent. Every subject, i, may be classified by response into three mutually exclusive groups, A_{obs} – observed, A_{mis} – missing, and A_{imm} – immune, where:

$$\begin{aligned}
A_{obs} &= \{\ i\text{: } i \text{ was Susceptible and Responded }\} \\
A_{mis} &= \{\ i\text{: } i \text{ was Susceptible but did not Respond in 12 calls }\} \\
A_{imm} &= \{\ i\text{: } i \text{ was Not Susceptible }\}
\end{aligned}$$

The probability that a subject is in each of these categories may be calculated as follows:

$$\begin{aligned}
P(i \in A_{obs}) &= P(Z_i = 1, R_i = 1, C_i = c_i) &&= \rho\pi(1-\pi)^{c_i - 1} \\
P(i \in A_{mis}) &= P(Z_i = 1, R_i = 0, C_i > 12) &&= \rho(1-\pi)^{12} \\
P(i \in A_{imm}) &= P(Z_i = 0) &&= 1 - \rho
\end{aligned}$$

The data indicates $m = 1429$ subjects in A_{obs} and $n - m = 969$ non-responsive subjects in $A_{mis} \cup A_{imm}$; $n = 2398$ is the estimated total number of subjects chosen to participate in the survey. Thus, the joint density of Z_{obs}, R and C_{obs} given ρ and π is:

$$f(Z_{obs}, R, C_{obs} | \rho, \pi) \propto \left[\rho^m \pi^m \,(1-\pi)^{(\sum_{i=1}^{m} c_i) - m}\right] \times \left[(1-\rho) + \rho(1-\pi)^{12}\right]^{n-m} \qquad (5)$$

The mixture model described by Equation 5 may be viewed as a special case of the non-response models discussed in Drew and Fuller (1981).

It would be useful to confirm that the above joint distribution accurately represents the response pattern of the susceptibles in the dataset. The MLE estimate for ρ is simply the proportion of respondents in the sample, which clearly underestimates ρ. Setting $U(0,1)$ prior distributions for both ρ and π and examining their joint posterior distribution by MCMC simulation, the posterior medians are found to be $\rho = .636$ and $\pi = .205$, with equal-tailed posterior credible intervals of (.613 , .659) and (.191 , .219) for ρ and π respectively. Figure 3 illustrates how the dataset might look after imputing the missing number of calls for our susceptible non-respondents based on these posterior medians. The problem previously displayed in Figure 2 has now been mostly eliminated.

[Figure 3 about here.]

While the Geometric distribution appears sufficient (after accounting for susceptibility), a referee questions the use of the Geometric distribution as it does not make use of possibly useful covariates. As explained above, the covariates we think would be most useful for this purpose were not collected. One alternative for modeling the response mechanism of the susceptibles is to use a discretized Gamma distribution. In cases where more complexity is necessary, the ν–Poisson (a two parameter Poisson which generalizes some well known discrete distributions, including the Geometric) of Shmueli, Minka, Kadane, Borle and Boatwright (2001) may also be considered.

5.2 Relating Non-Response to the Dependent Variable – The NI Model

Since the non-response of the susceptibles is described by the conditional Geometric distribution of the number of calls, the effect of the non-response of the susceptibles on the dependent variable may be considered by including the number of calls as an additional explanatory variable in the logistic regression likelihood. This will create two additional parameters in the logistic regression portion of the model, which are the coefficients of the number of calls, β_{call_j} in each of the linear equations η_{ij} described in equation (2).

Non-zero coefficients for the number of calls, then, would indicate that the dependent variable is not independent of the non-response mechanism, and, hence the non-response mechanism is non-ignorable. If these coefficients are zero, the non-response of the susceptibles is ignorable. Conclusions made here rely upon the underlying modeling assumption that the relationship among the number of calls, the dependent variable and the other explanatory variables considered is the same for the respondents and susceptible non-respondents. Including the number of calls in the logistic regression portion of the model does not address the immune subjects, since there will never be the realization of a successful call to them.

The full pseudo-likelihood for the NI model (or, more precisely, the susceptible NI model) is the product of the non-response and logistic regression pieces:

$$\begin{aligned} L(\rho, \pi, \beta) &\propto \left[\rho^m \pi^m (1-\pi)^{(\sum_{i=1}^m c_i)-m}\right] \times \left[(1-\rho) + \rho(1-\pi)^{12}\right]^{n-m} \\ &\times \left[\prod_{i=1}^m \prod_{j=0}^2 \left(\frac{e^{\eta_{ij}}}{1+e^{\eta_{i1}}+e^{\eta_{i2}}}\right)^{y_{ij} w_i}\right] \end{aligned} \tag{6}$$

Note that the household and post-stratification weighting variable W_i is included here in an effort to account for whether proper stratification of the respondents may eliminate the necessity for the introduction of a mechanism to describe non-response.

5.3 Data Augmentation

Tanner and Wong (1987) suggest an iterative method for computation of posterior distributions when faced with missing data. This method applies whenever augmenting the dataset makes it easier to analyze and the augmented items are easily generated. Consider the following additional notation: Let S represent the total number of susceptible subjects in the sample.

$S = \sum_{i=1}^{n} Z_i$, $S \sim Binomial(\rho)$. Let X be the matrix of explanatory variables (including the number of calls) for all the subjects selected to participate in the survey. Let $Y = (Y_1, \ldots, Y_n)$ be the vector of their responses. Partition X into $\{X_{obs}, X_{mis}, X_{imm}\}$ and Y into $\{Y_{obs}, Y_{mis}, Y_{imm}\}$. Also, by the memoryless property of the Geometric distribution, the distribution of the additional number of calls required to reach the subjects in A_{mis} is known, and may be expressed: $\forall i \in A_{mis}$, let $V_i = C_i - 12$, which is also distributed as a Geometric random variable with parameter π.

Now suppose that the true values of S, X_{mis}, & Y_{mis} were known. The likelihood could then be considered in the form:

$$
\begin{aligned}
&L(\rho, \pi, \beta | X_{obs}, X_{mis}, Y_{obs}, Y_{mis}, S, R) \\
\propto \quad & \left[(\rho\pi)^s (1-\pi)^{(\sum c_{sus})-s}\right] \times \left[(1-\rho)^{n-s}\right] \times \left[\prod_{i=1}^{s} \prod_{j=0}^{2} \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}}\right)^{y_{ij} w_i}\right],
\end{aligned}
\tag{7}
$$

where $\sum C_{sus} = \sum C_{obs} + \sum (V_i + 12)$ is the number of calls that would have been necessary to reach all susceptibles and the summands are taken over the appropriate range of subjects.

Although the true values of S, X_{mis}, & Y_{mis} are unknown, one may utilize what is known about the behavior of these variables to impute stochastically possible values for them within the MCMC algorithm. Given ρ, a value for S may be drawn from a truncated $Binomial(2938, \rho)$, where $1429 \leq S \leq 2398$. Given S, the number of subjects in A_{mis} is known. For each of these subjects in A_{mis} a value $V_i \sim Geometric(\pi)$ may be drawn, which results in an imputation for the number of calls needed to reach each susceptible but unreached subject. The relationships among the number of calls and the other explanatory variables may then be exploited to impute values for the rest of X_{mis}. Specifically, the missing values of Age and K-risk are imputed by regressing Calls on Age and K-risk respectively and predicting from the resultant linear equations. Similarly, the missing values of Smoker and Bother are imputed via logistic regression on each, using Calls as the explanatory variable. Here the model assumptions are checked using the respondents data, and an assumption is being made that these same relationships hold for the susceptible non-respondents. Note that these regression and logistic regression equations are fit in the Bayesian context (e.g. Gelman, Carlin, Stern and Rubin, 1998) and necessitate the inclusion of additional parameters, β_I, in the MCMC process which describe these relationships (see Appendix B for more detail). We chose this imputation plan in the interest of the efficiency of the full MCMC algorithm. An alternative would be impute the missing values for a particular explanatory variable conditional on all the remaining variables (e.g. Rubin, 1996). Finally, Y_{mis} may be predicted by utilizing the imputed values of X_{mis} and the relationship described in the logistic regression model. In the interest of the exchangeability of the susceptible non-respondents in the absence of subsequent stratification information, we apply a weight of 1.0 to all the imputed Y_{mis} values; an alternative here would be to impute the sex and household size of the susceptible non-respondents, in addition to their age, and apply the weighting procedure described in Appendix A to the imputed Y_{mis}.

5.4 Sampling from the Posterior Distribution

The full MCMC simulation consists of a Metropolis algorithm supplemented in every iteration with the data augmentation described above. An outline of the MCMC algorithm used may be found in Appendix B. Convergence was assessed utilizing the method of Hiedelberger and Welch (1983) as described in Cowles and Carlin (1996). MacEachern and Berliner (1994) assert that, under loose conditions, subsampling the MCMC simulated values to account for autocorrelation will result in poorer estimators. Following their suggestion, all simulated values, after an appropriate burn-in period, were used in the analysis that follows.

6 Choice of Prior Distributions

In the evaluation of possible prior distributions for the parameters of both the NI and MAR models, the goal of the comparison of the various models was taken into consideration. The choice of prior distributions for the parameters was made from the perspective of the MAR belief. Two possibilities were examined.

The first option is built around the utilization of the Phase I & II surveys. Since these surveys were similar to and were completed prior to the Phase III survey which comprises our data, information contained in these first two surveys may be utilized in the construction of priors. The same dependent variable was contained in the Phase I & II dataset, along with the variables Smoker, Age, and K-risk. A logistic regression model was compiled from the Phase I & II data to describe the relationship between the opinion on workplace smoking and these three explanatory variables. Normal priors were constructed for the coefficients of these three variables centered at their MLE's, but with increased standard error. The error terms were increased due to three factors: i) There was a three year span between the Phase II and Phase III surveys; opinions may have changed over that time, possibly as a result of the impact of the bylaw. ii) The MLE's were calculated under the same MAR assumption being evaluated. iii) Prior to the collection of the Phase III data, there existed the possibility that other explanatory variables would be included in the model; in the presence of other variables, the effect of these three could be altered. Although the variances were increased, the means were not changed, since it was unknown, *a priori*, in what direction any change might occur. Since the available Phase I & II data contained no information about the Calls or Bother variables, the coefficients of these were assigned a diffuse Normal(0,9) prior. For clarity, this option will be referred to as the "Phase I & II prior" in this analysis.

In the second option Normal(0,9) priors are assigned to each of the logistic regression coefficients. One motivation for this choice is that, for the same three reasons the error terms were increased above, the variables common to the Phase I & II and Phase III surveys are not exchangeable. Thus, construction based on the Phase I & II results would be inappropriate. This option will be referred to as the "Central prior".

The choice to use Normal(0,9) distributions here is for convenience. Centering the prior at zero gives equal weight to either direction of the relationship. We believe the choice of a variance of nine to be adequate without being overly diffuse. The use of improper priors could lead to a Markov Chain Monte Carlo simulation that never converges, and, as Natarajan and Kass (2000) show, an overly diffuse proper prior may behave like an improper one. In Section (7.2), we offer a sensitivity analysis to evaluate how the results are effected by the choice of prior.

The non-response parameters of the NI model, ρ and π , were treated the same under both prior options. There was no additional information available about the probability of a successful call or the probability of susceptibility. Thus, ρ and π were each assigned a $U(0,1)$ prior.

The data augmentation parameters found in each of the logistic regression equations, β_l, were were independently given diffuse Normal(0,9) priors. For each linear regression equation found in the data augmentation process, the coefficients, β_r, and variance, σ_r^2, were set to $p(\beta_r, \sigma_r^2) \propto 1/\sigma_r^2$, the standard non-informative prior distribution (e.g. Gelman, et al., 1998). Note that the closed forms of the posterior distributions of the linear regression parameters are known and may be drawn from directly.

7 Results

First, the validity of the MAR assumption is examined through the coefficients of the number of calls variable. Then, the NI model is evaluated with respect to sensitivity to the choice of prior. Finally, the magnitude of the impact of a faulty MAR assumption for this dataset is investigated by illustrating the change in the odds of response.

7.1 Coefficients for the Number of Calls

[Figure 4 about here.]

For both the Phase I & II and Central priors, Figure 4 displays the posterior density (solid line) and 95% credible interval estimates (dotted lines) of the coefficient of the calls variable in η_{i1} in the NI model, and compares them to the point $\beta_{call_1} = 0$ (dashed lines). The results clearly indicate this coefficient differs from zero. We also find a non-zero result in η_{i2}, where, using the Phase I & II prior, the 95% HPD credible interval for β_{call_2} is (-0.03613, 0.11595).

The non-zero coefficient of C_i demonstrates a dependence between the number of calls and the subject's opinion on smoking in the workplace. Thus, the dependent variable and the non-response mechanism are not independent under the conditions discussed in Section 5.2. This result implies that an assumption that the missing observations are missing at random prior to accounting for the non-response mechanism is incorrect for this dataset.

There is a hint in Figure 3 that the probability of a successful call decreases as the call number increases. To verify the assumption that the relationship between the number of calls and the log odds of response is linear, a second Bayesian NI model was constructed. This model split the calls variable into two, $C_i I_{\{C_i<7\}}$ and $C_i I_{\{C_i\geq 7\}}$, based on whether the number of calls were fewer than seven. The posterior distributions of the coefficients of these two variables were then compared and evidence that they are essentially different was not found. In particular, for η_{i1} the 95% credible interval for $C_i I_{\{C_i\geq 7\}}$ contained the same interval for $C_i I_{\{C_i<7\}}$, and for η_{i2} the 95% credible intervals strongly overlapped.

7.2 Sensitivity to Priors

Would different prior distributions, either on the calls coefficient or on the others, make a difference in the effect illustrated above? Table 1 displays 95% HPD credible intervals for the coefficient of the calls variable in the first logit equation of the NI model for six different priors. The priors include the Phase I & II and Central priors as well as four others—labeled options 3, 4, 5, and 6. Options 3 and 4 resemble the Central prior except that they change the prior distribution on the coefficient of the number of calls to Normal(1,9) and Normal(-1,9) respectively. Option 5 places Normal(0,9) priors on β_{call_1}, β_{age_1}, and $\beta_{b.USUL_1}$, a Normal(1,9) prior on β_{01}, a Normal(.5,9) prior on β_{K-risk_1}, a Normal(-1,9) prior on β_{S_1} and Normal(-.5,9) priors on β_{SQ_1} and $\beta_{b.NO_1}$. Option 6 takes the Central Prior and reduces all the variances from nine to two.

[Table 1 about here.]

Under all six priors, Table 1 demonstrates that the coefficient of the calls variable in the first logit equation clearly differs from zero. The finding that the missing data mechanism is non-ignorable for this dataset does not appear to be effected by the choice of prior among these options.

7.3 Effect on Odds of Response

Given the failure of the MAR assumption shown above, it is of interest to question the relevance of the error that using the MAR assumption would create. The magnitude of the error induced by a faulty MAR assumption may be illustrated by examination of its effect on the odds ratio $p_1(x_i)/p_0(x_i)$. First, we consider the effect on a typical respondent profile. The modal respondent was a non-smoker between the ages of 25-35 years old who was usually bothered by second-hand smoke, had a K-risk of 11 and could be reached in 2 calls. We label this modal respondent as Subject 1. Table 2 demonstrates the change in posterior odds for Subject 1 when called 13 times.

[Table 2 about here.]

The Subject 1 column Table 2 indicates a dramatic difference in the posterior odds when the non-response mechanism is taken into consideration. For this typical respondent profile, when the number of calls is increased from two to thirteen the posterior odds of choosing "Smoking should not be permitted at all" over "Smoking should be permitted in restricted areas only" increases by 52.18% under the Phase I & II prior and 57.84% when using the Central prior. This is dramatic evidence of the relationship between the dependent variable and the non-response mechanism.

Are the results for the modal subject above typical? Table 2 also displays the effects on the odds of response under the NI model for three additional test subject profiles for each of the six different priors considered above. Subject 2 is a fifty year old non-smoker who is always bothered by smoke and has a perfect "K-risk" score. Subject 3 is a 27 year old former smoker who is not bothered by smoke and has a "K-risk" score of seven. Subject 4 is a 40 year old smoker who is not bothered by smoke and has a "K-risk" score of three. On multiple subjects with multiple priors, Table 2 consistently shows the same result. Increasing the number of calls to greater than 12 will increase the posterior odds of choosing category "1" over category "0". For each of the test subjects and priors found in Table 2, the increase was between 52.18% and 58.41%.

Similar results were found when examining the odds of choosing the "Smoking should not be restricted at all" category over the "Smoking should be permitted in restricted areas only" category. Using test subjects which were a current and a former smoker (Subjects 3 and 4 above), the posterior odds increased 46.7% when the number of calls was increased from 2 to 13 under the Phase I & II prior.

7.4 Effect on Probability of Response

With the shift in posterior odds illustrated above comes a corresponding shift in the estimated probabilities that a subject will respond in a particular category. Among the respondents, 57.45% chose category "0", 40.64% chose category "1", and 1.91% chose category "2". The number of non-respondent susceptibles have a posterior median of 469, with a 95% credible interval of (25, 944). On average, 55.88% of the simulated non-respondent susceptibles chose category "0", 40.03% chose category "1", and 4.08% chose category "2". While, for categories "0" and "1", the average values for the non-respondent susceptibles do fall within the 95% confidence intervals for the proportions of the respondents in these categories, the point estimates for each category shift when the non-response mechanism is included in the model. In comparing the category “2” results, we estimate that non-respondents are twice as likely to favor no restrictions on smoking (category “2”) than are respondents. While the low number of subjects found in category “2” are unlike to provoke a change in workplace smoking law, the increase noted in the non-respondents in this category serves as an example of how the lack of proper consideration of the non-respondents could lead to flawed conclusions about the data.

8 Conclusion

Section 7 demonstrates that, for the dependent variable of interest in this dataset, an assertion that the missing observations are missing at random, prior to accounting for the missing data mechanism, is incorrect, assuming the relationship among the relevant variables is the same for all susceptible subjects. Furthermore, the use of a faulty MAR assumption in the evaluation of this dependent variable risks serious error in the calculation of the posterior odds and in any conclusion drawn from them. In order to perform a proper evaluation of the opinion on smoking in the workplace in Toronto in early 1993 via the dependent variable of interest in this survey, it is necessary to account for the non-response mechanism in the model structure.

In this analysis, only one simple piece of information, the number of calls, was utilized. A more complete treatment could have been made, had more information been available. Knowledge of the exact number of calls to the non-respondents, instead of a minimum, and the time of day of the calls could have enabled this analysis to be more precise. In addition, knowledge of the type of non-response, refusal or non-availability, and the number of of times the non-respondents were actually contacted could have allowed for better classification of the non-respondents. Groves and Couper (1998) point out that statistical errors arising from non-availability and those arising from refusals are likely to differ. As they further comment, the evaluation of how efforts to seek cooperation effect measurement error is an important area of research.

The results illustrated above apply only to this one dependent variable assessing smoking in the workplace in this one dataset. Given the perception that smoking has become less socially acceptable over recent years, it would be reasonable to think that non-response error due to questions about smoking may be more severe than other topics. A comparison of non-response bias including various smoking related questions, and others which do not concern smoking may be found in Biemer (2001); this comparison lends no credence to the idea that non-response error is unique to questions relating to smoking.

Although the above results make no implications about the missing data mechanisms in other surveys, there is a clear demonstration here that blindly assuming that the respondents of a survey constitute a random subsample of the population for the variables of interest can be an unwise choice. Information, available at the time of data collection, can enable the evaluation of whether or not the mechanism which causes the non-response is ignorable. In light of this observation, then, it should be of interest to those who work with such data to make use of the available information pertaining to the non-response in the evaluation of that data and to make such information available to others who utilize the dataset. As a general matter, we believe that the collection and analysis of data on where and how respondents were found, as well as how difficult they were to find, is an important future direction for survey methodology and practice.

Acknowledgments

This research was funded by National Science Foundation Grant DMS-9801401. The authors thank Shelley Bull for her many helpful comments and suggestions and for assistance in the acquisition of the data and John Eltinge and the anonymous referees and Associate Editor for their valuable comments.

Data from the Attitudes Toward Smoking Legislation, which was funded by Health and Welfare Canada, were made available by the Institute for Social Research at York University. The data were collected by the Institute for Social Research for Dr. Linda Pederson of the University of Western Ontario, Dr. Shelley Bull of the University of Toronto and Dr. Mary Jane Ashley of the University of Toronto. The principal investigators, the Ontario Ministry of Health and the Institute for Social Research bear no responsibility for the analyses and interpretations presented here.

A Post-stratification Weighting

HHW_i is the household weight of subject i as described in Northrup (1993).

- Let m = the number of respondents.
- Let r = the cumulative number of adults in the responding households.
- Let h_i = the number of adults in subject i's household.
- $HHW_i = h_i \cdot m/r$.

Proportions in the sample falling into the following age groups were calculated for both male and female respondents: 18-24 years, 25-44 years, 45-64 years, and over 65 years old. These proportions were then compared to the age/sex distribution in Metropolitan Toronto.

- Let p_{1i} = the proportion of adult Metropolitan Toronto residents falling into the same age/sex category as subject i, as per the 1991 Census.

- Let p_{2i} = the proportion of survey respondents with the same age and sex categories as subject i.

- $W_i = HHW_i \cdot p_{1i}/p_{2i}$, where W_i is the final post-stratification weight used in the analysis.

B MCMC Implementation

The full MCMC simulation for the NI model consists of a Metropolis algorithm supplemented with the data augmentation described in Section 5.3. The following is an overview of the MCMC algorithm. Variables used below are defined in Section 5. At each iteration t,

1. Draw ρ_t from $Beta(s_{t-1} + 1, 2398 - s_{t-1} + 1)$.

2. Impute s_t from $Binomial(\rho_t) \geq 1429$.

3. Impute C_{mis_t}: draw $(s_t - 1429)$ v_i's from $Geometric(\pi_{t-1})$ and $\forall c_i \in c_{mis}, c_i = v_i + 12$.

4. Draw π_t from $Beta(s_t + 1, \sum c_{sus_t} - s_t + 1)$.

5. Impute values for the rest of X_{mis} by utilizing the relationships with the number of calls, as described in section 5.3.

6. Update the additional parameters used in the data augmentation of X_{mis}.

    - Update linear regression parameters, β_r and σ_r, by drawing directly from the closed form of their posteriors.

    - Update logistic regression parameters, β_l, using a Metropolis step on each.

7. Impute Y_{mis_t}: $\forall y_i \in y_{mis}$, draw y_i from a $Multinomial(p_0(x_i), p_1(x_i), p_2(x_i))$.

8. Update each β_{kj} using a Metropolis step on the conditional likelihood and a Normal jump function.

References

Biemer, P. P. (2001). *Survey Nonresponse*, chap. Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. Wiley and Sons.

Bull, S. (1994). *Case Studies in Biometry*, chap. 16, Analysis of Attitudes toward Workplace Smoking Restrictions, 249–270. New York: Wiley and Sons.

Cowles, M. K. and Carlin, B. P. (1996). "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association*, 91, 883–904.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society*, B39, 1–38.

Drew, J. H. and Fuller, W. A. (1981). "Nonresponse in Complex Multiphase Surveys." In *ASA Proceedings of the Section on Survey Research Methods*, 623–628. American Statistical Association (Alexandria, VA).

Eltinge, J. L. and Yansaneh, I. S. (1997). "Diagnosis for Formation of Nonresponse Adjustment Cells, With and Application to Income Nonresponse n the U.S. Consumer Expenditure Survey." *Survey Methodology*, 23, 33–40.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1998). *Bayesian Data Analysis*, chap. 14, Generalized Linear Models. London: Chapman & Hall.

Groves, R. M. and Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley and Sons.

Hiedelberger, P. and Welch, P. (1983). "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research*, 31, 1109–1144.

Little, R. J. A. (1986). "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review*, 54, 139–157.

Little, R. J. A. and Rubin, D. B. (1987). *Statictical Analysis with Missing Data*. New York: Wiley and Sons.

MacEachern, S. N. and Berliner, L. M. (1994). "Subsampling the Gibbs Sampler." *The American Statistician*, 48, 188–189.

Maller, R. and Zhou, X. (1996). *Survival Analysis with Long Term Survivors*. Chichester, UK: Wiley and Sons.

Natarajan, R. and Kass, R. E. (2000). "Reference Bayesian Methods for Generalized Linear Mixed Models." *Journal of the American Statistical Association*, 95, 227–237.

Northrup, D. A. (1993). "Attitudes Towards Workplace Smoking Legislation: A Survey of Residents of Metropolitan Toronto, Phase III, 1992/93 Technical Documentation." Tech. rep., Institute for Social Research, York University. Unpublished.

Pederson, L. L., Bull, S. B., and Ashley, M. J. (1996). "Smoking in the workplace: Do smoking patterns and attitudes reflect the legislative environment?" *Tobacco Control*, 5, 39–45.

Pederson, L. L., Bull, S. B., Ashley, M. J., and Lefcoe, N. M. (1989). "A Population Survey on Legislative Measures to Restrict Smoking in Ontario: 3. Variables Related to Attitudes of Smokers and Nonsmokers." *American Journal of Preventive Medicine*, 5, 313–322.

Potthoff, R. F., Manton, K. G., and Woodbury, M. A. (1993). "Correcting for Nonavilability Bias in Surveys by Weighting Based on Number of Callbacks." *Journal of the American Statistical Association*, 88, 1197–1207.

Rubin, D. B. (1976). "Inference and Missing Data." *Biometrika*, 63, 581–592.

— (1996). "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, 91, 473–489.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2001). "Using Computational and Mathematical Methods to Explore a New Distribution: The ν-Poisson." Tech. Rep. 740, Department of Statistics, Carnegie Mellon University.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, 82, 528–549.

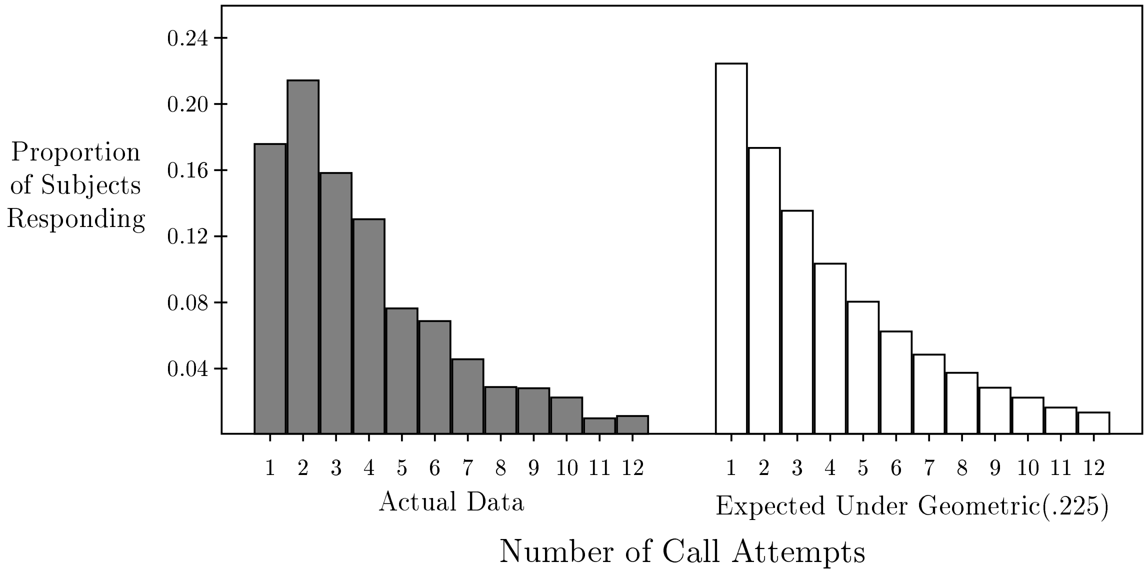

Figure 1: Comparison of the actual survey data for successful calls in the first 12 attempts to expected results based on a Geometric(.225) distribution for the number of calls needed to complete the survey.

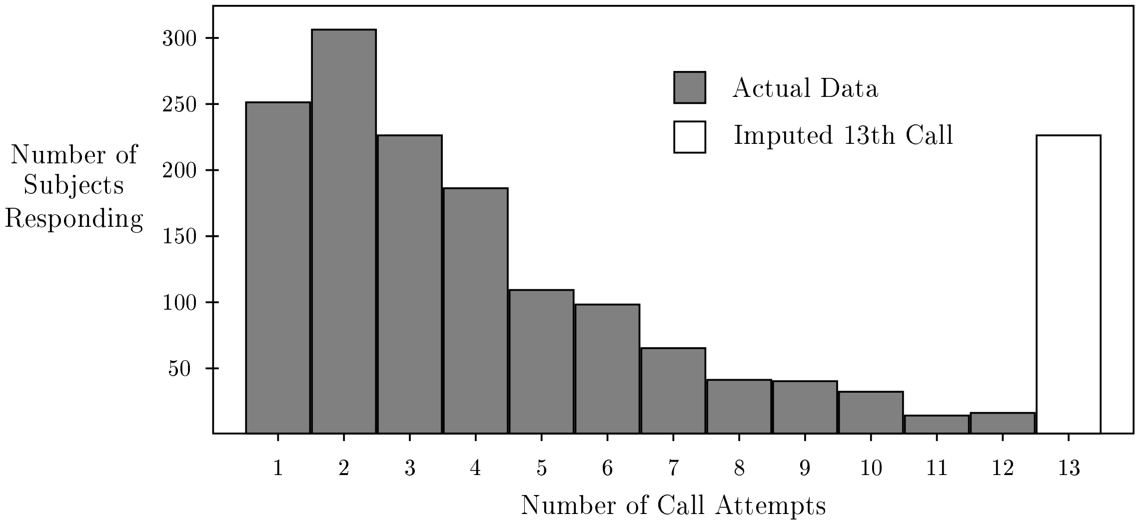

Figure 2: Display of the actual number of successful calls on each attempt through the first 12 and the expected number of successful calls on the 13th attempt. The expectation for the 13th call is based on a Geometric(.225) distribution to model the number of calls until the survey is completed.

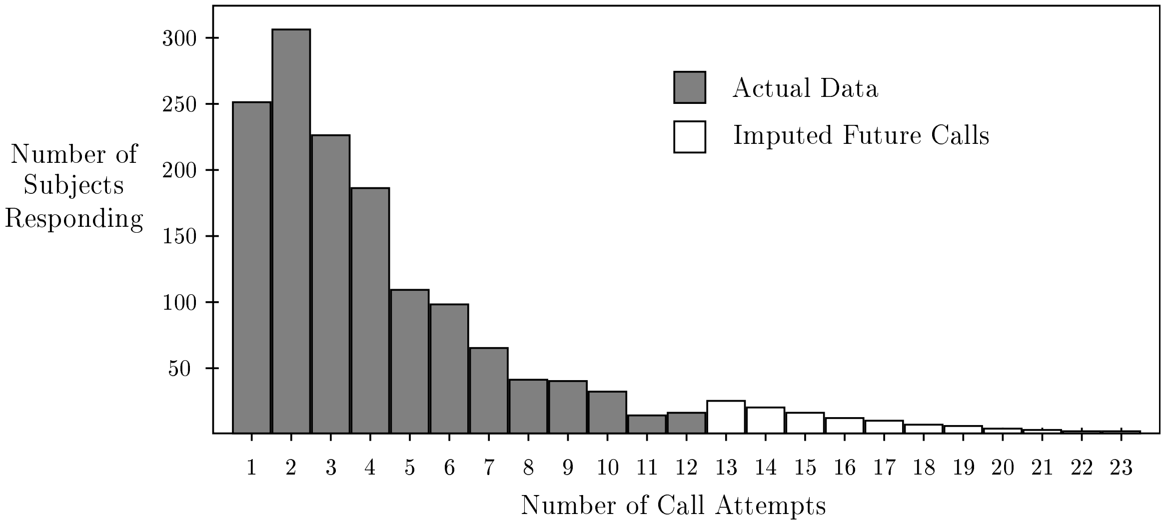

Figure 3: Display of the actual number of successful calls on each attempt through the first 12 and the imputed number of successful calls for call attempts 13 and higher. Imputed values are based on a probability of a successful call of .205 and a probability of susceptibility of .636.

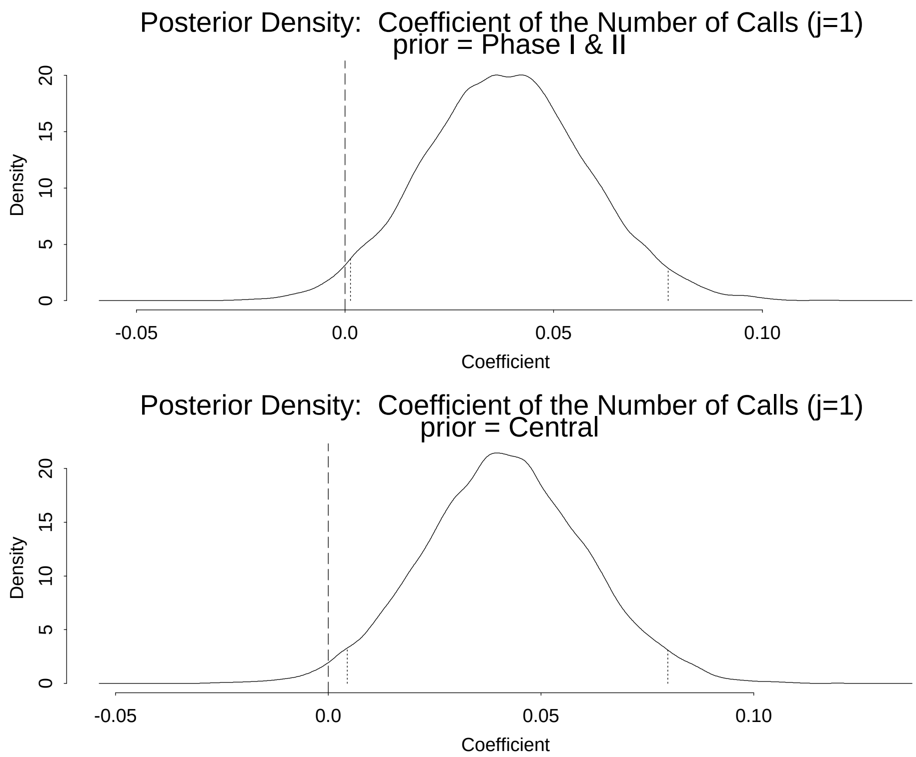

Figure 4: Display of β_{call_1}, the coefficient of the calls variable in η_{i1}: posterior density (solid line) and 95% equal tailed credible interval (dotted line), compared to $\beta_{call_1} = 0$ (dashed line).

Table 1: 95% HPD credible intervals for β_{call_1} under six different prior distributions

Prior	Coefficient of the number of Calls "C_i" in η_{i1}	
	95% intervals	
	Lower Bound	Upper Bound
Phase I&II	0.00129	0.07746
Central	0.00446	0.07980
Option 3	0.00447	0.07983
Option 4	0.00441	0.07975
Option 5	0.00440	0.07970
Option 6	0.00436	0.07944

Table 2: Comparison of the odds of response for 4 typical subjects. Posterior medians were used as the point estimates for the coefficients in the Bayesian models; the MLE was used for the frequentist model.

	Subject 1	Subject 2	Subject 3	Subject 4
Smoker	No	No	Former	Yes
Age	30	50	27	40
Bother	Usually	Always	No	No
K-risk	11	12	7	3
Model	Odds Y=1/Y=0			
MAR MLE	0.674	2.105	0.457	0.396
MAR Phase I & II prior	0.703	4.487	0.209	0.116
NI Phase I & II prior: 2 calls	0.640	4.024	0.202	0.108
NI Central prior: 2 calls	0.593	4.442	0.162	0.102
Option 3: 2 calls	0.594	4.449	0.162	0.102
Option 4: 2 calls	0.592	4.435	0.162	0.101
Option 5: 2 calls	0.590	4.423	0.161	0.101
Option 6: 2 calls	0.590	4.426	0.161	0.101
NI Phase I & II prior: 13 calls	0.974	6.128	0.308	0.165
NI Central prior: 13 calls	0.936	7.013	0.256	0.160
Option 3: 13 calls	0.937	7.026	0.256	0.161
Option 4: 13 calls	0.934	7.000	0.255	0.160
Option 5: 13 calls	0.930	6.975	0.254	0.159
Option 6: 13 calls	0.931	6.980	0.254	0.160