RESEARCH REPORT SERIES
(Statistics #2012-14)

Effects of Missing Data On Modeling Enumeration Status In The U.S. Census

Ryan Janicki
Eric Slud

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 24, 2012

Effects of Missing Data On Modeling Enumeration Status In The U. S. Census*

Ryan Janicki and Eric Slud
Center for Statistical Research and Methodology
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233, USA

Keywords: missing data, sensitivity analysis, nonignorable nonresponse, census coverage, census enumerations

Abstract

The Census Coverage and Measurement program at the U. S. Census Bureau uses dual system estimation to measure the accuracy of the decennial census. Construction of the dual system estimator involves first estimating the percentage of correctly enumerated persons in the decennial census within different domains. This estimation is complicated by the presence of unresolved census enumerations (missing data) for which enumeration status (correct or erroneous) can not be determined. Furthermore, there is concern that the propensity to respond depends on enumeration status, that is, that the missing data are not missing at random. This paper is an exploration of different missing data models and their effect on the prediction of enumeration status, and in particular a comparison of missing at random and not missing at random data models.

1 Introduction

The goal of the decennial census is to count every person and housing unit in the United States once and in the correct place. This goal can not be completely achieved since some people are omitted that should have been included in the decennial census, while some individuals that were enumerated in the census are duplicates, are recorded in the wrong location, are not residents, or are fictitious.

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau. The authors would like to thank Doug Olson and Jerry Maples for their careful review of this paper and helpful comments.

To measure the coverage of the decennial census, the Census Coverage and Measurement (CCM) program conducts a second, operationally independent post-enumeration survey, which can be compared to the original census and used to measure coverage error. A sample of block clusters is taken, the people in the sample are interviewed, and further effort is expended to resolve enumeration status in the census of persons enumerated in this sample. The residents in this sample of block clusters are called the P-sample. The sample of enumerations in the decennial census corresponding to households located in the P-sample block clusters is called the E-sample. The information contained in the P- and E-sample can be used both to estimate net error and to improve the coverage of future censuses.

The technique of using independent surveys to estimate a population total is known as *dual-system estimation*, and is based on the capture-recapture method (Seber, 1982; Hogan, 1993). A *dual-system estimator* of the population total can be constructed based on the number of matches between the two surveys. The form of the dual-system estimator used by the CCM program since the 2000 decennial census is (Cantwell and Ikeda, 2003)

$$\widehat{N} = (C - I)\,\widehat{p}_{ce}\left(\frac{1}{\widehat{p}_{match}}\right), \tag{1}$$

where C is the official census count, I is the number of whole-person imputations, $\widehat{p}_{ce}$ is the estimated probability of correct enumeration in the census, and $\widehat{p}_{match}$ is the estimated probability that a person in the P-sample matches to someone enumerated in the census. The dual system estimator $\widehat{N}$ is computed separately within different domains.

This paper is primarily concerned with estimating the probability of correct enumeration p_{ce} in the census.[1] A correct enumeration (CE) is the record of a person that is complete and was counted exactly once and in the correct location in the census. Examples of erroneous enumerations (EE) include people counted multiple times, counted in the wrong place, (for example, because their usual residence was elsewhere, such as in a college dormitory), or fictitious records.

The E-sample can be matched and compared to the P-sample to determine which census records are correctly enumerated and which records are erroneously enumerated within different domains. For the majority of the E-sample, there is sufficient information to determine a record's enumeration status. However, for 5% (4.69% weighted) of the E-sample, not enough data could be collected to determine whether a census enumeration is correct or erroneous. For example, it is possible that a person moved after the decennial census, but before an interviewer is able to follow up with that person, so that it is not possible to obtain the information necessary to determine whether that person was correctly enumerated. A census record for which enumeration status cannot be determined is said to be unresolved. Since the variable of interest is not observed for unresolved enumerations, these census enumerations can be thought of as missing data.

[1] From 1980 – 2000, CCM used post-stratification to estimate p_{ce} and in 2010 logistic regression was used.

When estimating the proportion of correct enumerations within a domain, the usual assumption is that the unresolved enumerations are stochastically equivalent to the resolved census enumerations, after conditioning on observable covariates. This is the case of ignorable nonresponse, or of a missing at random missing data mechanism. However, the missing at random assumption is not verifiable in the absence of auxiliary distributional assumptions on the joint behavior of enumeration status, resolved status, and covariates, and there is concern that resolved status could depend on the unobservable enumeration status, that is, the situation of nonignorable nonresponse, or of a not missing at random missing data mechanism. The goal of this paper is to investigate the sensitivity of CE estimates to different missing data methods, in particular methods which allow for the possibility that missingness and enumeration status are correlated, and to understand the sensitivity of estimates and imputations to different missing data models and missing data assumptions. In Section 2, methods of estimation in the presence of missing data are discussed. Section 3 describes the available data and selection of predictors for the missing data models. A summary of estimates and model sensitivity is given in Section 4, and concluding remarks are made in Section 5.

2 Estimation methods

Let $\mathcal{U}_E = \{1, \ldots, N\}$ denote the set of census records, consisting of N labelled individuals living in housing units. $\mathcal{U}_D$ is the set of census *data-defined* records which is a subset of $\mathcal{U}_E$ which meet certain rule-based definitions and are deemed sufficiently detailed that each must correspond to a unique individual person. With each record i in $\mathcal{U}_D$, there is an associated value $Y_i \in \{1, 0\}$ corresponding to whether individual i was correctly enumerated or erroneously enumerated in the census, respectively, and a vector of covariates $\mathbf{X}_i$ consisting of characteristics of the individual and the census record. A sample $\mathcal{S}_E$, called the E-sample, is taken from $\mathcal{U}_D$, and to each individual $i \in \mathcal{S}_E$, there is an associated survey weight $w_i = 1/P\left(i \in \mathcal{S}_E\right)$.

The interest is in estimating $P\left(Y = 1 \mid \mathbf{X}_i\right)$. The response variable Y and the covariates $\mathbf{X}$ may be related through a vector of parameters $\boldsymbol{\beta}$ and a functional form $p_i\left(\boldsymbol{\beta}\right) = P\left(Y_i = 1 \mid \mathbf{X}_i, \boldsymbol{\beta}\right)$. Based on this relationship, an estimating function $\boldsymbol{\Psi} = \boldsymbol{\Psi}\left(Y_i, \boldsymbol{X}_i; \boldsymbol{\beta}\right)$ can be specified. The *census estimating function* $\boldsymbol{\Psi}$ has the same dimension as the vector of parameters $\boldsymbol{\beta}$, and the census parameter $\boldsymbol{\beta}_C$ is defined implicitly as the solution to the *census estimating equation*

$$\boldsymbol{\Psi}_C = \sum_{i \in \mathcal{U}_D} \boldsymbol{\Psi}\left(Y_i, \mathbf{X}_i; \boldsymbol{\beta}\right) = \mathbf{0}. \tag{2}$$

The form of the estimating function $\boldsymbol{\Psi}$ may be motivated by the form of the data or the model assumptions that the investigator is willing to make and will be discussed further in Section 2.1. While a model ξ may be used to motivate the form of the estimating function, it is not necessary for the full model to be specified for $\boldsymbol{\Psi}$ to produce valid (consistent) estimates, so long as $E_\xi\left(\boldsymbol{\Psi}\right) =$

$\mathbf{0}$, $E_\xi\left(\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right) < \infty$, and $E_\xi\left(\partial\boldsymbol{\Psi}/\partial\boldsymbol{\beta}\right)$ is nonsingular, and instead it may be that only the conditional mean of the distribution is specified, for example. The tradeoff for having only partial model specification and the corresponding robustness from lack of model dependency is that there could be loss of efficiency in the estimates.

The census estimating function (2) can not be directly used when only a sample is observed, rather than values from the entire finite population. It can, however, be estimated using the observed sample data $\mathcal{S}_E$, through the use of a Horvitz-Thompson estimator of (2),

$$\boldsymbol{\Psi}_S = \sum_{i\in\mathcal{S}_E} w_i \boldsymbol{\Psi}\left(Y_i, \mathbf{X}_i; \boldsymbol{\beta}\right). \tag{3}$$

The estimating function $\boldsymbol{\Psi}_S$ in equation (3) is called the *sample estimating function*. Note that $E\left(\boldsymbol{\Psi}_S\right) = \boldsymbol{\Psi}_C$, where the expectation is taken with respect to the sample design, so that $\boldsymbol{\Psi}_S$ is an unbiased estimator of $\boldsymbol{\Psi}_C$. It was shown in Godambe and Thompson (1986a) that $\boldsymbol{\Psi}_S$ is the optimal estimating function for $\boldsymbol{\Psi}_C$, in terms of minimizing the mean squared distance to $\boldsymbol{\Psi}_C$ over the class of estimating functions $\widetilde{\boldsymbol{\Psi}}$ such that $E(\widetilde{\boldsymbol{\Psi}}) = \boldsymbol{\Psi}_C$. An estimate of the census parameter $\boldsymbol{\beta}_C$ is $\widetilde{\boldsymbol{\beta}}$, the solution to the sample estimating equation $\boldsymbol{\Psi}_S = \mathbf{0}$.

Now, suppose for some values $i \in \mathcal{S}_M \subset \mathcal{S}_E$, the response variable Y_i is not observed due to nonresponse. Let $R_i = 1$ if $i \in \mathcal{S}_E \setminus \mathcal{S}_M$ and $R_i = 0$ if $i \in \mathcal{S}_M$. For the moment, suppose $\pi_i = P\left(R = 1 \mid Y_i, \mathbf{X}_i\right)$ is known. It was shown in Godambe and Thompson (1986b) that the estimating function

$$\boldsymbol{\Psi}^* = \sum_{i\in\mathcal{S}_E} w_i \frac{R_i}{\pi_i} \boldsymbol{\Psi}\left(Y_i, \mathbf{X}_i; \boldsymbol{\beta}\right) \tag{4}$$

is the optimal estimating function for $\boldsymbol{\beta}$ in the presence of missing data over the class of estimating functions $\widetilde{\boldsymbol{\Psi}}$ such that $E(\widetilde{\boldsymbol{\Psi}} \mid \mathbf{Y}, \boldsymbol{X}) = \boldsymbol{\Psi}_S$. The statistic $\widehat{\boldsymbol{\beta}}$ which is the solution to the estimating equation $\boldsymbol{\Psi}^* = \mathbf{0}$ can be used as an estimator for $\boldsymbol{\beta}_C$.

2.1 Choice of estimating function

Since Y_i is binary, a sensible choice for the form of $p_i\left(\boldsymbol{\beta}\right)$ is a class of generalized linear models (McCullagh and Nelder, 1989) with suitable link function g; that is, conditional on $\mathbf{X}_i$ and $\boldsymbol{\beta}$,

$$g\left(E\left(Y \mid \mathbf{X}_i, \boldsymbol{\beta}\right)\right) = g\left(P\left(Y = 1 \mid \mathbf{X}_i, \boldsymbol{\beta}\right)\right) = \mathbf{X}_i\boldsymbol{\beta} = \sum_{j=1}^{p} X_{ij}\beta_j.$$

Let $p_i\left(\boldsymbol{\beta}\right) = P\left(Y = 1 \mid \mathbf{X}_i, \boldsymbol{\beta}\right) = h\left(\mathbf{X}_i\boldsymbol{\beta}\right)$, where $h = g^{-1}$ is the inverse of the link function. The estimating function

$$\boldsymbol{\Psi}\left(Y_i, \mathbf{X}_i; \boldsymbol{\beta}\right) = \frac{Y_i - p_i\left(\boldsymbol{\beta}\right)}{p_i\left(\boldsymbol{\beta}\right)\left(1 - p_i\left(\boldsymbol{\beta}\right)\right)} \frac{\partial p_i\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}$$

can be used for inference on $\boldsymbol{\beta}$. One popular choice of link function is the logistic link, $g_L(x) = \log(p/(1-p))$ with inverse $h_L(x) = \text{expit}(x) = e^x/(1+e^x)$. The corresponding estimating function is

$$\boldsymbol{\Psi}_L(Y_i, \mathbf{X}_i; \boldsymbol{\beta}) = (Y_i - p_i(\boldsymbol{\beta}))\mathbf{X}_i.$$

The complementary log-log link $g_C(x) = \log(-\log(1-p))$ with inverse $h_C(x) = 1 - e^{-e^x}$ is also of interest in this work. The estimating function based on the complementary log-log link is

$$\boldsymbol{\Psi}_C(Y_i, \mathbf{X}_i; \boldsymbol{\beta}) = -\frac{\log(1-p_i(\boldsymbol{\beta}))}{p_i(\boldsymbol{\beta})}(Y_i - p_i(\boldsymbol{\beta}))\mathbf{X}_i.$$

2.2 Missing data models

The response probabilities π_i in (4) are rarely known in practice, and must be estimated (Kim and Kim, 2007). Different choices of models for the probability of nonresponse π_i are investigated in this section. There are three types of nonresponse (Little and Rubin, 2002). If missingness does not depend on any data values, then the data are said to be *missing completely at random* (MCAR). If missingness depends only on observed data values, then the missing data mechanism is said to be *missing at random*. If the missingness depends on unobservable data, then the missing data mechanism is said to be *not missing at random* (NMAR). Five different missing data models for $\pi_i(\boldsymbol{\alpha}) = P(R = 1 \mid Y_i, \mathbf{X}_i, \boldsymbol{\alpha})$ were considered:

- missing completely at random (MCAR)

$$\pi_i(\boldsymbol{\alpha}) \equiv \pi,$$

- missing at random (MAR)

$$\pi_i(\boldsymbol{\alpha}) = \text{expit}\left(\alpha_0 + \mathbf{X}_i^T\boldsymbol{\alpha}_1\right),$$

- not missing at random 1 (NMAR1)

$$\pi_i(\boldsymbol{\alpha}) = \text{expit}(\alpha_0 + \alpha_1 Y_i),$$

- not missing at random 2 (NMAR2)

$$\pi_i(\boldsymbol{\alpha}) = \text{expit}\left(\alpha_0 + \alpha_1 Y_i + \mathbf{X}_i^T\boldsymbol{\alpha}_2\right),$$

- not missing at random 3 (NMAR3)

$$\pi_i(\boldsymbol{\alpha}) = \text{expit}\left(\alpha_0 + \alpha_1 Y_i + \mathbf{X}_i^T\boldsymbol{\alpha}_2 + Y_i\mathbf{X}_i^T\boldsymbol{\alpha}_3\right).$$

For the MCAR model,

$$\widehat{\pi} = \frac{\sum_{i \in \mathcal{S}_E} w_i R_i}{\sum_{i \in \mathcal{S}_E} w_i}, \tag{5}$$

and $\boldsymbol{\beta}$ can be estimated by using the estimating function (3). Note that when the missing data are missing completely at random, $\pi_i \equiv \pi$ for all i, and π_i does not affect estimation of $\boldsymbol{\beta}$. For the MAR model, $\boldsymbol{\alpha}$ can be estimated by solving the system

$$\boldsymbol{\Phi} = \sum_{i \in \mathcal{S}_E} \frac{w_i}{\pi_i(\boldsymbol{\alpha})} \left(R_i - \pi_i(\boldsymbol{\alpha})\right) \begin{bmatrix} 1 \\ \mathbf{X}_i \end{bmatrix} = \mathbf{0},$$

and $\boldsymbol{\beta}$ is estimated by using (3), with π_i replaced by $\widehat{\pi}_i = \pi_i(\widehat{\boldsymbol{\alpha}})$.

For the different NMAR models, estimation of $\boldsymbol{\alpha}$ is more difficult, because while R_i is observed for all $i \in \mathcal{S}_E$, Y_i is not observed if $R_i = 0$. If there were no missing data, $\boldsymbol{\alpha}$ could be estimated in the NMAR1 model, for example, by solving the system

$$\boldsymbol{\Phi}^* = \sum_{i \in \mathcal{S}_E} \frac{w_i}{\pi_i(\boldsymbol{\alpha})} \left(R_i - \pi_i(\boldsymbol{\alpha})\right) \begin{bmatrix} 1 \\ Y_i \end{bmatrix} = \mathbf{0}. \tag{6}$$

Since Y_i is not observable for all $i \in \mathcal{S}_E$, an adjustment has to be made to $\boldsymbol{\Phi}^*$ for inference on $\boldsymbol{\alpha}$. One possibility is to replace Y_i in equation (6) with its expectation $p_i(\boldsymbol{\beta}) = E(Y_i \mid \mathbf{X}_i, \boldsymbol{\beta})$ under the assumed model. Define $\boldsymbol{\Phi}$ to be the estimating function given in equation (6), with Y_i replaced by $p_i(\boldsymbol{\beta})$. Estimating functions for the other NMAR models can be constructed in a similar way by replacing each occurrence of Y_i by its expectation $p_i(\boldsymbol{\beta})$; this substitution does not affect the unbiasedness of the estimating function as its expectation remains equal to zero. However, the solution $\widehat{\boldsymbol{\alpha}}$ can still be biased in small and moderate samples, and the form of the plug-in for Y_i can affect both the bias and the variability of the estimators. Note that $\boldsymbol{\Phi}$ and the estimating function in equation (4) depend on both parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, so that inference must be made by solving the system $\left(\boldsymbol{\Psi}^T, \boldsymbol{\Phi}^T\right)^T = \mathbf{0}$. In the examples in this paper, this system was solved using the Newton-Raphson algorithm.

Using the inverse of the probability of response as in equation (6) to modify the estimating equations in the presence of nonresponse was considered in Robins et al. (1994). This inverse probability weighting is in the spirit of Horvitz and Thompson (1952), and creates new weighting classes with large weights given to respondents which are similar to nonrespondents, under the assumed model. The extension to nonignorable nonresponse using inverse probability weighted estimating functions was considered in, for example, Rotnitzky and Robins (1997) and Sharfstein et al. (1999).

The ignorable model is embedded in each of the nonignorable models through the submodel $\alpha_1 = 0$ for models NMAR1 and NMAR2 and through the submodel $\alpha_1 = 0$ and $\boldsymbol{\alpha}_3 = \mathbf{0}$ for model NMAR3. Because of this, it is tempting to consider tests for the hypothesis $\alpha_1 = 0$ (or $\alpha_1 = 0$ and $\boldsymbol{\alpha}_3 = \mathbf{0}$) to determine

whether an ignorable or a nonignorable model is more appropriate for a given data set and a given model specification.

It was shown in Rotnitzky and Robins (1997) in model NMAR1, if the true value of the parameter is $\alpha_1 = 0$, that there does not exist a regular asymptotically linear (RAL) estimator of $\boldsymbol{\beta}$ (that is, an estimator that is asymptotically Gaussian, with a $\sqrt{n}$ rate of convergence, such that the convergence is locally uniform). It can be shown that the same phenomenon occurs in models NMAR2 and NMAR3. The reason for the non-existence of RAL estimators in the NMAR models is that the matrix $E_\xi \left(\partial \boldsymbol{\Psi} / \partial \boldsymbol{\beta}\right)|_{\alpha_1=0}$ is singular, hence a hypothesis test of $\alpha_1 = 0$ (or $\alpha_1 = 0$ and $\boldsymbol{\alpha}_3 = \mathbf{0}$) cannot be constructed in the usual way (using, for example, a Wald test statistic or a Rao-score test statistic) to distinguish between a NMAR model and a MAR submodel since there is no longer a $\sqrt{n}$ rate of convergence. See also Molenberghs et al. (2008). However, while inference depends critically on model assumptions, nonignorable models can still be useful as part of a sensitivity analysis, to better understand how the specification of the missing data mechanism can affect the analysis.

3 Summary of data and variable selection

The sample $\mathcal{S}_E$ consists of 5681 block clusters with a total of 370,505 people. For each record, there is associated an 800 dimensional vector of covariates describing, for example, characteristics of the individual, geographic location, and data collection methods, which are observable for all persons in sample, regardless of resolved status. Due to the high dimensionality of the data set, variable selection was done in the interest of having a more parsimonious and interpretable model while maintaining predictive power.

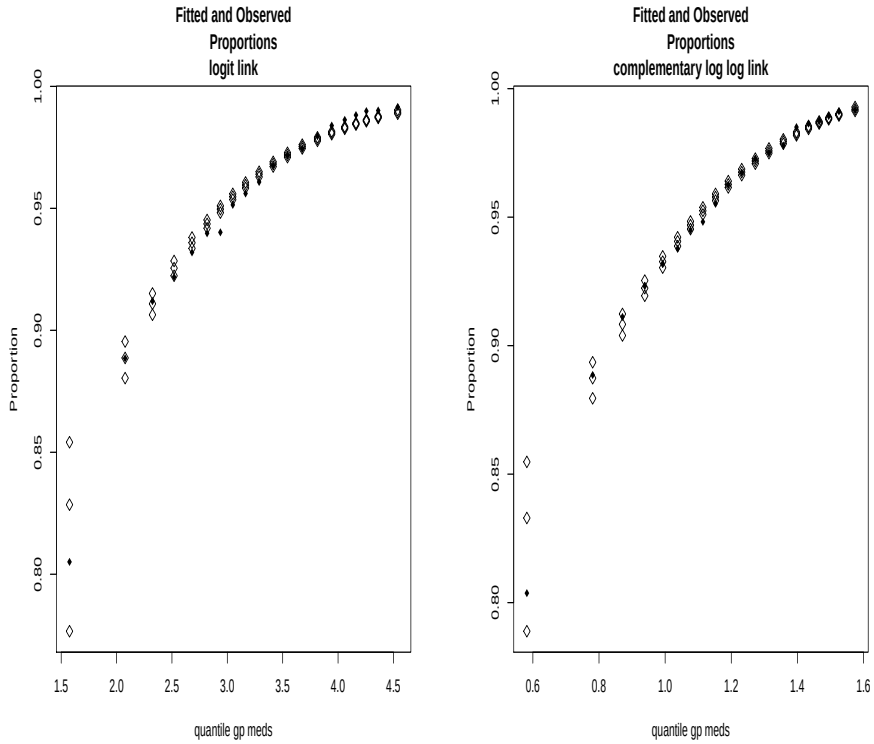
Stepwise variable selection (Venables and Ripley, 2002, p. 175) was used to choose a set of predictive variables. For this exploratory data analysis, the survey weights were ignored, and the data was treated as a sample of independent random variables. The group lasso of Meier et al. (2008) was also investigated for model selection, but it was found that the selected models were generally the same as models selected using stepwise variable selection.

The stepwise variable selection procedure fits a sequence of nested models by either adding or removing a covariate, depending on which step minimizes a chosen objective function. For modeling correct enumeration rate, a BIC penalty was used, so that the objective function was

$$BIC = -2\text{log-likelihood} + p \log n,$$

where p is the number of parameters in the model and n is the sample size. A final model is chosen when no added covariate and no removed covariate can reduce the BIC.

The log-likelihood depends on the choice of link function, and initially, the logistic link was used. However, the model chosen using stepwise variable selection, a logistic link, and a BIC penalty did not appear to result in adequate



(a) logit link

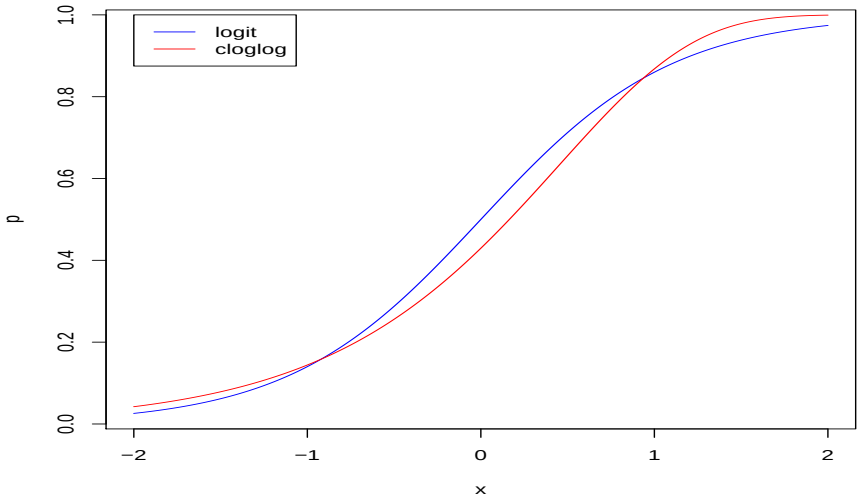
(b) complementary log-log link

Figure 1: fitted and observed proportions

fit, as can be seen in Figure 1(a). Figure 1(a) compares the predicted and observed probability of correct enumeration, by first dividing the linear predictors $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ into 20 quantile groups, then plotting the predicted and observed values in each quantile group. The open diamonds in each column represent the 25th, 50th, and 75th percentiles of the predicted values in each quantile group and the filled diamond in each column represents the mean of the observed data corresponding to the predicted values in each quantile group. This type of figure can be used as a graphical tool for assessing the goodness-of-fit of the model. Further discussion about "binning" as a diagnostic tool in binary regression can be found in Gelman and Hill (2007, ch. 5).

In Figure 1(a), the mean of the observed values are below the 25th percentile of the predicted values in the middle bins, and above the 75th percentile of the predicted values in the upper bins, suggesting the logistic link is inappropriate for this data set. This is most likely due to the fact that in most domains the average probability of correct enumeration was large, and the logistic link function increases too slowly, resulting in predictions that were too small for

Figure 2: Comparison of standardized link functions



high probability events.

Alternatively, the complementary log-log link was considered. Figure 2 compares the shape of the logistic link function and the complementary log-log link. The complementary log-log link increases much more rapidly than does the logistic link, and was more appropriate for this problem. Model selection was done using stepwise regression and a BIC penalty, but with the complementary log-log link in the log-likelihood. A comparison of the predicted and observed probabilities are shown in Figure 1(b).

The final set of covariates selected using stepwise variable selection and the complementary log-log link is denoted by $\mathbf{X}_i$, and consists of the 15 variables and 11 interactions (for a total of 75 coefficients)

- type of housing unit (huTyp) (single-unit; multi-unit; trailer; other)
- proxy type (PROX) (household member on 4 – 1; household member moved in after 4 – 1; neighbor or other proxy; other)
- relationship type (relType) (member of the nuclear family; adult child of the householder; other member of the household)
- participation rate (btPart) (0.5 – 0.92, treated as continuous)
- eligible for supplemental nonresponse followup universe (supnrfu) (address is not a valid decennial address, eligible to be part of the supplemental nonresponse follow up universe; address is a valid decennial address, eligible to be part of the supplemental nonresponse follow up universe; not eligible to be part of the supplemental nonresponse follow up universe)

- (msatea) (MSA = L and TEA = 1 or 6; MSA = M and TEA = 1 or 6; MSA = S and TEA = 1 or 6; MSA = X and TEA = 1 or 6; MSA = L, M, or S and TEA = 2 or 7; MSA = X and TEA = 2 or 7; TEA = 3 or 5)
  - MSA[2] (L = one of the 12 largest; M = not one of the 12 largest and $>$ 500,000; S = $<$ 500,000; X = 0)
  - TEA[3] (1 = Mailout / Mailback; 2 = Update / Leave; 3 = List / Enumerate; 5 = Rural Update / Enumerate; 6 = Military; 7 = Urban Update / Leave)
- composition of household (typeHH) (married-couple family; other family; non-family with householder living alone; non-family with householder not living alone; occupied, but does not fulfill other requirements for other types)
- log of the number in household (palt) (treated as continuous)
- (repmail) (blanketed replacement mailing block; targeted replacement mailing block; not targeted replacement mailing block)
- age (0 – 17; 18 – 29; 30 – 49; 50 +)
- mailReturn (yes; no)
- sex
- characteristic imputation flag (charImpFlag) (all characteristics reported; at least one characteristic imputed)
- date of birth reported indicator (dobInd) (day, month, and year of birth reported; at least one not reported)
- race / ethnicity (race[4]) (American Indian or Alaska Native on Reservation; off-reservation American Indian or Alaska Native; Hispanic; non-Hispanic Black; native Hawaiian or Pacific islander; non-Hispanic Asian; non-Hispanic White)
- interaction terms (relType * palt; huTyp * typeHH; relType * btPart; supnrfu * mailReturn; palt * mailReturn; palt * cmAge; typeHH * palt; relType * typeHH; typeHH * mailReturn; btPart * palt; palt * sex)

For the missing data models, the set of covariates $\mathbf{X}_i$ was used as the universe of possible predictors and stepwise variable selection with a logistic link function was done on this set with a BIC penalty. The selected predictors for nonresponse, denoted by $\widetilde{\mathbf{X}}_i$, are huTyp, PROX, relType, btPart, supnrfu, msatea, typeHH, palt, repmail, age, sex, charImpFlag, dobInd, domain, relType

[2]MSA = metropolitan statistical area

[3]TEA = type of enumeration area

[4]race is modeled using the seven-category Race-Origin Domain described in Mulligan and Davis (2012)

Table 1: Summary of predicted probability of being resolved $\widehat{P}(R=1 \mid \mathbf{X})$

	Percentiles						
Model	0.01	0.10	0.25	0.50	0.75	0.90	0.99
MCAR	0.953	0.953	0.953	0.953	0.953	0.953	0.953
MAR	0.747	0.896	0.944	0.971	0.982	0.988	0.992
NMAR1	0.845	0.915	0.940	0.962	0.975	0.979	0.982
NMAR2	0.747	0.896	0.944	0.971	0.982	0.988	0.993
NMAR3	0.720	0.941	0.972	0.988	0.994	0.996	0.998

* palt, and palt * age. For model NMAR3, the covariates used for the 'nonignorable' interaction terms $Y * X_i$ are the variables relType, charImpFlag, dobInd, and huTyp, which were among the first few variable selected in the stepwise procedure for choosing a missing data model.

Solving the different estimating equations under missing data models MCAR, MAR, NMAR1 and NMAR2 was easily done using the Newton-Raphson algorithm. However, solving the estimating equation using model NMAR3 was more challenging. It was found that the more nonignorable interaction terms included in the model, the more difficult it was to find a root of the corresponding estimating equation; the Newton-Raphson algorithm became more sensitive to the choice of starting value, and the size of each step needed to be small to avoid large jumps which made the iterations computationally intractable, that is the solution was found iteratively using the steps

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - \rho \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi} \left(\boldsymbol{\theta}^{(n)} \right) \right)^{-1} \boldsymbol{\Psi} \left(\boldsymbol{\theta}^{(n)} \right)$$

for small values of ρ, rather than the usual Newton-Raphson algorithm with $\rho = 1$. Due to the small jump sizes, the computational time is greatly increased.

4 Summary of results

This section summarizes the parameter estimates and predicted values under the different models. All results in this section are based on the generalized linear model with a complementary log-log link function, the covariates described in Section 3 for the correct enumeration model, and the five different missing data models described in Section 2.2.

Table 1 summarizes the marginal distribution of the predicted probability of being resolved

$$\widehat{F}_n(t) = \frac{1}{\sum_i w_i} \sum_i w_i I \left\{ \widehat{P}(R=1 \mid \mathbf{X}_i) \leq t \right\}$$

under each of the five missing data models. For model MCAR, the probability of being resolved does not depend on any covariate or response variable, hence

Table 2: Mean predicted correct enumeration rate within domain, unresolved only

	Missing Data Model				
	MCAR	MAR	NMAR1	NMAR2	NMAR3
huTyp M	0.9136	0.9144	0.8558	0.9135	0.8986
huTyp O	0.8303	0.8319	0.7378	0.8307	0.8041
huTyp S	0.9483	0.9484	0.9113	0.9481	0.9290
huTyp T	0.9142	0.9145	0.8573	0.9140	0.8978
race 1	0.9111	0.9114	0.8523	0.9110	0.8968
race 2	0.8951	0.8952	0.8304	0.8945	0.8746
race 3	0.9244	0.9249	0.8728	0.9243	0.9073
race 4	0.9229	0.9235	0.8710	0.9227	0.9022
race 5	0.8971	0.8982	0.8369	0.8974	0.8799
race 6	0.9334	0.9342	0.8872	0.9337	0.9188
race 7	0.9286	0.9290	0.8811	0.9285	0.9116
Total	0.9256	0.9261	0.8758	0.9255	0.9080

the probability of being resolved is the same for all individuals and can be calculated using equation (5). The NMAR models considered are conditional on the response variable Y. Since $\widehat{P}(R = 1 \mid Y, \mathbf{X})$ can not calculated for the sampled individuals for which Y is missing, Table 1 gives the distribution of the marginal probabilities $\widehat{P}(R = 1 \mid \mathbf{X}) = \sum_{y=0}^{1} \widehat{P}(R = 1 \mid Y = y, \mathbf{X}) \times \widehat{P}(Y = y \mid \mathbf{X})$.

Notice that the distribution of the predicted probability of being resolved using the MAR model is nearly identical to the distribution under the NMAR2 model. Recall that model MAR is a submodel of model NMAR2 when $\alpha_1 = 0$. The estimated value of α_1 in model NMAR2 is 0.074. Since there are 42 other variables used in the model to predict resolved status, the effect of α_1 in the predicted values of resolved status was minimal. The only real difference between the predicted values is in the tails, and even these differences are small. Further discussion of the estimated parameter values is given in Section 4.2.

Table 2 shows the mean predicted correct enumeration rate of the unresolved census enumerations within different domains. The domains used in Table 2 are the housing unit types (huTyp) and race categories. Two items stand out in Table 2. First, there is hardly any difference in the mean predicted probabilities of correct enumeration within the presented domains for models MCAR, MAR and NMAR2. While the lack of difference in the predicted values using models MAR and NMAR2 can be attributed to the near-zero estimated coefficient of Y in model NMAR2, the lack of difference in models MCAR and MAR is surprising. Two factors may account for this: first, there is a large observed sample that can be used to fit the models, and a relatively small amount of missing data. Second, there is a large set of covariates which can be used as predictors, and stepwise variable selection was done to select a subset for the model. The

lack of effect of the MAR missing data model on the predicted probabilities of correct enumeration suggest that the covariates used in the correct enumeration model have sufficient predictive power to eliminate the effect of the missing data mechanism in this model.

The second interesting point in Table 2 is the comparison of models NMAR1 and NMAR3 with the other models. Recall that intuitively, it should be expected that it is more difficult to resolve a correct enumeration than an erroneous enumeration, so that if the missing data mechanism does depend on enumeration status, a NMAR assumption should reduce the probability of correct enumeration over a MAR or MCAR assumption. While the results presented in Table 2 do conform to intuition, in that the average predicted probability of correct enumeration is reduced, it is somewhat surprising that the effects on CE rate are as small as they are. The missing data mechanism specified in model NMAR1 is that enumeration status is the *only* variable which affects the propensity to respond. Hence NMAR1 can be thought of as an extreme case. Yet even in this extreme case, the average predicted probability of correct enumeration within the unresolved census enumerations is only reduced from 0.9256 to 0.8758. Again, this suggests good predictive power in the covariates used to model enumeration status, mitigating effects of the specification of the missing data model.

4.1 Variance estimation

There are two general strategies for variance estimation – Taylor series linearization and replication methods. Let $\boldsymbol{\theta}^T = \left(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T\right)$ and $\boldsymbol{\Upsilon}^T = \left(\boldsymbol{\Psi}^T, \boldsymbol{\Phi}^T\right)$. Using a Taylor series linearization, the variance of $\widehat{\boldsymbol{\theta}}$ can be estimated by (Fuller, 2009, p. 68)

$$Var\left(\widehat{\boldsymbol{\theta}}\right) \approx \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Upsilon}\left(\widehat{\boldsymbol{\theta}}\right)\right)^{-1} \widehat{Var}_{HT}\left(\boldsymbol{\Upsilon}\left(\widehat{\boldsymbol{\theta}}\right)\right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Upsilon}\left(\widehat{\boldsymbol{\theta}}\right)^T\right)^{-1}. \tag{7}$$

The Horvitz-Thompson variance estimator of $\boldsymbol{\Upsilon}$ requires knowledge about the joint inclusion probabilities, which are not available. However, $\widehat{Var}_{HT}\left(\boldsymbol{\Upsilon}\right)$ can be approximated by (Morel, 1989; Natarajan et al., 2008)

$$\widehat{Var}\left(\boldsymbol{\Upsilon}\left(\boldsymbol{\theta}\right)\right) = \sum_{i=1}^{n} \left(\boldsymbol{\Upsilon}_{i\cdot} - \boldsymbol{\Upsilon}_{\cdot\cdot}\right)\left(\boldsymbol{\Upsilon}_{i\cdot} - \boldsymbol{\Upsilon}_{\cdot\cdot}\right)^T,$$

where $\boldsymbol{\Upsilon}_{ij} = \boldsymbol{\Upsilon}\left(Y_{ij}, \mathbf{X}_{ij}; \boldsymbol{\theta}\right)$, $\boldsymbol{\Upsilon}_{i\cdot} = \sum_{j=1}^{n_i} \boldsymbol{\Upsilon}_{ij}$, and $\boldsymbol{\Upsilon}_{\cdot\cdot} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \boldsymbol{\Upsilon}_{ij}$. Here $i = 1, \ldots, n = 5681$ indexes the block clusters and $j = 1, \ldots, n_i$ indexes the records in the ith block cluster.

Alternatively, replication-based variance estimates can be made. Rao and Tausi (2004) give a delete-one-cluster Jackknife method for estimating the variance of estimators obtained by solving survey-weighted estimating equations. It was found that since there are 5681 clusters in this problem, that this method is

Table 3: 'nonignorable' coefficients

Model	variable	estimate	se	95% CI
NMAR1	Y	4.1420	0.2587	(3.6349, 4.6491)
NMAR2	Y	0.0740	0.5485	(−1.0011, 1.1491)
NMAR3	Y	2.2788	0.6108	(1.0816, 3.4760)
	$Y * huTypO$	0.1659	0.5076	(0.8290, 1.1608)
	$Y * huTypS$	1.7220	0.3459	(−1.0440, 2.4000)
	$Y * huTypT$	0.6712	0.5758	(0.4574, 1.7998)
	$Y * relTyp2$	−0.8419	0.4206	(−1.6663, −0.0175)
	$Y * relTyp3$	−0.8727	0.2047	(−1.2739, −0.4715)
	$Y * dobInd$	−1.6253	0.4975	(−2.6004, −0.6502)
	$Y * charImpFlag$	−0.1612	0.2481	(−0.6475, 0.3251)

too computationally demanding. The less computationally demanding delete-a-group jackknife of Kott (2001) was implementable, with 100 groups in this problem. It was found that the Taylor series linearization variance estimates were very similar to the delete-a-group jackknife variance estimates, so only the former are reported. For example, for model NMAR2, the mean absolute difference of the 118 estimated standard errors using Taylor series linearization and the estimated standard errors using the delete-a-group jackknife was 0.0056, while the maximum absolute difference was 0.0373.

4.2 Analysis of nonignorable terms

Table 3 shows the estimates of the 'nonignorable' parameter α_1 in the nonresponse models NMAR1, NMAR2, and the parameters α_1 and $\boldsymbol{\alpha}_3$ in model NMAR3, along with the estimate of the standard errors computed using Taylor series linearization as in equation (7).

The 95% confidence intervals shown in the last column of Table 3 are based on the normal approximation. Using such an approximation it is simple to construct a test statistic, for example the Wald statistic, to test the hypothesis that $\alpha_1 = 0$. However, as discussed in Section 2.2, it is not clear how such a test statistic should be interpreted, since the parameter estimates are no longer regular when the true model is the missing at random submodel of the not missing at random model. Similarly, caution should be taken in interpreting the confidence intervals, particularly when the interval includes 0, such as the interval for the estimate of the coefficient for Y in model NMAR2.

For model NMAR1, the estimated coefficient of 4.142 along with an estimated standard error of 0.2587 indicates a strong influence on enumeration status on the propensity to respond. As was discussed earlier in this section, the nonresponse model NMAR1 is an extreme model, in the sense that resolved status depends only on enumeration status, which is most likely not a reasonable assumption. This model is, however, useful as a reference for comparison to the other models, in the spirit of a sensitivity analysis.

In model NMAR2, since the estimated coefficient of Y is 0.0740, the predicted probabilities are nearly identical to the predicted probabilities using the MCAR or MAR models. While there is little change in the predicted probabilities of correct enumeration, there is a noticeable change in the estimated standard errors. The inclusion of a single nonignorable term in the missing data model on average increased the estimated standard errors of the 75 coefficients in the correct enumeration model by 1.1%, and increased the estimated standard errors of the 42 coefficients in the missing data model by an average of 2.7%.

Because the estimated coefficient of Y is close to zero in model NMAR2, it is tempting to conclude that extending model NMAR1 and conditioning on observable covariates eliminates the nonignorable effect and that a missing at random model is appropriate. However, as was discussed in Section 2.2, if the model is correctly specified and the truth is that the nonignorable parameter is exactly equal to zero, the estimate of the coefficient will not have the typical $\sqrt{n}$ rate of convergence, so that the usual test statistics may not be valid. Because of the unusual behavior of estimators when the nonignorable coefficients are simultaneously equal to zero and the possibility of model misspecification, it is important to investigate larger models with more nonignorable interaction terms to see if a nonignorable effect appears.

Missing data model NMAR3 is an extension of model NMAR2, in that the probability of response depends on the response variable, the observed covariates, and the interaction between the two. The estimated coefficient of Y and the estimated coefficients of the interaction terms $Y * X_i$ can be seen in Table 3. Inclusion of additional interaction terms causes the magnitude of the estimates of the nonignorable coefficients to increase, resulting in predicted probabilities of correct enumerations which fall between the predictions using models NMAR1 and NMAR2 in Table 2.

Conceptually, it is simple to expand the NMAR missing data models with more interaction terms to investigate nonignorable nonresponse associated with different covariates or domains. However, computationally, the inclusion of more interaction terms involving the response variable Y made it much more difficult to find a root to the estimating equation $\boldsymbol{\Upsilon} = \mathbf{0}$. In particular, inclusion of the variable $Y * PROX$ caused difficulties, as the coefficient of the term $Y * PROX3$ increased rapidly at each iteration of the Newton-Raphson algorithm. It is possible that a root to the estimating equation does not exist when certain interactions are included in the missing data model. Another difficulty in expanding the missing data model with more interaction terms to investigate which 'nonignorable' interactions are influential is that the estimated standard errors increase rapidly. The 'more nonignorable' the nonresponse model becomes the more uncertainty is introduced, and there is no unresolved data for which the response variable Y is observed with which to fit the model.

5 Conclusion

This paper investigated missing data alternatives for modeling correct enumeration status in the decennial census. Because it was expected that if the propensity to respond is affected by enumeration status, then being erroneously enumerated should reduce the probability of response, it was not surprising that nonignorable nonresponse models reduced the predicted probabilities, both overall and within different domains. What was something of a surprise was the moderate degree to which nonignorable modeling reduced the predicted probabilities, with the most extreme nonignorable assumptions only reducing the overall predicted probability of correct enumeration from 0.9256 to 0.8758 among the unresolved census enumerations. However, within the domains considered, the effect could be larger (for example, huTyp 0 in Table 2), and it is possible that within finer cross-classifications that were not considered, the effect could be even more dramatic. Some explanations for the relative lack of sensitivity to the choice of missing data model could be the large overall sample size, the small percentage of missing data, and the rich collection of available predictors.

In this paper, only the effect of missing data on modeling correct enumerations was considered. While interesting in its own right, the probability of correct enumeration is only one piece of the dual system estimator in (1). The analysis in the paper can be repeated to model the probability of matching to the census. These estimates could then be combined to better understand the effects of missing data on the dual system estimator. Future work involves investigating the effect of missing data alternatives on the dual system estimator.

Another possibility for future work is to investigate the effects of missing data on the different types of erroneous enumerations. Starting in 2010, the CCM program produced estimates of the four components of census coverage: correct enumerations, erroneous enumerations, whole-person imputations, and omissions. The erroneous enumeration estimate was 10.04 million people, with 8.52 million people coming from duplication and 1.52 million coming from other reasons (Keller and Fox, 2012). Nonignorable missing data models can be used to investigate the sensitivity of the estimate of erroneous enumerations due to duplication to different missing data models.

References

Robert M. Bell and Michael L. Cohen, editors. *Coverage Measurement in the 2010 Census*. The National Academies Press, Washington, D. C., 2009.

Patrick J. Cantwell and Michael Ikeda. Handling missing data in the 2000 accuracy and coverage evalutation survey. *Surv. Methodol.*, 29(2):139 – 153, 2003.

Wayne A. Fuller. *Sampling Statistics*. Wiley, New Jersey, 2009.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge University Press, New York, 2007.

V. P. Godambe and M. E. Thompson. Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.*, 54(2):127 – 138, 1986a.

V. P. Godambe and M. E. Thompson. Some optimality results in the presense of nonresponse. *Surv. Methodol.*, 12:29 – 36, 1986b.

Howard Hogan. The 1990 post-enumeration survey: operations and results. *J. Amer. Statist. Assoc.*, 88(423):1047 – 1060, 1993.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47:663 – 685, 1952.

Don Keathley, Anne Kearney, and William Bell. ESCAP II: analysis of missing data alternatives for the accuracy and coverage evaluation. Technical report, U. S. Census Bureau, 2001. Available at `http://www.census.gov/dmd/www/pdf/Report12.PDF`.

Andrew Keller and Tyler Fox. Components of census coverage for the household population in the united states. Technical report, DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-04, 2012. Available at `http://www.census.gov/coverage_measurement/pdfs/g04.pdf`.

Jae Kwang Kim and Jay J. Kim. Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.*, 35(4):501 – 514, 2007.

Philip S. Kott. The delete-a-group jackkife. *J. Off. Statist.*, 17(4):521 – 526, 2001.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis With Missing Data*. Wiley, New Jersey, second edition, 2002.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, New York, second edition, 1989.

Lukas Meier, Sara van der Geer, and Peter Bühlmann. The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B*, 70(1):53 – 71, 2008.

Geert Molenberghs, Caroline Beunckens, Cristina Sotto, and Michael G. Kenward. Every missing not at random model has a missing at random counterpart with equal fit. *J. Roy. Statist. Soc. Ser. B*, 70(2):371 – 388, 2008.

G. Morel. Logistic regression under complex survey design. *Surv. Methodol.*, 15:203 – 223, 1989.

James Mulligan and Peter P. Davis. 2010 Census Coverage Measurement: description of race / Hispanic origin domain. Technical report, DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-45, 2012.

Sundar Natarajan, Stuart R. Lipsitz, Garrett Fitzmaurice, Charity G. Moore, and Rene Gonin. Variance estimation in complex survey sampling for generalized linear models. *Appl. Statist.*, 57:75 – 87, 2008.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. N. K. Rao and M. Tausi. Estimating function jackknife variance estimators under stratified multistage sampling. *Comm. Statist. Theory Methods*, 33(9): 2087 – 2095, 2004.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89(427):846 – 866, 1994.

Andrea Rotnitzky and James Robins. Analysis of semi-parametric regression models with non-ignorable non-response. *Statist. Med.*, 16(1):81 – 102, 1997.

G. A. F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Edward Arnold, United Kingdom, second edition, 1982.

Daniel O. Sharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.*, 94(448):1096 – 1120, 1999.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics With S*. Springer, New York, fourth edition, 2002.