Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Robust Bayesian Predictive Inference for the Finite Population Quantile of a Small Area

Balgobin Nandram, and Jai Won Choi[1]

Abstract

We use a robust Bayesian method to analyze data with possibly *nonignorable* nonresponse and selection bias. A *robust* logistic regression model is used to relate the response indicators (Bernoulli random variable) to the covariates, which are available for everyone in the finite population. This relationship can adequately explain the difference between respondents and nonrespondents for the sample. This robust model is obtained by expanding the standard logistic regression model to a mixture of Student's t distributions, thereby providing propensity scores (selection probability) which are used to construct *adjustment* cells. The nonrespondents' values are filled in by drawing a random sample from a kernel density estimator, formed from the respondents' values within the adjustment cells. Prediction uses a linear spline rank-based regression of the response variable on the covariates by areas, sampling the errors from another kernel density estimator; thereby further robustifying our method. We use Markov chain Monte Carlo (MCMC) methods to fit our model. The posterior distribution of a quantile of the response variable is obtained within each sub-area using the order statistic over all the individuals (sampled and nonsampled). We compare our robust method with recent parametric methods

KEY WORDS: Logistic regression, Metropolis-Hastings sampler, Order statistics, Propensity scores, Rank-based method, Student's t distribution.

1. Introduction

The purpose of this work is to predict the percentile of Body Mass Index (BMI) for the finite population of children and adolescents, post-stratified by county for each domain formed by age, race and sex and to investigate what adjustment needs to be made for nonresponse and selection bias, using National Health and Nutrition Examination Survey (NHANES III) data. Nandram and Choi (2005, 2006) fit hierarchical Bayesian models to accommodate such a nonresponse mechanism. We seek more robust models, robust to distributional assumptions and outliers.

Greenlees, Reece and Zieschang (1982) developed a normal-logistic regression model, a nonignorable nonresponse model within the selection approach, to impute missing values in the Current Population Survey when the probability of response depends on the variable being imputed. Nandram and Choi (2005) extend this model to accommodate small domains for the NHANES III data. The main contribution in Nandram and Choi (2005) is a Bayesian predictive inference of the finite population mean using a spline regression model in which the logarithm of the BMI values are modeled. Nandram and Choi (2006) make four new contributions. First they make inference about the more appropriate finite population percentiles. Second, they show that the logarithmic transformation is the best in a selected set within the Box-Cox family. Third, they demonstrate small effects of clustering. Fourth, they show how to account for the selection probabilities.

Our key idea of this paper is to robustify the logistic regression model based on the relation between the logistic and the Student's t densities. The first four moments of a standard logistic density (i.e., zero location and unit scale) are the same as that of a Student's t density with location zero, scale $\gamma = \pi\sqrt{7/27}$ and nine degrees of freedom. That is, if x/γ has a standard Student's t density, then x has approximately a standard logistic density; see Mudholkar and George (1978) for the original discussion on this issue, and Albert and Chib (1993) for an application. This result

[1] Balgobin Nandram; Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280The opinions Jai Won Choi. Expressed in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics.

creates an avenue for our robust analysis on nonresponse. Let $I_i = 1$ if the i^{th} individual responds and $I_i = 0$ otherwise, for a sample of size n, and let $\underline{z}_i$ denote covariates. Then, the standard logistic regression model with $I_i \mid \underline{\beta} \overset{ind}{\sim} Bernoulli\{e^{\underline{z}_i'\underline{\beta}}/(1+e^{\underline{z}_i'\underline{\beta}})\}$, is approximately the same as $I_i \mid \underline{\beta} \overset{ind}{\sim} Bernoulli\{\mathrm{T}_{\eta}(\underline{z}_i'\underline{\beta}/\gamma)\}$, where $\mathrm{T}_{\eta}(\cdot)$ is the cumulative distribution function of the Student's t random variable on $\eta = 9$ degrees of freedom (i.e., the regression coefficients $\underline{\beta}$ in the two models are approximately the same). Thus, robustification is obtained by placing weights on different values of η with a substantial weight at $\eta = 9$ degrees of freedom; thereby forming a more flexible class of models.

In addition, we robustify our procedure in the prediction process itself. Instead of assuming normality, we use a linear rank-based procedure (Hettmansperger 1984) to fit a linear model of logarithm of BMI on the covariates, requiring no distribution assumptions; see Potvin and Roff (1993) for interesting comments about normality. It is clear that this new procedure is an important advance over the normal-logistic model of Greenlees, Reece and Zieschang (1982) and Nandram and Choi (2005, 2006).

The major difference with our previous work is that we use propensity scores to study "nonignorable" nonresponse. Our approach here is even different from our earlier non-parametric approach with Dirichlet process prior (see Nandram and Choi 2004). The propensity scores incorporate differences between respondents and nonrespondents through covariates; see Little and Rubin (2002) for a discussion on nonresponse and Rosenbaum (2002) for a similar discussion on observational studies. The propensity score simply describes the response probability of an individual as a function of the covariates. A standard reasonable assumption is that the response indicators follow a logistic regression model in which the response probability (propensity score) is a function of the covariates. The propensity scores are used to form adjustment cells, and the nonrespondents' BMI values are filled in by drawing samples from a kernel density estimator (Silverman 1986).

The notion of nonignorability is a somewhat different from that of Little and Rubin (2002). The estimated propensity scores are a function of the response indicators and the covariates, unrelated to the observed responses. But, the unobserved responses are a function of the estimated propensity scores and the covariates, and therefore, they are a function of the response indicators. However, the covariates may not discriminate between respondents and nonrespondents. Thus, our models are very flexible because they are allowed to capture some degree of nonignorability.

The purpose of this paper is to describe a robust method to obtain propensity scores to fill in nonrespondents and to predict the finite population percentiles (at risk of overweight; overweight) for small domains formed by age, race and sex. In Section 2 we discuss the the NHANES III data that we study. In Section 3, we review the previous hierarchical Bayesian model (Nandram and Choi 2006) for nonignorable nonresponse through the selection approach. In Section 4 we provide our robust Bayesian methodology. In Section 5 we present data analysis.

2. Main Features of the NHANES III Data

Nonresponse occurs in both the interview and examination stages of NHANES III survey, October 1988 through September 1994 (see National Center for Health Statistics, 1994 for detail). The interview nonresponse arises from sampled persons who did not participate in the interview. Some individuals, who were already interviewed and included to participate in the health examination, missed the examination at home or at the MEC, thereby missing all or part of the examination. In all our previous work, "nonresponse" refers to a missing BMI value for those sampled "persons" whose age, sex and race information was obtained. We note also that for children and adolescents (2-19 years old) the observed nonresponse rate is $100 \times (1606/6791) \approx 24$%.

We study the BMI data for four age classes (02-04, 05-09, 10-14, 15-19 years). Recalling that there are 560 ($35 \times 2 \times 2 \times 4$) domains, the sample sizes on the average are very small per domain e.g., $6791/560 \approx 12$ for the sample and $5185/560 = 9$ for the respondents. Many domains by county are too small (i.e., many domains do not have any individuals in the sample) for any meaningful analysis. Therefore, the small areas are formed by crossing age, race and sex within counties. As in Nandram and Choi (2005, 2006) our models are constructed at county level, and, at the same time, age, race and sex are represented as covariates; inference is made for each domain formed by crossing age, race and sex within county. There are differential probabilities of selection by age, race and sex, from a screening procedure, oversampling some age group and race.

3. The Bayesian Spline Regression Model: A Review

We give a review of the spline regression model of Nandram and Choi (2005), and describe how Nandram and Choi (2006) incorporate correlation and selection probabilities into the model.
There are data from $\ell = 35$ counties and each county has N_i (known) individuals. We assume a probability sample of n_i individuals is taken from the i^{th} county. Let s denote the set of sampled units and ns the set of nonsampled units. Let I_{ij}, $i = 1,2,..., \ell$ and $j = 1,2,..., N_i$, be the response indicator for the j^{th} individual within the i^{th} county in the population. Also, let x_{ij} denote the BMI value, possibly transformed (e.g., the logarithmic transformation). Note that r_{ij} and x_{ij} are all observed in the sample s; x_{ij} are unknown, and r_{ij} are not needed, in ns. Let $r_i = \sum_{j=1}^{n_i} r_{ij}$ (i.e., r_i is the number of sampled individuals that responded in the i^{th} county). For convenience, we express the BMI x_{ij} as $x_{i1}, x_{i2}, \ldots, x_{ir_i}, x_{ir_i+1}, \ldots, x_{in_i}$ in s, $x_{in_i+1}, \ldots, x_{iN_i}$ in ns, and $(N_i - n_i)$ non-sampled persons for the i^{th} county.

A key point that we note for what follows is that the r_i individuals are not necessarily random respondents from the n_i sampled individuals. This is the nonresponse bias we need to address. It is clear that we need to predict the BMI values x_{ij} for (a) the nonrespondents in s and (b) the individuals in ns. Thus, for the finite population of N_i individuals, we need a Bayesian predictive inference for the $100\eta, 0 < \eta < 1$, percentile of the finite population of BMI values for each age-race-sex domain within the i^{th} county. For example, for the i^{th} county, let $\underline{x}_i = (\underline{x}_i^{(s,r)}, \underline{x}_i^{(s,nr)}, \underline{x}_i^{(ns)})'$, where $\underline{x}_i^{(s,r)}$ is observed BMI values of the sampled respondents, and both $\underline{x}_i^{(s,nr)}$, the BMI values of the sampled nonrespondents, and $\underline{x}_i^{(ns)}$, the nonsampled BMI, are not observed. Then, the 100η percentile of the i^{th} county is the $[\eta N_i]^{th}$ order statistic ($[\cdot]$ is the nearest integer to ηN_i) among the N_i components of $\underline{x}_i$. Because only $\underline{x}_i^{(s,r)}$ is observed, Nandram and Choi (2005) develop a Bayesian selection and a Bayesian pattern mixture model to predict the finite population mean BMI for each domain. Note that only the $\underline{x}_i^{(s,r)}$ are observed; $\underline{x}_i^{(s,nr)}$ and $\underline{x}_i^{(ns)}$ are to be predicted. This is an enormous task for three reasons: The finite population is large, there are covariates, and transformation is used.

To accommodate the BMI values and the nonresponse indicators, Nandram and Choi (2006) used the selection nonresponse model which has two parts. For part 1 of the selection model, the response depends on the BMI as $I_{ij} \mid x_{ij}, \underline{\beta}_i \overset{ind}{:} Bernoulli\left\{ e^{\beta_{0i}+\beta_{1i}x_{ij}} / (1 + e^{\beta_{0i}+\beta_{1i}x_{ij}}) \right\}$, where $\underline{\beta}$ is bivariate normal with appropriate priors for the mean, variance, and correlation. For the part 2 of the selection model, the single most important predictor of BMI is age, with race and sex playing a relatively minor role, and there is also a need to understand the relationship between BMI and age, race and sex. For $i = 1, \ldots, \ell, j = 1, \ldots, N_i$, we let $z_{ij0} = 1$ for an intercept, $z_{ij1} = 1$ for non-black and $z_{ij1} = 0$ for black, $z_{ij2} = 1$ for male and $z_{ij2} = 0$ for female, $z_{ij3} =$ $z_{ij1} \times z_{ij2}$ for the interaction between race and sex, and we let $\underline{z}_{ij}' = (z_{ij0}, z_{ij1}, z_{ij2}, z_{ij3})$. Also, let a_{ij} denote the age of the j^{th} individual within the i^{th} county. Generically, letting $c^+ = 0$ if $c \le 0$ and $c^+ = c$ if $c > 0$,

$w_{ij1}=1, w_{ij2}=(a_{ij}-8)^{+}, w_{ij3}=(a_{ij}-13)^{+}$, for a spline regression of BMI on age adjusting for race and sex, we take $x_{ij}=\sum_{t=1}^{3}(\underline{z}_{ij}'\underline{\alpha}_t+\nu_{ti})w_{ijt}+e_{ij},\ e_{ij}\mid\sigma_3^2 \overset{iid}{:} Normal(0,\sigma_3^2)$ with proper priors and hyperpriors. See Appendix B (Nandram and Choi 2005) for the details of how to fit the selection model using Markov Monte Carlo methods.

Prediction of the finite population percentiles is straightforward. We first predict $\underline{x}^{(ns)}=(x_{ij}, j=n_i+1,\dots,N_i)$. Then, $f(\underline{x}^{(ns)}\mid\underline{x}^{(obs)})=\int\{\prod_{j=n_i+1}^{N_i}f(x_{ij}\mid\Omega)\}\,\pi(\Omega\mid\underline{x}^{(obs)})d\Omega$, where $x_{ij}\mid\Omega\colon Normal\{\sum_{t=1}^{3}(\underline{z}_{ij}'\underline{\alpha}_t+\nu_{ti})w_{ijt},\sigma_3^2\}$ and Ω is the set of all parameters including $\underline{x}_i^{(s,nr)}$. Thus, it is straightforward to predict $x_{in_i+1},\dots,x_{iN_i}$. First we take a sample of size M from the posterior distribution, $\{\Omega^{(h)}:h=1,\dots,M\}$, and second, we fill in $x_{i,n_i+1},\dots,x_{iN_i}$ using this. (Once the transformed BMI values are predicted, they can be easily retransformed to the original scale.) Letting $\underline{x}_i^{(h)}$ denote the vector of all N_i iterated values, we order these components to obtain the $[\eta N_i]^{th}$ value $x_{[\eta N_i]}^{(h)}$. Thus, we have a "random sample" $x_{[\eta N_i]}^{(h)}, h=1,\dots,M$, from the posterior density of the $[100\eta]$ percentile.

4. Robust Bayesian Methodology

Now, we describe the robust method to predict the finite population percentiles. First, we use robust logistic regression model to obtain propensity scores. Second, using these propensity scores to form adjustment cells, we fill in the nonrespondents. Third, we fit a linear spline rank-based, regression model of the logarithmic BMI values (respondents and nonrespondents) on the relevant covariates and use this model also to predict the nonsampled BMI values. We also include the selection probabilities. We will call the model described in this section the *mixture model* or *robustified logistic regression model*. Note that the logistic regression model with random effects is a special case of this model when the degrees of freedom is nine.

4.1 Robust Logistic Regression

We assume that

$$I_{ij}\mid\{\underline{\beta},\nu_i,\eta=a_r\}\overset{ind}{:}\ Bernoulli[\mathrm{T}_{a_r}\{(\underline{z}_{ij}'\underline{\beta}+\nu_i)/\gamma\}],\ j=1,\dots,n_i, \tag{1}$$

where $\mathrm{T}_{a_r}(\cdot)$ is the cumulative distribution function of the Student's t density on a_r degrees of freedom,

$$\nu_i\mid\sigma^2\overset{iid}{:}\ Normal(0,\sigma^2), i=1,\dots,\ell, \tag{2}$$

$$Pr(\eta=a_r)=\omega_r, r=1,\dots,R,\sum_{r=1}^{R}\omega_r=1, \tag{3}$$

where $\gamma=\pi\sqrt{7/27}$ and $\{(a_r,\omega_r), r=1,\dots,R\}$ is specified with R the number of values of η. [Note that it is convenient to use R here, and later it will be used to denote ranks; also η was used in Section 3 for percentiles.] In our application we take $a_1=3, a_2=6, a_3=9, a_4=12, a_5=25, a_6=50$. Here $a_1=3$ is near a Cauchy density, $a_3=9$ is the approximate logistic density and $a_6=50$ is near a normal density. The construction in (1), (2) and (3) is a prescription of R models, the r^{th} model having probability ω_r. Note that the ν_i are area effects, and they form a common stochastic process. Moreover, note that the regression coefficients are not allowed to change over the models; thus facilitating a borrowing of strength across the counties and this is moderated by the ν_i. Also, it is desirable to identify a single set of regression parameters as in (1); in this way we can avoid model averaging, an undesirable complex feature. We only allow the degree of freedom a_r to depend on the model.

Finally, a priori we assume that $\underline{\beta}$ and σ^2 are independent with

$$\underline{\beta}\text{: } Normal(\underline{\theta}_0, \Delta_0) and\ p(\sigma^2) = 1/(1+\sigma^2)^2,\ \sigma^2 > 0, \tag{4}$$

where $\underline{\theta}_0$ and Δ_0 are specified, and we avoid the ill-behaved prior distribution σ^{-2}: $Gamma(.001, .001)$ (see Gelman 2006). We specify $\underline{\theta}_0$ and Δ_0 by performing a standard logistic regression of the response indicators on the covariates, taking $\underline{\theta}_0$ as the point estimator of $\underline{\beta}$ and Δ_0 as the covariance matrix inflated one hundred times.

Note that setting $\omega_1 = 1, a_1 = 9$ gives the logistic regression model, and letting σ^2 go to zero, in the limit the ν_i become point masses at 0 (i.e., there will be no county effects). Doing both simultaneously produces the logistic regression model without county effects. We will call the model specified by (1)-(4) the robustified logistic regression model, and it is the basis for robust prediction of the finite population percentiles.

We fit the model using the Metropolis-Hastings sampler (e.g., see Appendix B of Nandram and Choi 2006 for the selection model). We obtain a sample $(\underline{\beta}^{(h)}, \eta^{(h)}, \underline{\nu}^{(h)}), h = 1, \ldots, M = 1000$ which we use for inference about the finite population percentiles. The joint posterior density is $p(\underline{\beta}, \underline{\nu}, \sigma^2 \mid \underline{I})$. To run the Metropolis-Hastings sampler, we need the conditional posterior densities for priors, $p(\underline{\beta} \mid \underline{\nu}, \sigma^2, \underline{I})$, $p(\underline{\nu}_i \mid \underline{\beta}, \underline{\nu}_{(i)}, \sigma^2, \underline{I})$, and $p(\sigma^2 \mid \underline{\beta}, \underline{\nu}, \underline{I})$ (see Nandram and Choi, 2006).

We need to predict the nonrespondents' BMI values and the BMI values of the nonsample population. The methods to fill in the missing values for the nonrespondents and the nonsample population are both robust. First, we consider the nonrespondents. At the h^{th} iterate, the propensity scores (selection probability) are $T_{\eta^{(h)}}\{(\underline{z}_{ij}\underline{\beta}^{(h)} + \nu_i^{(h)})/\gamma\}, j = 1, \ldots, n_i, i = 1, \ldots, \ell$, where $\eta^{(h)}$ is one of the $a_r, r = 1, \ldots, R$. For the h^{th} iterate, we have partitioned all propensity scores (for respondents and nonrespondents in all counties) into five strata (adjustment cells). We have computed the five quintiles (.2, .4, .6, .8 to form the five strata) of the propensity scores, and we allocated both the r_i respondents and the $(n_i - r_i)$ nonrespondents to these five strata. So we now know which stratum each individual (respondent or nonrespondent) belongs, albeit with uncertainty.

For each stratum, we construct a kernel density estimator of the n observed values. Then, letting L_s denote the number of respondents in the s^{th} adjustment cell, and order the logarithm of the BMI values for the respondents within the s^{th} adjustment cell as $\tilde{x}_{sj}, j = 1, \ldots, L_s$. Then, to estimate the density of the $\tilde{x}_{sj}$, we use the kernel density estimator,

$$g(\tilde{x}) = \sum_{j=1}^{L_s} \frac{1}{L_s} \frac{1}{h_{opt}} \phi\left(\frac{\tilde{x} - \tilde{x}_{sj}}{h_{opt}}\right),$$

where $\phi(\cdot)$ is the standard normal density and $h_{opt} = \frac{1.06}{L_s^{1/5}} min(STD, IQR/1.34)$ with STD and IQR respectively the standard deviation and interquartile range of the $\tilde{x}_{sj}, j = 1, \ldots, L_s$, s=1,...,5; see Silverman (1986, pg. 47).

Then, to fill in a BMI value for a nonrespondent, we draw a random value from $g(\tilde{x})$. Specifically, we draw random a sample of size one from the labels $j = 1, \ldots, L_s$, say j', and then draw the logarithm of the BMI value from the normal distribution with mean $\tilde{x}_{sj'}$ and standard deviation h_{opt}. We obtain the BMI value x_{sj} by retransforming $\tilde{x}_{sj}$ (i.e., $x_{sj} = exp(\tilde{x}_{sj})$. The entire process is repeated independently for all the nonrespondents in each of the five adjustment cells. At the h^{th} iterate, we have done so for all nonrespondents in all strata. Note that this process includes uncertainty about the formation of the strata as well because the adjustment cells are formed at each iterate.

Second, a similar procedure, but without using the adjustment cells, can be applied to the BMI values of the nonsampled individuals, which are needed to obtain the finite population percentiles. Specifically, we use a linear spline rank-based estimation procedure. In passing, we note that a less robust alternative is the least squares procedure. However, the least sqaures procedure has less computational burden, and we have found small differences between the linear spline rank-based and the least squares procedures; we prefer the linear spline rank-based procedure.

We use a notation similar to Part 2 of the selection model for the spline regression in Section 3. For $i = 1, \ldots, \ell, j = 1, \ldots, N_i$, we let $z_{ij0} = 1$ for an intercept, $z_{ij3} = 1$ for non-black and $z_{ij3} = 0$ for black, $z_{ij4} = 1$ for male and $z_{ij4} = 0$ for female, $z_{ij5} =$ $z_{ij3} \times z_{ij4}$ for the interaction between race and sex. Here $z_{ij1} = (a_{ij} - 8)^+$ and $z_{ij2} = (a_{ij} - 13)^+$ with again $c^+ = 0$ if $c \le 0$ and $c^+ = c$ if $c > 0$. Note the small change in the definition of $\underline{z}'_{ij} = (z_{ij0}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}, z_{ij5},)$, with the second and third components for the two nonconstant splines in age.

We specify the following model for all individuals in the finite population (i.e., sample respondents and nonrespondents and nonsample individuals) At the h^{th} iterate letting $\tilde{x}_{ij}$ denote the logarithm of the BMI values, we use the model

$$\tilde{x}_{ij}^{(h)} = \underline{z}'_{ij} \underline{\tilde{\beta}}_i + e_{ij}, i = 1, \ldots, \ell, j = 1, \ldots, N_i,$$

where e_{ij} are independent and identically distributed. Note that $\tilde{x}_{ij}^{(h)}, j = 1, \ldots, r_i$ do not change with h because they are observed, and the unobserved $\tilde{x}_{ij}^{(h)}, j = r_i + 1, \ldots, n_i$ for the nonrespondents in the sample are obtained using the kernel density estimator as described for the nonrespondents, without any specific distribution assumptions. That is, both the BMI values of the respondents and the nonrespondents are used to estimate the $\underline{\tilde{\beta}}_i$. Also, note that the parameters of this model are not related to those in the logistic regression model.

We obtain the rank-based estimators $\underline{\tilde{\beta}}_i^{(h)}$ in Appendix B (Nandram and Choi, 2006). Then, using the rank-based estimators, we obtain the residuals $\tilde{e}_{ij}^{(h)} = \tilde{x}_{ij}^{(h)} - \underline{z}'_{ij} \underline{\tilde{\beta}}_i^{(h)}, i = 1, \ldots, \ell, j = 1, \ldots, n_i$. In a manner similar to the nonrespondents, within the i^{th} county and at the h^{th} iterate, we have constructed a kernel density estimator using the $\tilde{e}_{ij}^{(h)}, j = 1, \ldots, n_i$. Thus, to fill in the BMI values for the nonsampled population, we draw a random sample of size $N_i - n_i$ using the kernel density estimator, formed by the $\tilde{e}_{ij}^{(h)}$ from the i^{th} county, $i = 1, \ldots, \ell$. Then, $x_{ij}^{(h)} = \exp\{\tilde{x}_{ij}^{(h)}\}$, where $\tilde{x}_{ij}^{(h)} = \tilde{e}_{ij}^{(h)} + \underline{z}'_{ij} \underline{\tilde{\beta}}_i^{(h)}, i = 1, \ldots, \ell, j = n_i + 1, \ldots, N_i$, are obtained.

Thus, conditional on the observed BMI values, we have obtained the BMI values of the nonrespondents and the nonsampled individuals. Therefore, the finite population quantiles are now easy to obtain as in the spline regression model; see Section 3.

Finally, we have also included the selection probabilities in this robust framework. We note the inclusion of the selection probabilities is a challenging problem, because there is no simple linear relation between the selection probabilities and the BMI values, but a preliminary investigation shows that there is important information about BMI in these selection probabilities. See Appendix B (Nandram and Choi, 2006) on how to include the selection probabilities.

5. Data Analysis

In this section, we present an analysis of the BMI data using five models. The first two models are the selection nonresponse model with spline regression (1) (Nandram and Choi 2005) and the selection nonresponse model with the spline regression and selection probabilities (2) (Nandram and Choi 2006). These first models are parametric models. The other three models are the logistic regression model (3), robustified logistic regression model (4) and robustified logistic regression model with selection probabilities (5). These are nonparametric models in the sense that there are no distribution assumptions on the BMI values of the nonrespondents and the nonsample population. For inference, we have presented posterior mean (PM), posterior standard deviation (PSD), numerical standard deviation (NSE), and 95% credible interval. In Tables 1 and 2 we have studied inference of the regression coefficients, and in Table 3 we have predicted the 85^{th} and 95^{th} percentile BMI values by model and age for white males. In Table 1 we have studied how the regression coefficients of the robustified logistic regression model (4) change with the weights, but not the selection probabilities. As is evident, the changes with the weights are small. These differences in weights do not affect inference markedly. Thus, we have used for all our calculations $\omega_3 = .5$ and $\omega_k = (1-\omega_3)/5, k \neq 3$.

We have compared the regression coefficients from the logistic regression model (3), the robustified logistic regression model (4), and the robustified logistic regression model (5) with selection probabilities. In Table 2 there are very little differences between these three models when one considers the 95% credible intervals. The posterior quantities should be exactly the same for the robustified logistic regression model with and without selection probabilities; the small differences are due to Monte Carlo errors. The differences between logistic regression (3) and the robustified logistic models (4, 5) with and without selection probabilities should be larger than the differences between these two latter models (4, 5). More importantly, age and race are discriminators of respondents and nonrespondents. For example, under the robust model (5) with selection probabilities the 95% credible intervals for the regression coefficient of age and race are $(0.932, 1.054)$ and $(-0.305, -0.201)$.

Next, we discuss our results for prediction of the finite population 85^{th} and 95^{th} percentiles. While there are important differences among the counties (Nandram and Choi 2006), we exemplify our results using county 11. However, we now compare all five models to see the effects of robustification and inclusion of the selection probabilities. These results are presented for white males in Table 3. We discuss in detail the predictions for white males in Table 3. We summarize the similarities and differences between the two parametric models (first two) and the three nonparametric models (last three). We note three similarities. [a.] The posterior means of the 85^{th} percentile and the 95^{th} percentile are respectively very similar for all five models, and the expected increase from age group 1 to 4 is similar. As it must be, the posterior means of the 95^{th} percentile are larger than the posterior means of the 85^{th} percentile. [b.] The selection probabilities do two things. Both the posterior means and the posterior standard deviations are smaller under the models with selection probabilities than under the models without the selection probabilities. That is, the selection probabilities correct for some selection bias and help to improve precision. [c.] As is expected, the posterior standard deviations for both the 85^{th} and 95^{th} percentiles are generally larger under the robust models than under the parametric models; there are some exceptions. Under the parametric models the largest posterior standard deviations occur at age group 4; whereas the largest posterior standard deviations occur at age group 3 under all robust models except for the robustified logistic regression model when the 95^{th} percentile is being estimated.

References

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, **88**, 669-679.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D.(1982), "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, **77**, 251-261.

Hettmansperger, T. P. (1984), *Statistical Inference Based on ranks*. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*. New York: Wiley.

Mudholkar, G. S. and George, E. O. (1978), "A remark on the shape of the logistic distribution," *Biometrika*, **65**, 667-668.

Nandram, B. and Choi, J.W. (2006), "A Bayesian Assessment of Obesity Among Children and Adolescents from Small Domains Under Nonignorable Nonresponse," *Technical Report*, Worcester Polytechnic Institute.

Nandram, B. and Choi, J.W. (2005), "Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES Data," *Survey Methodology*, **31**, 73-84.

Nandram, B. and Choi, J. W. (2004), "A Nonparametric Bayesian Analysis of a Proportion for a Small Area Under Nonignorable Nonresponse," *Journal of Nonparametric Statistics*, **16**, 821-839.

National Center for Health Statistics (1994), "Plan and Operation of the Third National Health and Nutrition Examination Survey," *Vital and Health Statistics Series 1*, **32**.

Potvin, C. and Roff, D. A. (1993), "Distribution-free and robust statistical methods: Viable alternatives to parametric statistics?" *Ecology*, **74**, 1617-1628.

Rosenbaum, P. R. (2002), *Observational Studies*. New York: Springer-Verlag.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Table 1: Sensitivity of the posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals for the regression coefficients (coef) of the robustified logistic model (4) without selection probabilities with respect to the choice of weights

ω_3	Coef	PM	PSD	NSE	Interval
.3	β_0	1.445	0.034	0.002	(1.378, 1.509)
	β_1	0.990	0.031	0.001	(0.931, 1.053)
	β_2	-0.252	0.025	0.001	(-0.300, -0.201)
	β_3	-0.007	0.024	0.001	(-0.055, 0.039)
	β_4	-0.014	0.026	0.001	(-0.063, 0.040)
.5	β_0	1.448	0.033	0.002	(1.387, 1.515)
	β_1	0.993	0.030	0.001	(0.931, 1.052)
	β_2	-0.253	0.027	0.001	(-0.303, -0.201)
	β_3	-0.007	0.025	0.001	(-0.059, 0.040)
	β_4	-0.015	0.026	0.001	(-0.064, 0.034)
.7	β_0	1.448	0.037	0.006	(1.399, 1.524)
	β_1	1.000	0.030	0.003	(0.943, 1.067)
	β_2	-0.246	0.025	0.003	(-0.311, -0.211)
	β_3	-0.004	0.025	0.002	(-0.054, 0.042)
	β_4	-0.024	0.022	0.002	(-0.065, 0.020)
.9	β_0	1.454	0.035	0.004	(1.401, 1.521)
	β_1	1.005	0.030	0.003	(0.956, 1.065)
	β_2	-0.254	0.028	0.003	(-0.313, -0.207)
	β_3	-0.008	0.026	0.002	(-0.061, 0.043)
	β_4	-0.017	0.024	0.002	(-0.056, 0.035)

NOTE. The logistic regression coefficients are intercept (β_0), age (β_1), race (β_2), sex (β_3) and the race-sex interaction (β_4). The weights in the robustified logistic regression model (4) without selection probabilities and with county effects are $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6)$; $\omega_k = (1-\omega_3)/5, k \neq 3$ and ω_3 is specified.

Table 2: Posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals for the regression coefficients (coef) by model

Model	Coef	PM	PSD	NSE	Interval
3	β_0	1.474	0.033	0.003	(1.413, 1.536)
	β_1	1.014	0.031	0.002	(0.952, 1.070)
	β_2	-0.260	0.029	0.002	(-0.316, -0.204)
	β_3	-0.006	0.027	0.002	(-0.060, 0.040)
	β_4	-0.017	0.024	0.002	(-0.055, 0.037)
4	β_0	1.448	0.033	0.001	(1.387, 1.515)
	β_1	0.993	0.030	0.001	(0.931, 1.052)
	β_2	-0.253	0.027	0.001	(-0.303, -0.201)
	β_3	-0.007	0.025	0.001	(-0.059, 0.040)
	β_4	-0.015	0.026	0.001	(-0.064, 0.034)
5	β_0	1.452	0.032	0.001	(1.389, 1.518)
	β_1	0.995	0.031	0.001	(0.932, 1.054)
	β_2	-0.252	0.027	0.001	(-0.305, -0.201)
	β_3	-0.007	0.026	0.001	(-0.059, 0.041)
	β_4	-0.015	0.026	0.001	(-0.064, 0.036)

NOTE. The regression coefficients are intercept (β_0), age (β_1), race (β_2), sex (β_3) and the race-sex interaction (β_4). Sex and the race-sex interaction are not important to discriminate between respondents and nonrespondents. Model - 1 (spline regression model); 2 (spline regression model with selection probabilities); 3 (logistic regression); 4 (robustified logistic regression); 5 (robustified logistic regression with selection probabilities). For the robustified logistic regression models (4, 5) $\omega_3 = .5$.The regression coefficients for Models 1 and 2 are not comparable to Models 3-5; see Nandram and Choi (2005).

Table 3: Posterior means (PM), posterior standard deviations (PSD) and 95% credible intervals for the 85^{th} and 95^{th} percentiles of BMI for *white males* from county 11 by model and age

		85th Percentile			95th Percentile		
Model	Age	PM	PSD	Interval	PM	PSD	Interval
1	1	18.0	0.23	(17.5, 18.4)	19.7	0.25	(19.2, 20.2)
	2	18.2	0.23	(17.8, 18.7)	19.9	0.25	(19.4, 20.4)
	3	22.6	0.43	(21.7, 23.4)	24.8	0.48	(23.8, 25.7)
	4	25.1	0.84	(23.4, 26.7)	27.5	0.92	(25.6, 29.3)
2	1	18.1	0.15	(17.8, 18.3)	19.4	0.16	(19.1, 19.8)
	2	18.2	0.15	(17.9, 18.5)	19.7	0.17	(19.3, 20.0)
	3	22.1	0.39	(21.3, 22.8)	23.8	0.42	(22.9, 24.6)
	4	24.3	0.78	(22.7, 26.0)	26.2	0.85	(24.5, 28.0)
3	1	17.9	0.35	(17.2, 18.6)	20.2	0.61	(19.2, 21.5)
	2	18.0	0.37	(17.3, 18.7)	20.4	0.64	(19.4, 21.8)
	3	20.8	0.70	(19.5, 22.3)	23.7	0.92	(22.1, 25.6)
	4	25.5	0.41	(24.8, 26.4)	29.0	0.76	(27.7, 30.6)
4	1	17.9	0.35	(17.2, 18.6)	20.3	0.62	(19.3, 21.6)
	2	18.0	0.35	(17.4, 18.8)	20.5	0.62	(19.4, 21.9)
	3	20.9	0.69	(19.7, 22.3)	23.7	0.95	(22.2, 25.9)
	4	25.6	0.40	(24.8, 26.4)	29.0	0.81	(27.7, 30.9)
5	1	17.1	0.22	(16.8, 17.6)	19.4	0.44	(18.6, 20.4)
	2	17.3	0.23	(16.9, 17.7)	19.5	0.45	(18.8, 20.5)
	3	20.1	0.39	(19.4, 20.9)	22.7	0.63	(21.7, 24.1)
	4	25.3	0.35	(24.7, 26.0)	28.6	0.74	(27.4, 30.3)

NOTE. Age: 1 - (2-4), 2 - (5-9), 3 - (10-14), and 4 - (15-19); Model: 1 - (spline regression model), 2 - (spline regression model with selection probabilities), 3 - (logistic regression model), 4 - (robustified logistic regression model), 5 - (robustified logistic regression with selection probabilities).