RESEARCH REPORT SERIES
(Statistics #2011-03)

A partition model for analyzing categorical data subject to non-ignorable non-response

Ryan Janicki
Donald Malec

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

Report Issued: March 17, 2011

A partition model for analyzing categorical data subject to non-ignorable non-response*

Ryan Janicki and Donald Malec
Center for Statistical Research and Methodology, U. S. Census Bureau

Abstract

In many surveys, the goal is to estimate the proportion of the population within different domains with a certain characteristic of interest. This estimation problem is often complicated by survey non-response and the difficulty in modeling the non-response mechanism. In this paper we develop a new method for analyzing categorical data with non-response when there is uncertainty about ignorability, which incorporates the idea that there are many a priori plausible ignorable and non-ignorable models. We consider saturated submodels of the full model, which may have a mixture of ignorable and non-ignorable components, and use Bayesian averaging to incorporate model uncertainty. This method is illustrated using data from the 2000 Accuracy and Coverage Evaluation Survey. A simulation study is used to evaluate the performance of the model and to compare the partition model to other popular non-ignorable Bayesian models.

1 Introduction

The goal of the decennial census is to count every housing unit and individual in the country once and in the correct place. Unfortunately, this goal cannot be completely achieved since some housing units and individuals are not counted, counted multiple times, counted in the wrong place, or it may be that a record does not correspond to a person. If a person falls into this first category, he is a census omission. If a census enumeration falls into one of the remaining categories, we say that the record is an erroneous enumeration (EE). Otherwise the record is considered to be a correct enumeration (CE).

To estimate the probability that a census record is a correct enumeration, a sample of census records is taken and follow-up interviews are conducted. For the majority of the sampled records there is sufficient information to determine if the enumeration is correct or erroneous. However, there is a portion of the sample which cannot be resolved. If the unresolved enumerations are very different than the resolved enumerations, the estimate of the probability that an enumeration is correct could be inaccurate. See Hogan (1992) or Chen et al. (2010) for further discussion on estimating enumeration error in the U. S. census.

The previous discussion gives an example of incomplete data where the unresolved census enumerations are missing. In this paper we consider the general problem of analyzing a data set consisting of a binary response variables Y_{ij}, which may be missing for certain records, that have been post-stratified based on a vector of categorical covariates $\mathbf{X}_{ij}$, which are observed for all records. R_{ij} is an indicator variable which is 1 if a record is resolved and 0 otherwise. Here, $i = 1, \ldots, I$ indexes the post-strata and $j = 1, \ldots, m_i$ indexes the records in the ith post-stratum. In our earlier example, $Y_{ij} = 1$ if a record is a CE and $Y_{ij} = 0$ if a record is an EE, $\mathbf{X}_{ij}$ is a vector of categorical variables describing the characteristics of a record, and R_{ij} is 1 or 0 if a record is resolved or unresolved, respectively. We do not observe Y_{ij} if $R_{ij} = 0$. Let $Y_i = \sum_{j=1}^{m_i} Y_{ij} R_{ij}$ be the total number of resolved correct enumerations in cell i and $R_i = \sum_{j=1}^{m_i} R_{ij}$ be the total number of resolved enumerations in cell i. A typical presentation of this type of data is given in Table 1.

*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 authors and not necessarily those of the U.S. Census Bureau. The authors would like to thank Richard Griffin and Yves Thibaudeau for their careful review of this paper and their helpful comments.

Table 1: Example

	Resolved		Unresolved		
	CE	EE	CE	EE	Total
Cell 1	Y_1	$R_1 - Y_1$	?	?	$n_1 - R_1$
Cell 2	Y_2	$R_2 - Y_2$	?	?	$n_2 - R_2$
...	...	...	...	...	...
Cell I	Y_I	$R_I - Y_I$	?	?	$n_I - R_I$

Let

$$\begin{aligned} p_i &= P\left(\text{CE} \mid \text{Cell} = i\right), \\ \pi_{i1} &= P\left(\text{Resolved} \mid \text{CE}, \text{Cell} = i\right), \\ \pi_{i0} &= P\left(\text{Resolved} \mid \text{EE}, \text{Cell} = i\right), \end{aligned}$$

$i = 1, \ldots, I$. If $\pi_{i1} = \pi_{i0}$, cell i is said to be *ignorable*. Said another way, using the terminology of Little and Rubin (2002), the missing data in cell i is *missing at random* (MAR), since the missing data mechanism depends only on which cell the observation belongs to, and not on the characteristic of interest. If $\pi_{i0} \neq \pi_{i1}$, cell i is said to be *non-ignorable*, or *not missing at random* (NMAR).

Our goal is to estimate p_i, π_{i1}, and π_{i0} for each $i = 1, \ldots, I$ and to impute values for the unresolved cell counts. Conceptually, the imputation of resolved status is only needed within sampled clusters or segments, which are completely enumerated (with respect to their frame) and do not require sample weights. However, when estimating the parameters of an underlying model, the sampled areas may need to be combined, and the sample design will need to be accounted for. Within a sampled area, it is usually assumed that the likelihood has the form

$$\prod_{i=1}^{I} \left\{ \binom{n_i}{R_i} \binom{R_i}{Y_i} \left(p_i \pi_{i1}\right)^{Y_i} \left((1-p_i)\pi_{i0}\right)^{R_i - Y_i} \left(p_i(1-\pi_{i1}) + (1-p_i)(1-\pi_{i0})\right)^{n_i - R_i} \right\}. \tag{1}$$

Note that there are $3 * I$ free parameters and $2 * I$ data points so that the parameters are not identifiable without further assumptions.

There is clearly a great difficulty in estimating the proportion of the population with the characteristic of interest when the sample contains missing data since there is so little information about $P\left(Y \mid R = 0\right)$ contained in the data. Several approaches to modeling categorical data with non-ignorable non-response have been proposed. Baker and Laird (1988) showed that certain non-ignorable log-linear models are identifiable and gave an EM algorithm for estimating the parameters. However, they noted that it can often be the case that the EM algorithm converges to a boundary solution giving unrealistic estimates. This phenomenon occurs when fitting the identifiable non-ignorable log-linear model (XY, YR) to the data set presented in Table 2 in Section 3.1; the estimated probabilities are $\widehat{\pi}_{i1} = 1$ for all i. The notation (XY, YR) means that there is dependence between the variables X and Y and the variables Y and R. For a two way contingency table with one supplemental margin, this is the only non-ignorable log-linear model with estimable parameters (Little and Rubin, 2002). Park and Brown (1994) gave a geometric argument for why boundary solutions can occur. They used a prior distribution on the model parameters and the EM algorithm to obtain parameter estimates that are pushed away from the boundary.

Stasny (1991) described an empirical Bayes method which makes use of hierarchical models and small area techniques to estimate a proportion, allowing for non-response. Nandram and Choi (2002) extended the work of Stasny (1991) by using a full Bayesian model and "centering" a non-ignorable model around an ignorable model. Nandram and Choi (2002) introduce parameters which control the extent of non-ignorability of the model. Forster and Smith (1998) constructed log-linear models that are also centered around an ignorable model and introduced parameters which control the extent of non-ignorability when there is uncertainty about the degree of non-ignorability.

Some authors have questioned the value of attempting to model the missingness mechanism for NMAR models. Rubin et al. (1995) argue that ignorable models can be quite accurate and much simpler than non-ignorable models. They believe that follow-up information tends to support the utility of the ignorable assumption so long as non-response is limited and good covariate information is available. Tsiatis (2006) does not believe that NMAR models are useful since the correctness of the model cannot be verified using the observed data. Molenberghs et al. (2001) argue that non-ignorable models are often simpler than ignorable models, can provide realistic results, and can therefore be used as a good starting point. They believe that since an ignorable model can usually be formulated as a special member of a general family of non-ignorable models, the concept of fitting a single model should be replaced by a sensitivity analysis, where several plausible non-ignorable models are contrasted.

It is certainly true that ignorable methods are usually simpler than non-ignorable methods and that it is impossible to verify the correctness of a non-ignorable model. However, it is also impossible to verify the correctness of an ignorable model, and an ignorable model may be undesirable or unrealistic since one can never completely exclude the possibility that the data generation process is governed by a NMAR model. For example, in the earlier census enumeration problem, intuition tells us that it should be easier to resolve a correct enumeration than an erroneous enumeration. In fact, it was shown in Molenberghs et al. (2008) that there does not exist an omnibus test of MAR versus NMAR. Hence any model depends on non-verifiable modeling assumptions.

In this paper we describe a novel method of estimation when there is uncertainty about ignorability, which can be used as a part of a sensitivity analysis. Our method of estimation tries to incorporate the idea that there are many a priori plausible ignorable and non-ignorable models. We consider both ignorable and non-ignorable parametric submodels of the likelihood given in equation (1) that use the data as fully as possible. We then average the estimates obtained from these submodels by considering which submodels are a posteriori most likely. Section 2 shows how to construct this partition model and how to obtain parameter estimates by averaging over a collection of saturated submodels. In Section 3, we analyze a data set from the 2000 Accuracy and Coverage Evaluation Survey (A.C.E.) using this partition model and compare the results to commonly used non-ignorable models. In Section 4 we conduct a simulation study to compare the different models we have considered. Concluding remarks are made in Section 5.

2 A non-ignorable partition model

Due to the non-identifiability of the full model in equation (1), there is not enough information in the observed data to estimate the parameters of interest. For this reason it makes sense to consider submodels which use the data as fully as possible, that is, to consider saturated submodels.

If each cell is ignorable then we have a saturated model since there are two free parameters for each cell, p_i and $\pi_i = \pi_{i1} = \pi_{i0}$, $i = 1, \ldots, I$, which can be estimated using, for example, maximum likelihood estimates. Another model assumption that can be made is that parameters are shared across cells. For example, if we believe there is similarity between cell i and cell j, then we could assume that $\pi_{i1} = \pi_{j1} = \pi_1$ and $\pi_{i0} = \pi_{j0} = \pi_0$. We then have a total of four free parameters in cells i and j, π_1, π_0, p_i, and p_j, hence an estimable, saturated model. This assumption expresses the belief that the missingness mechanism is the same for each of the two cells. More generally, we can obtain an identifiable, saturated model, if we know a priori that a cell is either ignorable, or that it shares parameters with exactly one other cell.

Our goal is to find a method which takes into account the fact that there are many a priori plausible saturated models and to avoid unrealistic boundary estimates. This naturally leads to the idea of partition models considered in, for example, Hartigan (1990), Malec and Sedransk (1992), and Consonni and Veronese (1995). Following Malec and Sedransk (1992), we say that a partition of a set $\mathcal{I} = \{1, 2, \ldots, I\}$ is a collection $M = \{M_1, M_2, \ldots, M_d\}$ of subsets of $\mathcal{I}$, which we call groups, where the M_j are disjoint with union $\mathcal{I}$. Let $M^I = \{m \in M : |m| = 1\}$ be the collection of singletons in M, and $M^{NI} = \{m \in M : |m| = 2\}$ be the collection of paired cells in M. A general saturated model can be written $M_S = \left\{M^{NI}, M^I\right\}$ consisting of n pairs of non-ignorable cells in $M^{NI} = \{(i_1, i_2), \ldots, (i_{2n-1}, i_{2n})\}$ and $I - 2n$ ignorable cells in $M^I = \{i_{2n+1}, \ldots, i_I\}$.

Starting from the full likelihood given in equation (1), we can write the likelihood for a saturated sub-

model, given the parameters $\boldsymbol{p}$, $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_0$, and a model partition M_S, as

$$f\left(\mathbf{R},\mathbf{Y}\mid \boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M_S\right)=f_{NI}\left(\mathbf{R},\mathbf{Y}\mid \boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M^{NI}\right)\times f_I\left(\mathbf{R},\mathbf{Y}\mid \boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M^{I}\right)$$

where

$$\begin{aligned}f_{NI}\left(\mathbf{R},\mathbf{Y}\mid \boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M^{NI}\right)=\prod_{j=1}^{n}\Bigg\{&\binom{n_{i_{2j-1}}}{R_{i_{2j-1}}}\binom{R_{i_{2j-1}}}{Y_{i_{2j-1}}}\binom{n_{i_{2j}}}{R_{i_{2j}}}\binom{R_{i_{2j}}}{Y_{i_{2j}}}p_{i_{2j-1}}^{Y_{i_{2j-1}}}\left(1-p_{i_{2j-1}}\right)^{R_{i_{2j-1}}-Y_{i_{2j-1}}}\\
&\times p_{i_{2j}}^{Y_{i_{2j}}}\left(1-p_{i_{2j}}\right)^{Y_{i_{2j}}-Z_{i_{2j}}}\pi_{j1}^{Y_{i_{2j-1}}+Y_{i_{2j}}}\pi_{j0}^{R_{i_{2j-1}}-Y_{i_{2j-1}}+R_{i_{2j}}-Y_{i_{2j}}}\\
&\times\left[p_{i_{2j-1}}\left(1-\pi_{j1}\right)+\left(1-p_{i_{2j-1}}\right)\left(1-\pi_{j0}\right)\right]^{n_{i_{2j-1}}-R_{i_{2j-1}}}\\
&\times\left[p_{i_{2j}}\left(1-\pi_{j1}\right)+\left(1-p_{i_{2j}}\right)\left(1-\pi_{j0}\right)\right]^{n_{i_{2j}}-R_{i_{2j}}}\Bigg\}\end{aligned}$$

and

$$f_I\left(\mathbf{R},\mathbf{Y}\mid \boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M^{I}\right)=\prod_{j=2n+1}^{I}\left\{\binom{n_{i_j}}{R_{i_j}}\binom{R_{i_j}}{Y_{i_j}}p_{i_j}^{Y_{i_j}}\left(1-p_{i_j}\right)^{R_{i_j}-Y_{i_j}}\pi_{i_j}^{R_{i_j}}\left(1-\pi_{i_j}\right)^{n_{i_j}-R_{i_j}}\right\}.$$

If we knew with certainty that the partition M_S was correct, we could use maximum likelihood estimates for the unknown parameters. However, since there may be many partitions under consideration and significant uncertainty as to which partition is correct, we instead use a Bayesian procedure for estimation. Given a partition M_S, let π_{j1} and π_{j0}, $j=1,\ldots,n$, be the parameters that are shared across cells and π_{i_j}, $j=2n+1,\ldots,I$, the parameters from the ignorable cells. For our prior specification we let $\pi_{j1}\overset{i.i.d.}{\sim}\text{Beta}\left(\alpha_1,\beta_1\right)$, $\pi_{j0}\overset{i.i.d.}{\sim}\text{Beta}\left(\alpha_0,\beta_0\right)$, $p_{i_j}\overset{i.i.d.}{\sim}\text{Beta}\left(a,b\right)$, and $\pi_{i_j}\overset{i.i.d.}{\sim}\text{Beta}\left(\alpha,\beta\right)$. This specification implies that the parameters associated to the cells M^I (or M^{NI}) are exchangeable. Then the prior distribution of the parameters given a partition M_S is

$$\begin{aligned}\pi\left(\boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0\mid M_S\right)=&\prod_{j=1}^{n}\left\{\frac{1}{B\left(\alpha_1,\beta_1\right)}\pi_{j1}^{\alpha_1-1}\left(1-\pi_{j1}\right)^{\beta_1-1}\times\frac{1}{B\left(\alpha_0,\beta_0\right)}\pi_{j0}^{\alpha_0-1}\left(1-\pi_{j0}\right)^{\beta_0-1}\right\}\\
&\times\prod_{j=2n+1}^{I}\left\{\frac{1}{B\left(a,b\right)}p_{i_j}^{a-1}\left(1-p_{i_j}\right)^{b-1}\times\frac{1}{B\left(\alpha,\beta\right)}\pi_{i_j}^{\alpha-1}\left(1-\pi_{i_j}\right)^{\beta-1}\right\},\end{aligned}$$

where $B\left(x,y\right)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$ and $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$ is the standard gamma function. The posterior distribution of the parameters given the data and a model is

$$\pi\left(\boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0\mid\mathbf{R},\mathbf{Y},M_S\right)=\frac{f\left(\mathbf{R},\mathbf{Y}\mid\boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M_S\right)\pi\left(\boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0\mid M_S\right)}{\iiint f\left(\mathbf{R},\mathbf{Y}\mid\boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0,M_S\right)\pi\left(\boldsymbol{p},\boldsymbol{\pi}_1,\boldsymbol{\pi}_0\mid M_S\right)d\boldsymbol{p}\,d\boldsymbol{\pi}_1 d\boldsymbol{\pi}_0}.\tag{2}$$

It can often be the case that the denominator of (2) does not have a closed form, so that inference must be made using Markov chain Monte Carlo techniques. However, in our example, we can use the binomial theorem and the expansion

$$\left[p\left(1-\pi_1\right)+\left(1-p\right)\left(1-\pi_0\right)\right]^{n-R}=\sum_{r=0}^{n-R}\binom{n-R}{r}p^r\left(1-\pi_1\right)^r\left(1-p\right)^{n-R-r}\left(1-\pi_0\right)^{n-R-r}\tag{3}$$

to simplify the calculations. Equation (3) allows us to represent the denominator of (2) as the integral of the sum of kernels of the beta distribution, so that (2) has a closed form representation and we can obtain the moments of the posterior distribution.

If i is an ignorable cell, then the posterior means are given by

$$E\left(p_i \mid \mathbf{R}, \mathbf{Y}, M_S\right) = \frac{B\left(Y_i + a + 1, R_i - Y_i + b\right)}{B\left(Y_i + a, R_i - Y_i + b\right)} = \frac{Y_i + a}{R_i + a + b}$$

$$E\left(\pi_i \mid \mathbf{R}, \mathbf{Y}, M_S\right) = \frac{B\left(R_i + \alpha + 1, n_i - R_i + \beta\right)}{B\left(R_i + \alpha, n_i - R_i + \beta\right)} = \frac{R_i + \alpha}{n_i + \alpha + \beta}.$$

If i and j are paired non-ignorable cells, the posterior means are given by

$$\begin{aligned} E\left(p_i \mid \mathbf{R}, \mathbf{Y}, M_S\right) = \sum_{r=0}^{n_i - R_i} \sum_{s=0}^{n_j - R_j} \Bigg\{ & \binom{n_i - R_i}{r} \binom{n_j - R_j}{s} B\left(Y_i + r + a + 1, n_i - Y_i - r + b\right) \\ & \times B\left(Y_j + s + a, n_j - Y_j - s + b\right) B\left(Y_i + Y_j + \alpha_1, r + s + \beta_1\right) \\ & \times B\left(R_i - Y_i + R_j - Y_j + \alpha_0, n_i - R_i - r + n_j - R_j - s + \beta_0\right) \Bigg\} / C_{ij}, \end{aligned}$$

$$\begin{aligned} E\left(\pi_1 \mid \mathbf{R}, \mathbf{Y}, M_S\right) = \sum_{r=0}^{n_i - R_i} \sum_{s=0}^{n_j - R_j} \Bigg\{ & \binom{n_i - R_i}{r} \binom{n_j - R_j}{s} B\left(Y_i + r + a, n_i - Y_i - r + b\right) \\ & \times B\left(Y_j + s + a, n_j - Y_j - s + b\right) B\left(Y_i + Y_j + \alpha_1 + 1, r + s + \beta_1\right) \\ & \times B\left(R_i - Y_i + R_j - Y_j + \alpha_0, n_i - R_i - r + n_j - R_j - s + \beta_0\right) \Bigg\} / C_{ij}, \end{aligned}$$

and

$$\begin{aligned} E\left(\pi_0 \mid \mathbf{R}, \mathbf{Y}, M_S\right) = \sum_{r=0}^{n_i - Y_i} \sum_{s=0}^{n_j - R_j} \Bigg\{ & \binom{n_i - R_i}{r} \binom{n_j - R_j}{s} B\left(Y_i + r + a, n_i - Y_i - r + b\right) \\ & \times B\left(Y_j + s + a, n_j - Y_j - s + b\right) B\left(Y_i + Y_j + \alpha_1, r + s + \beta_1\right) \\ & \times B\left(R_i - Y_i + R_j - Y_j + \alpha_0 + 1, n_i - R_i - r + n_j - R_j - s + \beta_0\right) \Bigg\} / C_{ij}, \end{aligned}$$

where

$$\begin{aligned} C_{ij} = \sum_{r=0}^{n_i - R_i} \sum_{s=0}^{n_j - R_j} \Bigg\{ & \binom{n_i - R_i}{r} \binom{n_j - R_j}{s} B\left(Y_i + r + a, n_i - Y_i - r + b\right) B\left(Y_j + s + a, n_j - Y_j - s + b\right) \\ & \times B\left(Y_i + Y_j + \alpha_1, r + s + \beta_1\right) B\left(R_i - Y_i + R_j - Y_j + \alpha_0, n_i - R_i - r + n_j - R_j - s + \beta_0\right). \end{aligned}$$

Let $\{M_1, \ldots, M_K\}$ be the set of models under consideration and let P be a prior probability measure on this set. The posterior probability that a model is correct can be computed using Bayes' formula:

$$P\left(M_l \mid \mathbf{R}, \mathbf{Y}\right) = \frac{P\left(\mathbf{R}, \mathbf{Y} \mid M_l\right) P\left(M_l\right)}{P\left(\mathbf{R}, \mathbf{Y}\right)} = \frac{P\left(\mathbf{R}, \mathbf{Y} \mid M_l\right) P\left(M_l\right)}{\sum_l P\left(\mathbf{R}, \mathbf{Y} \mid M_l\right) P\left(M_l\right)}, \tag{4}$$

where $P(\mathbf{R}, \mathbf{Y} \mid M) = \iiint f(\mathbf{R}, \mathbf{Y} \mid \boldsymbol{p}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_0, M)\, \pi(\boldsymbol{p}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_0 \mid M)\, d\boldsymbol{p}\, d\boldsymbol{\pi}_1 d\boldsymbol{\pi}_0$. If P is the uniform probability measure, that is, $P(M_i) = P(M_j)$ for all i and j so that all models are a priori equally likely, then (4) simplifies to

$$P(M_l \mid \mathbf{R}, \mathbf{Y}) = \frac{P(\mathbf{R}, \mathbf{Y} \mid M_l)}{\sum_l P(\mathbf{R}, \mathbf{Y} \mid M_l)}.$$

Because we are are assuming Beta priors for the model parameters we can again use the binomial theorem to expand the likelihood and calculate the probabilities given by (4) explicitly.

Once we have specified a probability measure on the set of models, we can use the posterior probabilities (4) to average the parameter estimates for the different models. For example,

$$\widehat{p}_i = E(p_i \mid \mathbf{R}, \mathbf{Y}) = \sum_l E(p_i \mid \mathbf{R}, \mathbf{Y}, M_l)\, P(M_l \mid \mathbf{R}, \mathbf{Y}).$$

Similar expressions can be found for $\widehat{\pi}_{i1}$ and $\widehat{\pi}_{i0}$.

The remaining problems are to choose the set of models and to specify a prior distribution for this set. In our examples, we consider the set of all possible saturated models, that is, all combinations of ignorable cells and cells that share parameters with exactly one other cell. The computations required for this method are reasonable if the number of cells is not too large. However, the number of possible saturated models increases rapidly as the number of cells increase. For example, if there are 10 cells, the number of possible saturated models is 9496. If there are 15 cells, the number of possible saturated models increases to 10,349,536.

In selecting a prior distribution for the set of possible models, we may believe that an ignorable model is a reasonable starting point for our analysis, but would like to allow for the possibility of non-ignorability and to incorporate some uncertainty about the degree of non-ignorability. This is consistent with the idea of "centering" a non-ignorable model around an ignorable model (Forster and Smith, 1998; Nandram and Choi, 2002). For this reason, instead of using a uniform prior on the set of models, it may make sense to put more prior weight on the ignorable components of the models. In our examples we assume each cell has a prior probability of 0.5 of being ignorable. The remaining partitions are then equally likely.

3 Data Example

3.1 Description of the data

In this section we apply the methodology described in Section 2 to a data set and compare the method to non-ignorable hierarchical Bayesian models which do not use partitioning. The data set is presented in Table 2 and is constructed using A.C.E. 2000 data post-stratified into 15 cells, so that the elements of each cell share similar characteristics and correct enumeration rate. See U.S. Census Bureau (2004) for a complete description of the A.C.E. 2000 design and Moldoff and Viehdorfer (2010) for the methods used to construct Table 2. Grouping according to correct enumeration rate was done under the assumption that the unresolved enumerations are missing at random. The variables used to construct this data set are before followup group ($B = 1 - 5$), date of birth reported ($D = 0, 1$), Hispanic origin ($H = 0, 1$), imputed characteristics ($I = 0, 1$), and proxy/non-proxy ($P = 0, 1$). The before followup group is a categorical variable ranging from 1 – 5 indicating whether a household is a match, a conflicting household, a partial household non-match, a whole household non-match, or a household for which there is insufficient information. The variable D is 1 or 0 corresponding to whether date of birth is reported or not reported, respectively, and the variable H is 1 or 0 corresponding to whether the person is of Hispanic origin or not. The P variable indicates whether information was collected by proxy; for example, if a neighbor reported information instead of the individual of interest. Finally, the variable I indicates whether imputation was needed for one or more characteristics.

3.2 Fit of the models to the data

In typical statistical problems in which some inference is to be made based on a sample of data, there is uncertainty in estimates due to the finiteness of the sample. In missing data problems, in addition to finite sample statistical imprecision, there is *statistical uncertainty* (Molenberghs et al., 2001) due to the lack

Table 2: A.C.E. 2000 data

Cell	Cell Composition	Resolved CE	Resolved EE	Unresolved #	Unresolved %
1	$B = 1/D = 1$	576833	6177	7681	1.3
2	$B = 1/D = 0/P = 1$	5477	698	482	7.2
3	$B = 1/D = 0/P = 0$	24956	1077	691	2.6
4	$B = 2/D = 1$	3875	157	1072	21.0
5	$B = 2/D = 0$	612	39	209	24.3
6	$B = 3/D = 0$	2336	50	314	11.6
7	$B = 3/D = 1/H = 1$	3475	30	393	10.1
8	$B = 3/D = 1/H = 0/O = 1$	9460	169	624	6.1
9	$B = 3/D = 1/H = 0/O = 0$	4748	62	799	14.2
10	$B = 4/D = 1$	36835	612	5502	12.8
11	$B = 4/D = 0/P = 1$	1226	57	350	21.4
12	$B = 4/D = 0/P = 0$	2460	66	472	15.7
13	$B = 5/I = 0$	2181	157	1550	39.9
14	$B = 5/I = 1/H = 0$	3132	479	4239	54.0
15	$B = 5/I = 1/H = 1$	409	38	641	58.9
Total		678015	9868	25019	3.5

of information in missing data. Consider "optimistic" estimates, where estimates are constructed as if all missing data has the characteristic of interest, and "pessimistic" estimates, where estimates are constructed as if all missing data does not have the characteristic of interest. Any reasonable point estimate should be between these two extremes. Unfortunately, there does not seem to be any way to distinguish between the estimators in this range. Let $\widehat{p}_L$ denote the lower pessimistic estimate and $\widehat{p}_U$ denote the upper optimistic estimate. Let $\widehat{p}_t = (1-t)\widehat{p}_L + t\widehat{p}_U$, $t \in [0,1]$ be the set of estimates in the interval $[\widehat{p}_L, \widehat{p}_U]$. Suppose we evaluate the quality of our statistic using the quadratic loss function. Then straightforward calculations give the risk of $\widehat{p}_t$ as

$$R\left(\widehat{p}_t, p\right) = E\left(\widehat{p}_t - p\right)^2 = \left((t-1)p(1-\pi_1) + t(1-p)(1-\pi_0)\right)^2 + O\left(1/n\right). \tag{5}$$

The first term on the right hand side of equation (5) is constant and gives the risk due to statistical uncertainty, that is, the risk due to the missingness of data. The risk due to statistical imprecision is of order $1/n$. This shows that the statistic $\widehat{p}_t$ can have substantial risk for any value of t, even for arbitrarily large values of n. The first term on the right hand side of equation (5) is minimized at $t^* = p\left(1-\pi_1\right) / \left[p\left(1-\pi_1\right) + \left(1-p\right)\left(1-\pi_0\right)\right]$. Unfortunately, t^* is a function of the parameters and is not estimable without some additional knowledge of the structure of the parameters.

A first step in the analysis is to determine what can be done without making any further model assumptions. The estimable quantities are functions of the probability of being a resolved, correct enumeration, $p\pi_1$, the probability of being a resolved, erroneous enumeration, $(1-p)\pi_0$, and the probability of being unresolved, $p\left(1-\pi_1\right) + \left(1-p\right)\left(1-\pi_0\right)$. Due to the non-identifiability of the likelihood, it is impossible to find estimates of the parameters p, π_1, and π_0 with properties such as unbiasedness or consistency. We can, however, find a range of estimates for the unknown parameters. Table 3 gives a range of plausible estimates for the unknown parameters with range endpoints corresponding to the optimistic and pessimistic estimates discussed previously. For cells with a high percentage of missing data there can be a wide range of reasonable point estimates.

To obtain a single point estimate we must place additional structure on the model. Under the assumption of ignorability, there are two free parameters in each cell, p_i and $\pi_i = \pi_{i1} = \pi_{i0}$. The ignorable estimates based on the data in Table 2, which can be calculated using the formulas

$$\widehat{p}_i^I = \frac{Y_i}{R_i}, \quad \widehat{\pi}_i^I = \frac{R_i}{n_i},$$

Table 3: Optimistic/pessimistic point estimates

	p_i		π_{i1}		π_{i0}	
Cell	Pessimistic	Optimistic	Pessimistic	Optimistic	Optimistic	Pessimistic
1	0.977	0.990	0.987	1.000	0.446	1.000
2	0.823	0.895	0.919	1.000	0.592	1.000
3	0.934	0.960	0.973	1.000	0.609	1.000
4	0.759	0.969	0.783	1.000	0.128	1.000
5	0.712	0.955	0.745	1.000	0.157	1.000
6	0.865	0.981	0.882	1.000	0.137	1.000
7	0.891	0.992	0.898	1.000	0.071	1.000
8	0.923	0.984	0.938	1.000	0.213	1.000
9	0.846	0.989	0.856	1.000	0.072	1.000
10	0.858	0.986	0.870	1.000	0.100	1.000
11	0.751	0.965	0.778	1.000	0.140	1.000
12	0.821	0.978	0.839	1.000	0.123	1.000
13	0.561	0.960	0.585	1.000	0.092	1.000
14	0.399	0.939	0.425	1.000	0.102	1.000
15	0.376	0.965	0.390	1.000	0.056	1.000

are presented in Table 4.

Non-ignorable models typically rely on prior assumptions which effectively reduce the dimension of the parameter space by specifying the distribution of the unknown parameters. We briefly describe three commonly used Bayesian models and use these models as a comparison to the partition model. One of the simplest non-ignorable models making the fewest assumptions assumes $p_i, \pi_{i1}, \pi_{i0} \overset{i.i.d.}{\sim} Beta(1,1)$ for $i = 1, \ldots, I$. This model specification uses a non-informative flat prior distribution and makes no assumptions about how the model parameters relate to one another. Also, there is no "borrowing strength" in this model specification, that is, the parameter estimates for a cell use only the data associated with that cell. We call this model the uniform model.

The uniform model can be generalized by letting $p_i \overset{i.i.d.}{\sim} Beta(a,b)$, $\pi_{i1} \overset{i.i.d.}{\sim} Beta(\alpha_1, \beta_1)$, and $\pi_{i0} \overset{i.i.d.}{\sim} Beta(\alpha_0, \beta_0)$ for $i = 1, \ldots, I$, where the hyperparameters a, b, α_1, β_1, α_0, and β_0 are fixed but unknown. This model was considered by Stasny (1991). The posterior distribution of $(a, b, \alpha_1, \beta_1, \alpha_0, \beta_0)$ given the data is calculated by integrating out the parameters $\boldsymbol{p}$, $\boldsymbol{\pi}_1$, and $\boldsymbol{\pi}_0$, and the hyperparameters are estimated using generalized maximum likelihood techniques. The distributions $Beta\left(\widehat{a}, \widehat{b}\right)$, $Beta\left(\widehat{\alpha}_1, \widehat{\beta}_1\right)$, and $Beta\left(\widehat{\alpha}_0, \widehat{\beta}_0\right)$ are used as the "true" prior distributions.

The empirical Bayes method of Stasny (1991) was developed before the MCMC explosion for Bayesian computation (Gelfand and Smith, 1990). It can now be easily extended to a full Bayes procedure by assuming additional distributions on the hyperparameters. This method was considered by Nandram and Choi (2002). The difficulty is in selecting appropriate prior distributions for the hyperparameters, as there are many reasonable non-informative prior distributions which could be used. For the analysis of this data set, the beta priors were reparameterized, and uniform and Pareto distributions were used as the distributions for the hyperparameters, so that

$$\begin{aligned} p_i &\overset{i.i.d.}{\sim} Beta(\mu\tau, (1-\mu)\tau), \\ \pi_{i1} &\overset{i.i.d.}{\sim} Beta(\mu_1\tau_1, (1-\mu_1)\tau_1), \\ \pi_{i0} &\overset{i.i.d.}{\sim} Beta(\mu_0\tau_0, (1-\mu_0)\tau_0), \\ \mu, \mu_1, \mu_0 &\overset{i.i.d}{\sim} U(0,1), \\ \tau, \tau_1, \tau_0 &\overset{i.i.d}{\sim} \frac{1}{(x+1)^2}. \end{aligned} \tag{6}$$

The mean of the posterior distribution is used as a point estimate. The posterior distribution does not have

Table 4: Comparison of model based estimates

	Ignorable			Uniform			EB			FB			Partition model		
Cell	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$
1	0.989	0.987	0.987	0.984	0.993	0.685	0.983	0.994	0.621	0.984	0.993	0.695	0.986	0.991	0.738
2	0.887	0.928	0.928	0.861	0.955	0.775	0.848	0.970	0.698	0.862	0.956	0.781	0.874	0.942	0.843
3	0.959	0.974	0.974	0.948	0.985	0.787	0.942	0.991	0.710	0.945	0.988	0.759	0.946	0.987	0.762
4	0.961	0.790	0.790	0.894	0.853	0.410	0.938	0.810	0.550	0.936	0.813	0.620	0.862	0.885	0.322
5	0.940	0.757	0.757	0.862	0.830	0.439	0.908	0.786	0.549	0.915	0.782	0.660	0.818	0.878	0.339
6	0.979	0.884	0.884	0.940	0.921	0.427	0.963	0.899	0.552	0.954	0.908	0.576	0.931	0.930	0.404
7	0.991	0.900	0.900	0.961	0.928	0.346	0.984	0.906	0.552	0.971	0.918	0.511	0.953	0.936	0.263
8	0.982	0.939	0.939	0.960	0.961	0.505	0.968	0.954	0.556	0.965	0.957	0.591	0.981	0.941	0.878
9	0.987	0.858	0.858	0.944	0.898	0.345	0.978	0.866	0.552	0.966	0.877	0.550	0.932	0.910	0.290
10	0.984	0.872	0.872	0.943	0.911	0.383	0.971	0.883	0.551	0.961	0.894	0.545	0.924	0.930	0.276
11	0.956	0.786	0.786	0.886	0.850	0.423	0.929	0.809	0.550	0.932	0.808	0.648	0.855	0.883	0.336
12	0.974	0.843	0.843	0.923	0.891	0.408	0.955	0.859	0.551	0.947	0.868	0.589	0.898	0.916	0.296
13	0.933	0.601	0.601	0.816	0.703	0.351	0.918	0.613	0.548	0.922	0.610	0.646	0.733	0.796	0.280
14	0.867	0.460	0.460	0.728	0.582	0.347	0.875	0.458	0.545	0.901	0.445	0.711	0.620	0.716	0.257
15	0.915	0.411	0.411	0.753	0.536	0.283	0.928	0.409	0.549	0.925	0.415	0.644	0.895	0.432	0.403

a closed form, so a Metropolis-Hastings MCMC algorithm must be used to approximate characteristics of the posterior distribution. We note that Nandram and Choi (2002) developed a more sophisticated version of model (6) that "centers" the model around the parameter $\gamma_i = \pi_{i1}/\pi_{i0}$, which they call the expansion model. However, we use the simpler model (6) as a point of comparison, since it is much easier to implement.

Table 4 presents a comparison of the point estimates made using the different Bayesian models described in this section and the estimates made using the partition model discussed in Section 2. All four methods indicate some degree of non-ignorability in each of the cells. Recall that our intuition led us to believe that it is more likely that a correct enumeration will be resolved than an erroneous enumeration, that is, that $\pi_{i1} > \pi_{i0}$ for $i = 1, \ldots, 15$. This relationship holds for each model in cells $1 - 12$, with the uniform model and the partition model producing estimates that are generally furthest from the ignorable estimates, resulting in the most pessimistic estimates. A comparison of the estimates for p_i shows that we could be overestimating the proportion of correct enumerations in some of the cells if the assumption of ignorability is incorrect.

Table 5: Summary: number of imputed correct enumerations

		Method				
Cell	# UR	Ignorable	Uniform	EB	FB	Partition
1	7681	7597	4436	3671	4496	5328
2	482	428	267	172	268	338
3	691	663	388	232	319	323
4	1072	1030	726	927	941	501
5	209	196	137	172	183	89
6	314	307	215	268	257	176
7	393	389	287	365	334	243
8	624	613	408	473	464	596
9	799	789	579	743	708	473
10	5502	5414	3877	4936	4686	2774
11	350	335	234	297	309	164
12	472	460	325	410	402	223
13	1550	1446	1038	1404	1439	663
14	4239	3675	2677	3785	4010	1731
15	641	587	425	605	611	564
Total	25019	23929	16019	18460	19427	14186

Table 6: Comparison of model based estimates for cells 13 – 15

	Ignorable			Uniform			EB			FB			Model averaging		
Cell	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$	$\widehat{p}_i$	$\widehat{\pi}_{i1}$	$\widehat{\pi}_{i0}$
13	0.933	0.601	0.601	0.816	0.703	0.351	0.760	0.745	0.189	0.866	0.650	0.375	0.670	0.850	0.157
14	0.867	0.460	0.460	0.728	0.582	0.347	0.651	0.629	0.193	0.783	0.518	0.355	0.497	0.836	0.150
15	0.915	0.411	0.411	0.753	0.536	0.283	0.684	0.569	0.133	0.817	0.474	0.307	0.913	0.412	0.416

Cells 14 and 15 using the empirical Bayes method and cells 13 – 15 using the full Bayes version give estimates contrary to intuition where $\widehat{\pi}_{i1} < \widehat{\pi}_{i0}$. One possible reason this may happen is that there is excessive shrinking to the prior mean in these models. In particular, the empirical Bayes method gives estimates of π_{i0} in cells 4 – 15 near 0.55 with very little variability between cells. It is of course also possible that the characteristics of persons belonging to cells 13 – 15 cause the true proportions to be counter to intuition.

Table 5 gives a summary of the results of both the ignorable and the non-ignorable methods by showing the number of imputed correct enumerations. There is clearly strong sensitivity to the selected model. The number of imputed correct enumerations ranges from a high of 23,929, or 96% of the unresolved data, using the ignorable model, to a low of 14,186, or 57% of the unresolved data, using the partition model.

3.3 Analysis of cells 13, 14, and 15

Because the empirical Bayes and full Bayes models produced estimates for the parameters in cells 13 – 15 that were unexpected, we decided to analyze this data separately. Analyzing these cells separately has the additional advantage of showing how dependent the estimates for the parameters in one cell are to the data in other cells. The estimates based on this subset of the data are given in Table 6.

The ignorable model and the uniform model do not "borrow strength" from other cells, that is, the parameter estimates for a cell are based only on the data contained in that cell. Therefore, these estimates are identical to the estimates calculated using the full data set.

The estimates based on this subset of data using the partition model are not drastically different than the estimates using the full data set. The estimates for the parameters in cells 13 and 14 show the same non-ignorability relationship and the estimates for the parameters in cell 15 indicate that the cell is nearly ignorable. The reason for these results is that most of the posterior weight (≈ 0.91) is given to the partition $M_S = \{(13, 14), 15\}$.

The most interesting thing to notice in Table 6 is how different the estimates for cells 13 – 15 based on the empirical Bayes and the full Bayes models are from those in Table 4. The inequality is now reversed so that $\widehat{\pi}_{i1} > \widehat{\pi}_{i0}$. This indicates that there may be excessive shrinking to the mean when using hierarchical Bayes methods.

4 Simulation Example

Due to the non-identifiability of the model, it is impossible to construct a goodness-of-fit statistic to verify the correctness of the assumption of ignorability or non-ignorability. Hence, in the previous data example, since we do not know the truth, we cannot say one model is more appropriate than another. In this section we conduct a simulation study to compare the different models.

A set of population parameter values used for the simulation study is presented in Table 7. The p_i were drawn from a uniform distribution and the π_i were drawn from a distribution giving probability of 1/2 each to 0.9 and to 0.1. Five of the cells in Table 7 are ignorable while five of the cells are non-ignorable. The non-ignorable cells 2, 4, and 5 share parameters such that $\pi_1 < \pi_0$, and non-ignorable cells 3 and 7 share parameters such that $\pi_1 > \pi_0$. This set of parameters most likely does not correspond to any realistic population. It does give a population with a mix of ignorable and non-ignorable cells with parameters that are rather extreme, with the true parameter values near the boundaries. Also, the non-ignorable cells are not uniform, in the sense that some non-ignorable cells are such that $\pi_{i1} > \pi_{i0}$, while some cells are such

Table 7: True probabilities

Cell	p_i	π_{i1}	π_{i0}	$nE(R)$
1	0.850	0.9	0.9	90.0
2	0.078	0.1	0.9	83.7
3	0.023	0.9	0.1	11.8
4	0.288	0.1	0.9	67.0
5	0.329	0.1	0.9	63.7
6	0.430	0.1	0.1	10.0
7	0.488	0.9	0.1	49.1
8	0.886	0.9	0.9	90.0
9	0.100	0.9	0.9	90.0
10	0.389	0.1	0.1	10.0

that $\pi_{i1} < \pi_{i0}$. This setup gives us a chance to compare the fit of the different models for a variety of cell types.

For our simulation study we generated a sample of size 100 in each of the 10 cells from the likelihood (1) using the probabilities in Table 7 and generated point estimates using each of the previously discussed methods. This procedure was repeated 100 times. All work was done using R. Tables 8 – 10 give the sample mean, standard deviation, 2.5th and 97.5th percentiles of the point estimates for each model. The goodness of fit statistics given are estimates of the maximum absolute deviation (MAX), $\max_i E\left|\widehat{\theta}_i - \theta_i\right|$, the root mean square error (RMSE), $\left(1/I\sum_{i=1}^{I} E\left(\widehat{\theta}_i - \theta_i\right)^2\right)^{1/2}$, and the mean absolute error (MAE), $1/I\sum_{i=1}^{I} E\left|\widehat{\theta}_i - \theta_i\right|$.

For individual cells, all models were capable of producing estimates that missed the true value of the parameter badly. The parameters associated with cell 3 proved to be the most difficult to estimate. This is due to the fact that most of the observations are missing ($\approx 88\%$) and that the characteristics of the unresolved enumerations are so different than those of the resolved enumerations. For a sample size of 100 observations in cell 3, we expect approximately 2 correct resolved enumerations, 10 erroneous resolved enumerations, 0 correct unresolved enumerations, and 88 erroneous unresolved enumerations. None of the methods estimated the parameters associated with cell 3 very well and some of the methods estimated these parameters exceptionally badly. For example, the empirical Bayes estimates were $(\widehat{p}_3, \widehat{\pi}_{31}, \widehat{\pi}_{30}) = (0.824, 0.058, 0.792)$ while the true value of the parameters are $(0.023, 0.9, 0.1)$.

No one model uniformly outperformed the other models. However, somewhat surprisingly, the uniform model came closest to being "best" for this example as all three goodness-of-fit statistics for each parameter were either smallest or next smallest. As we should expect, the ignorable maximum likelihood estimates are very accurate for the ignorable cells 1, 6, and 8 – 10. It is also no surprise that the estimates can be quite poor for the non-ignorable cells. Because the ignorable estimates for the five ignorable cells are so precise, the overall performance of the ignorable model was not bad, compared to the other models under consideration. The empirical Bayes and full Bayes models seemed to have the worst performance, especially for estimating the main parameter of interest, p. The partition model was not the best model in this example, but it was an improvement over the ignorable model.

This simulation study shows that the 2.5% – 97.5% ranges given in Tables 8 - 10 can do a poor job of covering the true value of the parameters, so that we can not necessarily use frequentist properties of the estimators to construct confidence intervals. We considered different methods for constructing uncertainty ranges. First, we looked at Bayesian credible intervals, which can be calculated for the uniform, empirical Bayes, and full Bayes models. Figure 1 gives the average 95% Bayesian credible intervals for the mean of the posterior distribution for these models. The true parameter values are represented by dashed lines.

The credible intervals generated from the uniform model had the best performance, as all intervals contain the true value of the parameter, with the exception of the interval for $\pi_{3,1}$. The cost is that, on average, the length of the interval is quite large. The average credible intervals based on the empirical Bayes and full Bayes methods miss the true value of the parameter too often to be useful. Not only do these intervals give

Table 8: Comparison of the estimates for p

	Ignorable				Uniform				EB				FB				Partition model			
Cell	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%
1	0.850	0.040	0.783	0.929	0.812	0.038	0.745	0.886	0.833	0.046	0.751	0.914	0.812	0.040	0.741	0.889	0.827	0.039	0.760	0.903
2	0.011	0.011	0.000	0.037	0.067	0.020	0.035	0.107	0.061	0.030	0.015	0.120	0.060	0.023	0.028	0.114	0.058	0.018	0.030	0.096
3	0.199	0.126	0.000	0.437	0.394	0.052	0.296	0.483	0.824	0.174	0.308	0.948	0.456	0.162	0.155	0.742	0.322	0.091	0.168	0.492
4	0.044	0.027	0.000	0.101	0.148	0.033	0.086	0.211	0.229	0.077	0.071	0.360	0.146	0.042	0.075	0.241	0.149	0.028	0.081	0.199
5	0.049	0.027	0.000	0.104	0.162	0.033	0.093	0.224	0.270	0.088	0.099	0.427	0.161	0.043	0.088	0.248	0.164	0.033	0.095	0.229
6	0.443	0.153	0.154	0.726	0.482	0.047	0.392	0.568	0.873	0.160	0.379	0.971	0.523	0.175	0.198	0.876	0.494	0.098	0.316	0.676
7	0.888	0.044	0.796	0.960	0.746	0.036	0.680	0.811	0.905	0.078	0.671	0.972	0.752	0.068	0.639	0.874	0.849	0.030	0.776	0.903
8	0.882	0.035	0.811	0.943	0.842	0.034	0.777	0.897	0.865	0.041	0.779	0.937	0.844	0.035	0.774	0.899	0.860	0.035	0.791	0.923
9	0.095	0.030	0.044	0.151	0.137	0.031	0.083	0.198	0.114	0.029	0.063	0.166	0.128	0.031	0.077	0.197	0.112	0.028	0.062	0.165
10	0.419	0.147	0.133	0.667	0.474	0.044	0.378	0.542	0.871	0.174	0.350	0.972	0.518	0.187	0.135	0.855	0.475	0.093	0.276	0.655
Max	0.400				0.371				0.801				0.433				0.361			
RMSE	0.200				0.167				0.369				0.209				0.177			
MAE	0.150				0.123				0.245				0.149				0.128			

Table 9: Comparison of the estimates for π_1

	Ignorable				Uniform				EB				FB				Partition model			
Cell	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%
1	0.893	0.030	0.825	0.940	0.919	0.021	0.887	0.953	0.902	0.037	0.817	0.963	0.925	0.023	0.870	0.961	0.904	0.024	0.846	0.939
2	0.836	0.039	0.760	0.910	0.332	0.059	0.252	0.441	0.297	0.129	0.082	0.576	0.423	0.106	0.223	0.618	0.448	0.090	0.298	0.649
3	0.117	0.033	0.060	0.185	0.193	0.030	0.139	0.249	0.058	0.084	0.003	0.283	0.295	0.134	0.035	0.571	0.147	0.032	0.082	0.213
4	0.667	0.042	0.600	0.745	0.344	0.061	0.224	0.452	0.196	0.112	0.037	0.459	0.433	0.107	0.210	0.625	0.392	0.073	0.257	0.512
5	0.635	0.051	0.540	0.735	0.340	0.058	0.217	0.450	0.182	0.128	0.030	0.471	0.431	0.105	0.196	0.629	0.392	0.067	0.255	0.520
6	0.102	0.033	0.050	0.175	0.214	0.032	0.156	0.290	0.081	0.092	0.015	0.316	0.309	0.136	0.066	0.567	0.137	0.041	0.063	0.218
7	0.491	0.050	0.400	0.590	0.606	0.043	0.528	0.691	0.487	0.082	0.375	0.717	0.623	0.072	0.497	0.770	0.515	0.045	0.427	0.599
8	0.897	0.030	0.835	0.940	0.924	0.020	0.881	0.954	0.905	0.035	0.832	0.963	0.928	0.024	0.877	0.962	0.906	0.026	0.848	0.944
9	0.899	0.033	0.835	0.960	0.664	0.076	0.519	0.793	0.768	0.076	0.608	0.889	0.723	0.068	0.600	0.837	0.826	0.079	0.660	0.930
10	0.100	0.032	0.040	0.155	0.210	0.031	0.150	0.263	0.078	0.092	0.015	0.348	0.305	0.141	0.050	0.590	0.134	0.040	0.058	0.205
Max	0.783				0.707				0.842				0.605				0.753			
RMSE	0.441				0.293				0.322				0.314				0.323			
MAE	0.316				0.223				0.199				0.253				0.228			

Table 10: Comparison of the estimates for π_0

	Ignorable				Uniform				EB				FB				Partition model			
Cell	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%	μ	σ	2.5%	97.5%
1	0.893	0.030	0.825	0.940	0.725	0.060	0.629	0.827	0.841	0.073	0.680	0.930	0.749	0.056	0.638	0.844	0.811	0.041	0.741	0.892
2	0.836	0.039	0.760	0.910	0.873	0.029	0.815	0.925	0.873	0.038	0.796	0.944	0.873	0.030	0.821	0.925	0.864	0.031	0.806	0.920
3	0.117	0.033	0.060	0.185	0.250	0.036	0.186	0.323	0.792	0.200	0.214	0.930	0.392	0.132	0.139	0.646	0.196	0.050	0.118	0.301
4	0.667	0.042	0.600	0.745	0.747	0.034	0.688	0.814	0.837	0.067	0.687	0.930	0.757	0.041	0.673	0.824	0.753	0.033	0.693	0.807
5	0.635	0.051	0.540	0.735	0.723	0.042	0.649	0.804	0.839	0.064	0.712	0.930	0.734	0.053	0.620	0.822	0.731	0.036	0.667	0.800
6	0.102	0.033	0.050	0.175	0.224	0.034	0.165	0.287	0.791	0.201	0.218	0.930	0.376	0.135	0.121	0.645	0.179	0.041	0.115	0.261
7	0.491	0.050	0.400	0.590	0.349	0.047	0.261	0.437	0.803	0.169	0.307	0.930	0.463	0.117	0.184	0.658	0.506	0.063	0.395	0.624
8	0.897	0.030	0.835	0.940	0.694	0.065	0.545	0.789	0.835	0.082	0.604	0.930	0.728	0.057	0.603	0.806	0.810	0.033	0.746	0.864
9	0.899	0.033	0.835	0.960	0.926	0.022	0.881	0.966	0.899	0.037	0.828	0.968	0.922	0.026	0.869	0.966	0.898	0.028	0.842	0.953
10	0.100	0.032	0.040	0.155	0.224	0.034	0.160	0.287	0.795	0.192	0.240	0.930	0.379	0.147	0.109	0.672	0.176	0.042	0.101	0.249
Max	0.391				0.249				0.703				0.363				0.406			
RMSE	0.172				0.162				0.460				0.236				0.164			
MAE	0.111				0.142				0.313				0.191				0.122			

poor coverage, they can be very narrow, giving a false perception of high precision. For example, the 95% average credible interval based on the empirical Bayes model for $\pi_{3,1}$ is $(0.008, 0.166)$ while the true value of the parameter is 0.9.

We also considered the range of optimistic and pessimistic estimates for use as a confidence interval. The first lines of Figure 1 give the average interval of the lower pessimistic and upper optimistic estimates. This interval performs well as it covers the true value of each parameter. It is, however, a rather conservative measure giving wide intervals.

5 Conclusion

In this paper we proposed a new partition model for analyzing categorical data with non-ignorable non-response. The partition model averages estimates from each saturated submodel to produce estimates for the full model. This model has good intuitive appeal when there is uncertainty about the reasonableness of

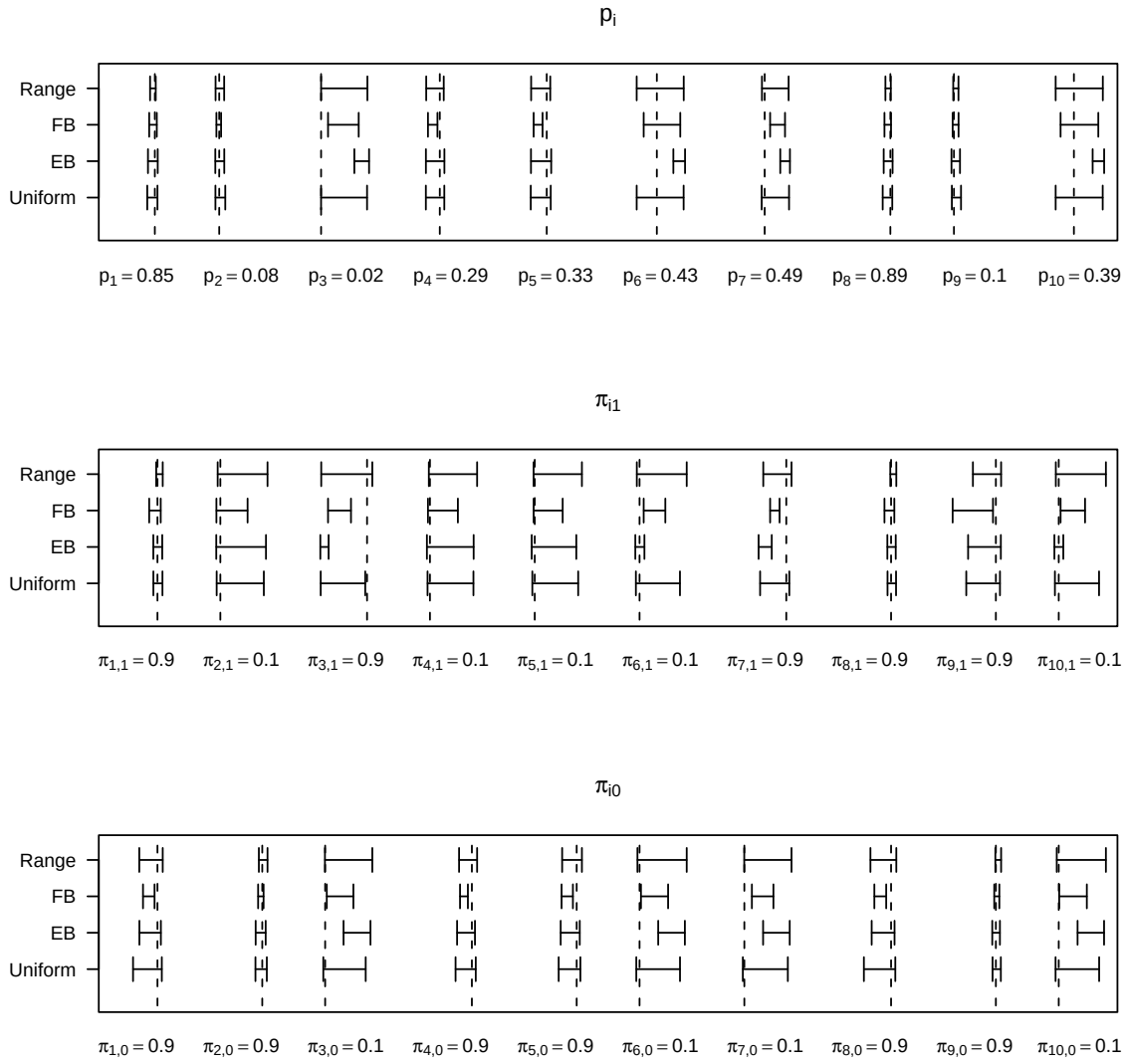

Figure 1: Average 95% credible intervals

an ignorability assumption and gives non-ignorable estimates that are not on the boundary. The degree of uncertainty about ignorability can be tuned through the prior distribution on the components of the model partition.

While the partition model is a useful tool for analyzing categorical data with possibly non-ignorable nonresponse, it is certainly not without its flaws. As with any model, since there is so much added uncertainty due to the missing data, it is impossible to produce estimates that are unbiased or consistent, or to use the data to validate the model. When analyzing this type of data, we feel that several models, such as the uniform model and the hierarchical Bayes models, should be considered and the estimates they produce compared as part of a sensitivity analysis. If trends in the estimates can be detected, it may make sense to abandon the ignorable model.

We have also considered different ways to construct confidence intervals for our parameter estimates. A range with endpoints given by pessimistic and optimistic estimates gives bounds for any reasonable model based estimates. It also provides good coverage as a conservative confidence interval, although the length of the interval is wider than one would hope. The Bayesian credible intervals can also be used, however, only

the credible intervals from the uniform model seemed reliable, and these intervals also tended to be wide.

References

Baker, S. G. and Laird, N. M. (1988), "Regression analysis for categorical variables with outcome subject to nonignorable nonresponse," *J. Amer. Statist. Assoc.*, 69, 62 – 69.

Chen, S. X., Tang, C. Y., and Mule, Jr., V. T. (2010), "Local post-stratification in dual system accuracy and coverage evaluation for the U. S. census," *J. Amer. Statist. Assoc.*, 105, 105 – 115.

Consonni, G. and Veronese, P. (1995), "A Bayesian method for combining results from several binomial experiments," *J. Amer. Statist. Assoc.*, 90, 935 – 944.

Forster, J. J. and Smith, P. W. F. (1998), "Model-based inference for categorical survey data subject to non-ignorable non-response," *J. R. Statist. Soc. B*, 60, 57 – 70.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, 85, 398 – 409.

Hartigan, J. A. (1990), "Partition models," *Comm. Statist.-Theory Meth.*, 19, 2745 – 2756.

Hogan, H. (1992), "The 1990 post-enumeration survey: an overview," *The American Statistician*, 46, 261 – 269.

Little, R. J. and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, Hoboken, New Jersey: Wiley, 2nd ed.

Malec, D. and Sedransk, J. (1992), "Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain," *Biometrika*, 79, 593 – 601.

Moldoff, M. M. and Viehdorfer, C. S. (2010), "Documentation of Research on Imputating an enumeration status for person component missing data," Tech. rep., U. S. Census Bureau, DSSD 2010 Census Coverage and Measurement Memorandum Series.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008), "Every missing not at random model has a missingness at random counterpart with equal fit," *J. R. Statist. Soc. B*, 70, 371 – 388.

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001), "Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case," *Appl. Statist.*, 50, 2373 – 2394.

Nandram, B. and Choi, J. W. (2002), "Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability," *J. Amer. Statist. Assoc.*, 97, 381 – 388.

Park, T. and Brown, M. B. (1994), "Models for categorical data with nonignorable nonresponse," *J. Amer. Statist. Assoc.*, 89, 44 – 52.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995), "Handling 'don't know' survey responses: the case of the Slovenian plebiscite," *J. Amer. Statist. Assoc.*, 90, 822 – 828.

Stasny, E. (1991), "Hierarchical models for the probabilities of a survey classification and nonresponse: an example from the national crime survey," *J. Amer. Statist. Assoc.*, 86, 296 – 303.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer.

U.S. Census Bureau (2004), "Accuracy and Coverage Evaluation of Census 2000," Tech. rep., U. S. Census Bureau, http://www.census.gov/prod/2004pubs/dssd03-dm.pdf.