HIERARCHICAL MODELS FOR RESPONSE TO THE BUSINESS AND ECONOMIC CENSUSES

Elizabeth A. Stasny, Steven N. MacEachern, Darryl T. Yamashita
Elizabeth A. Stasny, Dept. of Statistics, The Ohio State University, Columbus, OH 43210

KEY WORDS: Gibbs Sampler, Large Data Base, Missing Data

The Business and Economic Censuses are conducted every five years by the United States Bureau of the Census. In 1987, about 78% of all establishments in smaller multi-unit companies (fewer than 5000 employees) that were sent Census forms responded to the Census. Our goal in studying response by multi-unit establishments to the Census is to suggest strategies for improving this response rate. We propose hierarchical models that differentiate between companies that act on behalf of their establishments in responding to (or not responding to) the Census and those that expect their establishments to act individually with respect to the Census. The models are fit using Gibbs sampling techniques and adaptive rejection methods. Knowing whether or not companies are acting as a unit with respect to the Census would provide guidelines for targeting establishments or companies for programs to encourage response to the Census.

1. Introduction

Every five years the United States Bureau of the Census conducts the Business and Economic Censuses to collect basic economic information for approximately 12 million businesses in the United States. These Censuses are actually seven separate Censuses covering construction, manufacturing, mining, retail trade, wholesale trade, and selected service and transportation industries. Forms for the 1987 Business and Economic Censuses (hereafter referred to as "the Census") were mailed in January. The Census due date was February 15. To obtain information from as many units as possible, the Bureau of the Census sent nonresponding units reminders on March 1, April 1, May 1, June 1, and June 23. (See Zeisset (1990) for a discussion of the response patterns and follow-up activities for the 1987 Census.) Note that, as is the case for the decennial population census, it is required by law that businesses respond to the Census. Also like the population census, however, some businesses do not respond.

The observational unit for these Censuses is an establishment, defined as a single physical location where goods are sold or produced or services are rendered. A company may be made up of one or more establishments. About 1.1 million establishments in the United States belong to one of the 140,000 multi-unit companies in the 1987 Census. Since these multi-unit companies account for a disproportionate amount of the country's economic activity, it is particularly important that these companies be represented in the Census. The Bureau of the Census ensures that the Census information for the very largest companies is obtained. Thus, in this paper we consider multi-unit establishments having fewer than 5000 employees.

Establishments owned by the government or located outside the United States, for example in the Virgin Islands or Puerto Rico, were removed from the data base used in this study. The data consisted of 100,048 companies with a total of 531,530 establishments taken from the 1987 Census.

The Bureau of the Census is striving to improve both the response rate and the promptness of responses to these Censuses (see, for example, Zeisset, Mesenbourg, and Marske (1990)). In this paper, we present a hierarchical Bayes model for describing the response of multi-unit establishments to the Census. Our goal is to begin to understand nonresponse to the Census so that the Bureau of the Census may act early to encourage response from establishments that are likely nonrespondents. We also wish to determine which establishments are acting independently with respect to the Census and which companies are acting for their establishments since strategies for encouraging response will be different in these two cases.

In Section 2 we describe our hierarchical model for response to the Census by multi-unit establishments. These Bayes models are fit using Gibbs sampling with adaptive rejection. Details of the model fitting are discussed in Section 3. In Section 4 we present preliminary results from our model. Finally, in Section 5 we describe areas for future research.

2. A Hierarchical Model for Establishment Response to the Census

In this section we describe our hierarchical model for the process governing the response of multi-unit establishments to the Census. Since it is important to distinguish between establishments that are acting for themselves with respect to the Census and establishments whose companies act on their behalf, our model makes this distinction clear. Our unit of analysis is, therefore, the company. The data are the number of establishments in each company and the number of those establishments that respond to the Census.

In the following, let N be the total number of multi-unit companies. For $i = 1, 2, ..., N$ let

n_i = number of establishments for company i,

X_i = number of responding establishments out of n_i for company i,

$$Z_i = \begin{cases} 1 & \text{if company i acts on the Census for its establishments} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$Y_i = \begin{cases} 1 & \text{if company i responds to the Census} \\ 0 & \text{otherwise.} \end{cases}$$

Note that n_i and X_i are observed while Z_i and Y_i are unobserved variables. We may know, for example, that both establishments in a two-unit company responded but we do not know if they responded on their own or if the company responded for them. Similarly, if a two-establishment company decides that as company policy it will not respond to the Census, one of the establishments may not learn of that policy and may respond on its own.

In the first stage of our hierarchical model, each company decides if it is going to act on the Census for its establishments. Let ρ be the probability that a company acts for its establishments. Then, we assume that

$$Z_i \sim \text{Bernoulli}(\rho)$$

independently for $i = 1, 2, ..., N$. Given that company i decides to act for its establishments ($Z_i = 1$), then the company may respond to the Census or it may choose to ignore the Census. Let π be the probability that a company, acting for its establishments with respect to the Census, responds to the Census. We assume that

$$Y_i \mid (Z_i = 1) \sim \text{Bernoulli}(\pi)$$

independently for $i = 1, 2, ..., N$.

We have now divided multi-unit companies into three groups: 1) companies that act for their establishments and do respond to the Census (we call these "good" companies), 2) companies that act for their establishments and do not respond to the Census ("bad" companies), and 3) companies that let their establishments act individually on the Census ("ugly" companies). An establishment belonging to any one of these three types of companies may or may not respond to the Census. This is obvious in the case of an establishment acting on its own. In the case of good companies, a single establishment may not have a completed form in the Census because, for example, the form was overlooked or got lost, or the establishment went out of business. In the case of bad companies, a response may have been obtained from an establishment that decided to respond on its own. We will model the number of responding establishments for good, bad, and ugly companies, respectively, as

$$X_i \mid (Z_i = 1, Y_i = 1, n_i) \sim \text{Binomial}(p_{1i}, n_i),$$
$$X_i \mid (Z_i = 1, Y_i = 0, n_i) \sim \text{Binomial}(p_{2i}, n_i),$$
and
$$X_i \mid (Z_i = 0, n_i) \sim \text{Binomial}(p_{3i}, n_i)$$

independently for each company.

In the next stage of the hierarchical model, we assume that the p_{1i}, p_{2i}, p_{3i}, ρ, and π probabilities are sampled from Beta distributions as follows:

$$p_{1i} \sim \text{Beta}(w_1 v_1, w_1(1-v_1)),$$
$$p_{2i} \sim \text{Beta}(w_2 v_2, w_2(1-v_2)),$$
$$p_{3i} \sim \text{Beta}(w_3 v_3, w_3(1-v_3)),$$
$$\rho \sim \text{Beta}(w_4 v_4, w_4(1-v_4)), \text{ and}$$
$$\pi \sim \text{Beta}(w_5 v_5, w_5(1-v_5)).$$

Our parameterization of the Beta distributions is related to the standard α and β parameterization in that

$$v_k = \frac{\alpha_k}{\alpha_k + \beta_k} \text{ and}$$
$$w_k = \alpha_k + \beta_k$$

for $k = 1, 2, ..., 5$. We use this alternate parameterization to improve the convergence of our algorithm for fitting the model and

because it is easier to think about prior distributions in terms of the means of the distributions, $\alpha_k/(\alpha_k + \beta_k)$, and the variances ($\alpha_k + \beta_k$ is in the denominator of the expression for the variance of a Beta random variable) than in terms of the α_k and β_k themselves.

Following the example of Lehoczky and Schervish (1987), we assume that the v_k and w_k parameters are sampled, respectively, from Beta and Gamma distributions as follows:

$v_k \sim \text{Beta}(a_k, b_k)$ and
$w_k \sim \text{Gamma}(c_k, d_k)$

for $k = 1, 2, ..., 5$. Finally, we take the a_k, b_k, c_k, and d_k to be constants. For each of the w_k, we take $c_k = 25$ and $d_k = 4$. The values of a_k and b_k used in our model for the distribution of v_k are presented below.

Values for distribution of v_k

k	1	2	3	4	5
a_k	98	3	6	95	8
b_k	2	97	4	5	2

3. Gibbs Sampling for the Model

We used Gibbs sampling as described, for example, by Gelfand and Smith (1990) to fit our hierarchical model for the process governing the response of establishments to the Census. The Gibbs Sampler is an iterative stochastic technique useful for the evaluation of difficult posterior distributions, and is particularly useful when the model is hierarchical. The motivation behind this technique is that a Markov chain may be created with a state space equal to the parameter space of the posterior distribution and with a limiting distribution over this state space that coincides with the posterior distribution. A simulation of this Markov chain will then produce a sequence of parameter values that can be used to estimate posterior quantities, such as the posterior probability that a company falls in one of the three categories described in Section 2.

Our implementation of the Gibbs Sampler is largely standard. Our algorithm involves the following generations (in order) for each iteration of the sampler. The notation below follows the convention established in Gelfand and Smith (1990), where, for example, $[X|Y]$ represents the conditional distribution of X given Y. A bold face letter represents the entire vector.

1. $[Y_i, Z_i \mid p_{1i}, p_{2i}, p_{3i}, x_i, n_i]$ for $i = 1, ..., N$
2. $[p_{1i}, p_{2i}, p_{3i} \mid v, w, y_i, z_i, x_i, n_i]$ for $i = 1, ..., N$
3. $[\pi, \rho \mid y, z]$
4. $[v_i, w_i \mid p, y, z]$ for $i = 1, ..., 3$
5. $[v_4, w_4 \mid \rho]$ and $[v_5, w_5 \mid \pi]$.

The random variate generation for the first three steps is conventional. The first step is a multinomial generation designating the good, bad, or ugly status for each company. The second step consists of three generations of Beta variates corresponding to the response probabilities for establishments within each company. The third step consists of generations of two Beta variates corresponding to the probabilities of being a good, bad, or ugly company.

The fourth and fifth steps are more difficult. For the generation of v_i and w_i we can fix one of the parameters and treat the distribution of the other parameter as univariate. Since the posterior is not a familiar distribution, however, a numerical method must be used to generate random variates from this distribution. We use adaptive rejection sampling which was introduced by Gilks and Wild (1991) as a method for generating random numbers from non-standard distributions. Adaptive rejection sampling uses an upper and lower envelope of the log-density as its rejection criteria. If the rejection criteria are not met, the windows are refined to make the probability of rejection smaller.

One advantage of the parameterization that we have used for the Beta prior distributions for the p, π, and ρ parameters is that the parameters v_k and w_k are nearly independent in the conditional posterior. The same is not true for the usual parameterization based on α_k and β_k. This approximate independence in the posterior hastens the convergence of the Gibbs Sampler algorithm and results in more accurate estimates.

4. Preliminary Results

For each iteration of the Gibbs Sampler, values for $p_{1i}, p_{2i}, p_{3i}, Y_i$, and Z_i must be randomly generated for every company. Since the data contains 100,048 companies, the Gibbs Sampler must generate over 400,000 random numbers per iteration. To reduce the amount of computing time needed, our preliminary analyses use only a subset of the data which we selected by letting each company have a 0.10 probability of being

included in the subsample. This smaller data set contains 9899 companies.

We ran the Gibbs Sampler with two different sets of starting values. We assessed convergence by observing when the generated parameters for the two runs were no longer noticeably different. The two different runs started to produce similar values for all of the unknown variables before the 400th iteration. Figure 4.1 shows plots of ρ, π, v_1, and w_1 for iterations 400 through 7000 from the first run; the plots for v_2, w_2, v_3, and w_3 are similar to those for v_1 and w_1. The second run yielded similar results. These plots seem to indicate that the Gibbs Sampler has converged. Notice that even though the values of w_1 vary substantially, as do those for w_2 and w_3, the values generated for ρ and π stay approximately the same. This indicates that the values of these w_k do not greatly affect the ρ and π parameters.

Table 4.1 shows the values generated for ρ, π, v_1, w_1, v_2, w_2, v_3, and w_3 at iterations 3000, 5000, and 7000, as well as the average of the generated values over iterations 400 through 7000. Notice that about 97% of our multi-unit companies act on the Census for their establishments. Of those, about 74% are good companies that respond to the Census.

Using the generated parameter values from different points in the Gibbs Sampler, we can assess the appropriateness of our model. For each company with the same number of establishments, we generated values of x_i and compared them to the actual data. We looked at companies with $n_i = 2, 3, \ldots, 7$ establishments since there were at least 200 such companies for these values of n_i in our subsample. We generated the data as follows:

1.) Generate $z_i \sim$ Bernoulli(ρ) and $y_i \sim$ Bernoulli(π).
2.) Generate either p_{1i}, p_{2i}, or p_{3i}, depending on z_i and y_i, from $p_{ji} \sim$ Beta(v_j, w_j).
3.) Given p_{ji}, generate $x_i \sim$ Binomial(n_i, p_{ji}).

Table 4.2 lists the cell counts for the actual data and three different generated data sets for $n_i = 3$ and $n_i = 5$. We used the values from iterations 3000, 5000, and 7000 (as shown in Table 4.1) as parameters for the generation of the three data sets.

From Table 4.2 we see that for $n_i = 3$, the model overestimates the number of respondents and underestimates the number of nonrespondents. For $n_i = 5$, the model switches to underestimating the number of respondents and overestimating the number of nonrespondents. This suggests that the number of establishments in a company affects the response of an establishment and should be included in the model.

In our initial attempt to include number of establishments in the analysis, we partitioned the data set into four groups depending on n_i. The first group, Group 1, contained all companies with $n_i = 2$ and 3 establishments. Groups 2, 3, and 4 contain companies with $n_i = 4$ and 5, $n_i = 6, 7, \ldots, 10$, and $n_i \geq 11$ establishments respectively. These groups contain 6703, 1387, 997, and 782 companies

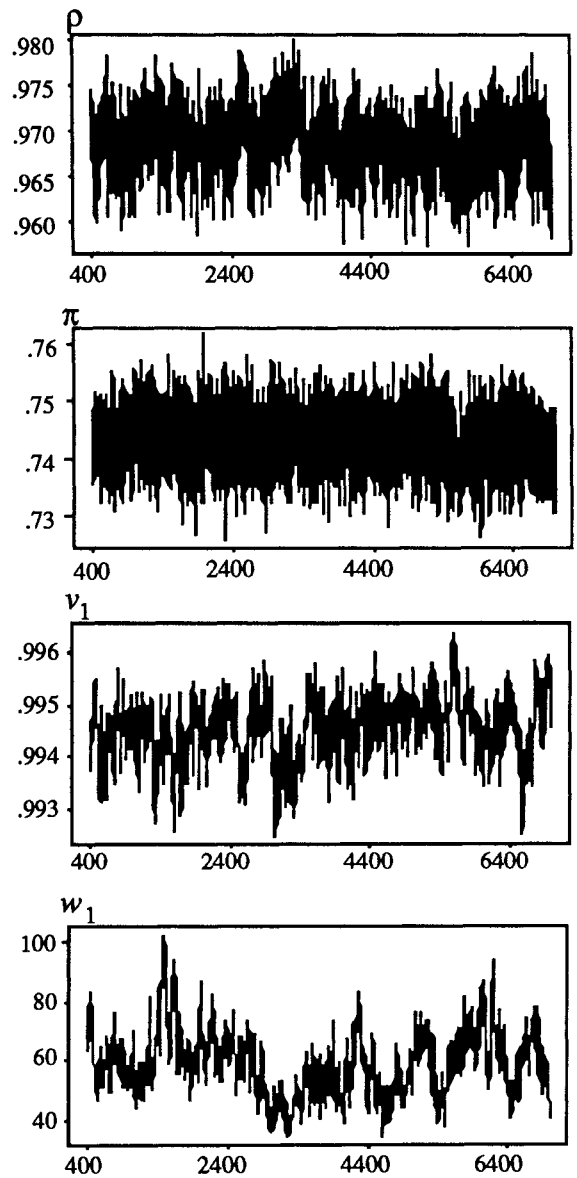

Figure 4.1: Line plots of ρ, π, v_1, and w_1 for iterations 400 to 7000 from the 1st run of the Gibbs Sampler.

respectively. The divisions for Groups 2, 3, and 4 were created to have at least two different sizes of companies within the group but still have similar numbers of cases in each group. Gibbs sampling was performed with 7500 iterations using the starting values from the first run on each of the groups to compare the various parameter estimates for the groups.

Table 4.3 presents the averages of the generated parameters for iterations 1000 to 7500 for Group 1 through Group 4. Notice that the average value of ρ, the probability that a company acts for its establishments, decreases as the number of establishments per company increases. This suggests that we might consider a model for the probability of responding that includes some sort of regression relationship between ρ and n. The relationship between number of establishments and π, the probability that a company acting for its establishments responds to the Census, is not as simple. Notice that the averages of w_1, w_2, and w_3 are approximately the same for Groups 1 and 2 and for Groups 3 and 4. These averages for Groups 1 and 2 are not far from 100, the mean of our prior distribution for the w_i's. We believe that this may be because there are fewer possible outcomes in terms of numbers of establishments that do and do not respond for these groups and, hence, it is difficult to estimate the many parameters in our model. Notice that for Groups 3 and 4, where there are more possible outcomes, the averages of the w_i's have moved away from the prior mean.

5. Future Work

We have presented our preliminary results for a hierarchical model describing the mechanism by which multi-unit establishments respond to the Business and Economic Censuses. Our results thus far lead us to believe that this modeling effort is worthwhile. They also suggest a number of areas for future work.

On the basis of the results described in Section 4, we believe that our model should be extended to account for the number of establishments in each company. We may also consider other possible covariates for our model. Although there is only limited information available about companies and their establishments when the Census forms are mailed, we do have information on the number of employees in each establishment, the

Table 4.1: Generated values for iterations 3000, 5000, 7000, and the average over iterations 400 through 7000.

	ρ	π	v_1	w_1	v_2	w_2	v_3	w_3
3000	.9702	.7418	.9937	37.33	.009636	49.15	.6358	84.21
5000	.9613	.7427	.9951	55.24	.007732	31.24	.6845	112.82
7000	.9648	.7303	.9947	41.21	.009367	71.11	.6675	84.14
Avg.	.9688	.7433	.9945	58.85	.009824	56.19	.6554	82.21

Table 4.2: Cell counts for actual data and generated data for $n_i = 3, 5$.

$n_i = 3$ x_i	actual	3000	5000	7000
0	447	423	425	447
1	37	19	18	32
2	23	49	57	48
3	1211	1227	1218	1191

$n_i = 5$ x_i	actual	3000	5000	7000
0	100	115	119	124
1	6	8	5	6
2	0	2	3	7
3	6	8	9	7
4	8	15	17	17
5	393	365	360	352

Table 4.3: Average of the generated parameters for Group 1 through Group 4.

	ρ	π	v_1	w_1	v_2	w_2	v_3	w_3
Group 1	0.977	0.722	0.994	105.1	0.01882	103.2	0.439	100.1
Group 2	0.972	0.786	0.997	102.3	0.01046	100.5	0.671	100.7
Group 3	0.970	0.744	0.994	56.4	0.00978	59.9	0.644	84.1
Group 4	0.961	0.811	0.993	53.9	0.00807	62.0	0.673	84.5

geographic location of each establishment, and the type of business carried out by each establishment. Number of establishments and number of employees per company could be included in a model directly. It will be more problematic to include location and type of business in the model since these variables may differ for individual establishments within a single company.

Since the main goals of our study include improving the response rates and reducing the time to respond to the Census, we plan to extend our analysis to a dynamic analysis that can be used to project future responses for establishments within a company based on the early returns from that company and other companies. This dynamic model should allow for the development of strategies that will improve the overall response rate to the Census. For example, if the predictive model suggests that an entire company is not responding to the Census, then follow-up mailings should be directed to the company rather than to individual establishments. In this way the model would help the Bureau of the Census take the most appropriate action to encourage response to the Census thereby saving time and money and potentially improving the response rate to the Census.

Acknowledgments

This work was supported in part through Joint Statistical Agreement 90-50 with the United States Bureau of the Census. The authors thank the Census for the opportunity to work on this problem. The authors take sole responsibility for the contents of this paper.

References

Abramowitz, M. and Stegun, I. (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Washington: U.S. Government Printing Office.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association*, **85**, 398-409.

Gilks, W. R. and Wild, P. (1991), "Adaptive Rejection Sampling for Gibbs Sampling", Technical Report No. UR-90-01, Medical Research Council Biostatistics Unit 5.

Lehoczky, J. P. and Schervish, M. J. (1987), "Hierarchical Modelling and Multi-level Analysis Applied to the National Crime Survey", paper presented at the Workshop on the National Crime Survey, July 6-17, 1987.

Zeisset, P. T. (1990), "Improving Response in the 1992 Economic Censuses", paper presented to the Census Advisory Committees of the American Marketing Association and the American Economic Association at the Joint Advisory Committee Meeting, October 18-19, 1990.

Zeisset, P. T., Mesenbourg, T. L., and Marske, R. A. (1990), "1987 Economic Censuses Advertising and Response Behavior Study", *Proceedings of the Census Bureau's 1990 Annual Research Conference*, 800-828.