Estimating Planned Sales Call Frequencies with Incomplete Information Using the EM Algorithm

Lan Ma Nygren* Lewis Coopersmith†

Abstract

We consider estimating planned sales call frequencies of a selling company with incomplete information caused by short recording durations in diary surveys. For practical reasons, it is necessary to keep the recording period short. Missing data occur when the recording period is not long enough to include observations with low call frequencies. We derive the maximum likelihood estimators of the multinomial cell probabilities for the planned sales call frequencies using the expectation maximization (EM) algorithm. We show that the EM algorithm estimators have good asymptotic properties in terms of both bias and mean squared error (MSE) and are more accurate and reliable than the estimators obtained by the naïve approach of treating the absence of a sales call as a non-called on respondent (i.e., zero frequency). The effect on the estimators when the number of frequency classes increases is also investigated.

Key Words: EM algorithm, Incomplete information, Multinomial cell probabilities, Sales call frequencies, Diary survey

1. Introduction

To estimate planned sales call frequencies of a selling company, diary surveys are often used to collect sample data. Most of these diaries have short durations such as a week because using long durations in these surveys makes it harder to recruit representative samples and may cause reporting (observation) fatigue which can jeopardize the reliability of the data collected. However, short recording durations may cause missing information for observations with low call frequencies. For example, to estimate monthly total sales calls received by physicians for each of many pharmaceutical companies, a random sample of physicians maintains a diary for a week listing sales calls they receive from various companies. For each call, the frequency of calls is also recorded, e.g., weekly, monthly, etc. When frequencies are low, say, monthly, there may be no call for a given company and frequency data are missing. The naïve approach of treating the absence of a sales call as a non-called on respondent with zero frequency may lead to estimates that are inefficient and biased. Traditional maximum likelihood method is unable to estimate the call frequencies with this type of missing information either. To illustrate, suppose the call frequencies of a pharmaceutical company can be classified into four categories: weekly, monthly, quarterly, and never, denoted by A_1, A_2, A_3, and A_4, respectively. Suppose the recording duration is a week. Denote the events that a physician receives a sales call in the recording duration and a physician does not receive a sales call in the recording duration by C and NC, respectively. The data structure is summarized in Table 1.

The multinomial cell probabilities p_i, for $i = 1, \ldots, 4$ are the parameters we are interested in estimating. The counts n_{2NC}, n_{3NC}, and n_{4NC} are regarded as latent (unobservable) since we are only able to observe their sum n_{NC}.

*Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08648

†Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08648

Table 1: Counts of Physicians And Cell Probabilities for Surveys That Measure Total Sales Calls

Categories	C	NC	Cell Prob.
Weekly	n_{1C}	0	$n_1 \backslash p_1$
Monthly	n_{2C}	(n_{2NC})	$n_2 \backslash p_2$
Quarterly	n_{3C}	(n_{3NC})	$n_3 \backslash p_3$
Never	0	(n_{4NC})	$n_4 \backslash p_4$
Total	n_C	n_{NC}	$n \backslash 1$

[a] Note: The latent frequencies of physicians are in parentheses.

Denote the conditional probabilities that a physician receives at least one sales call and a physician does not receive any sales calls in the previous week given that these physicians are in category A_i by γ_{1i} and γ_{2i}, respectively, for $i = 1, \ldots, 4$. It is self-explanatory that we have $\boldsymbol{\gamma}_{2\cdot} = (\gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}) = (0, \frac{3}{4}, \frac{11}{12}, 1)$.[1] The complementary rule of probability implies $\gamma_{1i} = 1 - \gamma_{2i},$ for $i = 1, \ldots, 4$, by which, we have $\boldsymbol{\gamma}_{1\cdot} = (\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}) = (1, \frac{1}{4}, \frac{1}{12}, 0)$. Let $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)$ be the parameter vector of cell probabilities. Denote the complete data vector of frequencies by $\mathbf{x} = (n_{1C}, n_{2C}, n_{3C}, n_{2NC}, n_{3NC}, n_{4NC})$. The complete-data likelihood function is given by

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{n_{1C}!(n_{2C}+n_{2NC})!(n_{3C}+n_{3NC})!n_{4NC}!} \prod_{i=1}^{3} (\gamma_{1i}p_i)^{n_{iC}} \prod_{j=2}^{4} (\gamma_{2j}p_j)^{n_{jNC}}. \tag{1.1}$$

From (1.1), the complete-data log likelihood, omitting terms that do not depend on p_i, for $i = 1, \ldots, 4$, is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{3} n_{iC} \log p_i + \sum_{j=2}^{4} n_{jNC} \log p_j. \tag{1.2}$$

To maximize this likelihood subject to the constraint that $\sum_{i=1}^{4} p_i = 1$, we introduce a Lagrange multiplier λ and maximize the Lagrangian function

$$Z = \sum_{i=1}^{3} n_{iC} \log p_i + \sum_{j=2}^{4} n_{jNC} \log p_j + \lambda\Big(1 - \sum_{k=1}^{4} p_k\Big). \tag{1.3}$$

On solving the five simultaneous equations obtained by setting the partial derivatives of (1.3) with respect to p_i, for $i = 1, \ldots, 4$, and λ, equal to zero, respectively, we find that the complete-data maximum likelihood estimate of the parameter $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) = \Big(\frac{n_{1C}}{n}, \frac{n_{2C}+n_{2NC}}{n}, \frac{n_{3C}+n_{3NC}}{n}, \frac{n_{4NC}}{n}\Big). \tag{1.4}$$

The first component in (1.4) provides the maximum likelihood estimate for p_1. However, we are unable to estimate p_j, $j = 2, 3, 4$ by (1.4) since the frequencies n_{2NC}, n_{3NC}, and n_{4NC} are unobservable.

[1] For the sake of simplicity, we assume that there are exactly four weeks in a month.

The naïve approach of treating the absence of a sales call as a non-called on respondent is equivalent of estimating p_2, p_3, and p_4 as $\hat{p_2} = \frac{n_{2C}}{n}$, $\hat{p_3} = \frac{n_{3C}}{n}$, and $\hat{p_4} = \frac{n_{NC}}{n}$, respectively, which is obviously inappropriate because their discrepancies from the corresponding estimates given in (1.4) may very likely cause big bias.

We propose an approach to estimate multinomial cell probabilities p_2, p_3, and p_4 when the data have latent frequencies as shown in Table 1. Our approach makes use of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). We investigate the asymptotic properties of the estimators obtained by the EM algorithm and compare them with the estimators obtained by the naïve approach in terms of the bias and mean squared error (MSE). The effect on the estimators when the number of frequency classes increases is also studied.

The rest of the paper is organized as follows. In Section 2, we derive the EM algorithm estimators for the multinomial cell probabilities. In Section 3, we derive the asymptotic variance-covariance matrix of the parameter estimators. In Section 4, we describe the simulation study used to investigate the asymptotic properties of the estimators. Section 5 discusses the results. Section 6 concludes the paper.

2. Estimating Planned Sales Call Frequencies via the EM Algorithm

The EM algorithm is an efficient iterative procedure to find maximum likelihood estimates and is particularly suitable for problems with incomplete data. Each iteration consists of an expectation step (E-step) followed by a maximization step (M-step). To apply the EM algorithm to the problem of estimating planned sales call frequencies p_2, p_3, and p_4 as discussed in the previous section, note that the complete-data likelihood in (1.1) has the regular exponential-family form

$$f(\mathbf{x}|\boldsymbol{\theta}) = b(\mathbf{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})\mathbf{t}(\mathbf{x})^T)/a(\boldsymbol{\theta}), \tag{2.1}$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\log(p_2/p_1), \log(p_3/p_1), \log(p_4/p_1))$ denotes a 1×3 vector parameter, $\mathbf{t}(\mathbf{x}) = (n_{2C} + n_{2NC}, n_{3C} + n_{3NC}, n_{4NC})$ denotes a 1×3 vector of complete-data sufficient statistics, and the superscript T denotes matrix transpose (see Equation (2.1) in Dempster et al. 1977). When the data are from a regular exponential family, a simple characterization of the EM algorithm described below applies. Let $\mathbf{y} = (n_{1C}, n_{2C}, n_{3C}, n_{NC})$ denote the observed data vector of frequencies and let $\boldsymbol{\theta}^{(v)}$ denote the current fit of $\boldsymbol{\theta}$ after v iterations. Notice that since p_1 can be estimated using the relationship $\hat{p}_1 = 1 - \hat{p}_2 - \hat{p}_3 - \hat{p}_4$, we only need to develop the EM algorithm for estimating p_2, p_3, and p_4. Therefore, from now on, the vector of unknown parameters $\boldsymbol{\theta}$ will be redefined as $\boldsymbol{\theta} = (p_2, p_3, p_4)$. The $(v+1)$th iteration consists of the following two steps:

E-step: Estimate the complete-data sufficient statistics $\mathbf{t}(\mathbf{x})$ by finding

$$\mathbf{t}^{(v)} = E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(v)}). \tag{2.2}$$

M-step: Determine $\boldsymbol{\theta}^{(v+1)}$ as the solution of the equations

$$E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) = \mathbf{t}^{(v)}. \tag{2.3}$$

As pointed out in Dempster et al. (1977), equations (2.3) are the familiar form of the likelihood equations for maximum-likelihood estimation given data from a regular exponential family. That is, if we were to suppose that $t^{(v)}$ represents the sufficient statistics computed from an observed $\mathbf{x}$ drawn from (2.1), then equations (2.3)

usually define the maximum likelihood estimator of $\boldsymbol{\theta}$. The EM algorithm proceeds by alternating (2.2) and (2.3) iteratively until $||\boldsymbol{\theta}^{(v+1)} - \boldsymbol{\theta}^{(v)}||$ is sufficiently small.

More specifically, on the $(v+1)$th iteration, in the E-step, to find the conditional expectation of the complete-data sufficient statistics $\mathbf{t}(\mathbf{x})$ given the observed data $\mathbf{y}$ and the current fit $\boldsymbol{\theta}^{(v)}$ for $\boldsymbol{\theta}$, notice that conditional on $\mathbf{y}$, effectively n_{NC}, the latent cell counts $(n_{2NC}, n_{3NC}, n_{4NC})$ follow a multinomial distribution with a total count of n_{NC} and cell probabilities

$$\Big(\frac{0.75p_2^{(v)}}{0.75p_2^{(v)} + \frac{11}{12}p_3^{(v)} + p_4^{(v)}}, \frac{\frac{11}{12}p_3^{(v)}}{0.75p_2^{(v)} + \frac{11}{12}p_3^{(v)} + p_4^{(v)}}, \frac{p_4^{(v)}}{0.75p_2^{(v)} + \frac{11}{12}p_3^{(v)} + p_4^{(v)}}\Big). \quad (2.4)$$

Therefore, the E-step yields

$$t_1^{(v)} = E(n_{2C} + n_{2NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{2C} + n_{2NC}^{(v)} \quad (2.5)$$

$$t_2^{(v)} = E(n_{3C} + n_{3NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{3C} + n_{3NC}^{(v)} \quad (2.6)$$

$$t_3^{(v)} = E(n_{4NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{4NC}^{(v)}, \quad (2.7)$$

where,

$$n_{2NC}^{(v)} = 0.75 n_{NC} p_2^{(v)} / P_{NC}^{(v)} \quad (2.8)$$

$$n_{3NC}^{(v)} = \tfrac{11}{12} n_{NC} p_3^{(v)} / P_{NC}^{(v)} \quad (2.9)$$

$$n_{4NC}^{(v)} = n_{NC} p_4^{(v)} / P_{NC}^{(v)}, \quad (2.10)$$

and $P_{NC}^{(v)} = 0.75p_2^{(v)} + \frac{11}{12}p_3^{(v)} + p_4^{(v)}$. The M-step requires the calculation of the first moments of the sufficient statistics $\mathbf{t}(\mathbf{x})$. In the Appendix A, using the properties of exponential families, we show that

$$E(t_j|\boldsymbol{\theta}) = np_{j+1}, \text{ j=1,2,3.} \quad (2.11)$$

On the $(v+1)$th iteration, the M-step is undertaken by letting the right-hand side of (2.11) equal to $t_j^{(v)}$ given in (2.5) - (2.6) and solving for $p_{j+1}^{(v+1)}$ for $j = 1, 2, 3$. This leads to the updated estimates

$$p_2^{(v+1)} = (n_{2C} + n_{2NC}^{(v)})/n \quad (2.12)$$

$$p_3^{(v+1)} = (n_{3C} + n_{3NC}^{(v)})/n \quad (2.13)$$

$$p_4^{(v+1)} = n_{4NC}^{(v)}/n. \quad (2.14)$$

The initial values used to start the EM algorithm are $p_2^{(0)} = n_{2C}/n, p_3^{(0)} = n_{3C}/n$, and $p_4^{(0)} = 1 - n_{1C}/n - p_2^{(0)} - p_3^{(0)}$.

To investigate the effect on the estimators when the number of frequency classes increases, we consider another scenario in which the company's call frequencies can be classified into seven categories, namely weekly, biweekly, monthly, bimonthly, quarterly, half-yearly, and never. Adopting the same notations C and NC to represent the events that a physician receives a sales call in the recording duration and a physician does not receive a sales call in the recording duration, respectively, the data structure is given in Table 2.

The conditional probabilities that a physician does not receive any sales calls in the previous week for each category are $\boldsymbol{\gamma}_{2\cdot} = (\gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}, \gamma_{25}, \gamma_{26}, \gamma_{27}) =$

Table 2: Counts of Physicians And Cell Probabilities for Surveys That Measure Total Sales Calls With Increased Number of Frequency Classes

Categories	C	NC	Cell Prob.
Weekly	n_{1C}	0	$n_1 \backslash p_1$
Biweekly	n_{2C}	(n_{2NC})	$n_2 \backslash p_2$
Monthly	n_{3C}	(n_{3NC})	$n_3 \backslash p_3$
Bimonthly	n_{4C}	(n_{4NC})	$n_4 \backslash p_4$
Quarterly	n_{5C}	(n_{5NC})	$n_5 \backslash p_5$
Half-yearly	n_{6C}	(n_{6NC})	$n_6 \backslash p_6$
Never	0	(n_{7NC})	$n_7 \backslash p_7$
Total	n_C	n_{NC}	$n \backslash 1$

[a] Note: The latent frequencies of physicians are in parentheses.

$(0, 0.5, 0.75, \frac{7}{8}, \frac{11}{12}, \frac{23}{24}, 1)$. Conditional on the observed sum of all latent cell counts n_{NC}, the latent cell counts

$$(n_{2NC}, n_{3NC}, n_{4NC}, n_{5NC}, n_{6NC}, n_{7NC})$$

follow a multinomial distribution with a total count of n_{NC}. Casting (2.2) in the current scenario, the $(v+1)$th iteration in the E-step yields the complete data sufficient statistics

$$t_1^{(v)} = E(n_{2C} + n_{2NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{2C} + n_{2NC}^{(v)} \tag{2.15}$$
$$t_2^{(v)} = E(n_{3C} + n_{3NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{3C} + n_{3NC}^{(v)} \tag{2.16}$$
$$t_3^{(v)} = E(n_{4C} + n_{4NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{4C} + n_{4NC}^{(v)} \tag{2.17}$$
$$t_4^{(v)} = E(n_{5C} + n_{5NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{5C} + n_{5NC}^{(v)} \tag{2.18}$$
$$t_5^{(v)} = E(n_{6C} + n_{6NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{6C} + n_{6NC}^{(v)} \tag{2.19}$$
$$t_6^{(v)} = E(n_{7NC}|\mathbf{y}, \boldsymbol{\theta}^{(v)}) = n_{7NC}^{(v)}, \tag{2.20}$$

where

$$n_{2NC}^{(v)} = 0.5 n_{NC} p_2^{(v)} / P_{NC}^{(v)} \tag{2.21}$$
$$n_{3NC}^{(v)} = 0.75 n_{NC} p_3^{(v)} / P_{NC}^{(v)} \tag{2.22}$$
$$n_{4NC}^{(v)} = \tfrac{7}{8} n_{NC} p_4^{(v)} / P_{NC}^{(v)} \tag{2.23}$$
$$n_{5NC}^{(v)} = \tfrac{11}{12} n_{NC} p_5^{(v)} / P_{NC}^{(v)} \tag{2.24}$$
$$n_{6NC}^{(v)} = \tfrac{23}{24} n_{NC} p_6^{(v)} / P_{NC}^{(v)} \tag{2.25}$$
$$n_{7NC}^{(v)} = n_{NC} p_7^{(v)} / P_{NC}^{(v)}, \tag{2.26}$$

and $P_{NC}^{(v)} = 0.5 p_2^{(v)} + 0.75 p_3^{(v)} + \frac{7}{8} p_4^{(v)} + \frac{11}{12} p_5^{(v)} + \frac{23}{24} p_6^{(v)} + p_7^{(v)}$.

Execution of the M-step on the $(v+1)$th iteration based on (2.3) yields the

updated estimates

$$p_2^{(v+1)} = (n_{2C} + n_{2NC}^{(v)})/n \tag{2.27}$$
$$p_3^{(v+1)} = (n_{3C} + n_{3NC}^{(v)})/n \tag{2.28}$$
$$p_4^{(v+1)} = (n_{4C} + n_{4NC}^{(v)})/n \tag{2.29}$$
$$p_5^{(v+1)} = (n_{5C} + n_{5NC}^{(v)})/n \tag{2.30}$$
$$p_6^{(v+1)} = (n_{6C} + n_{6NC}^{(v)})/n \tag{2.31}$$
$$p_7^{(v+1)} = n_{7NC}^{(v)}/n. \tag{2.32}$$

The initial values used to start the EM algorithm are $p_2^{(0)} = n_{2C}/n, p_3^{(0)} = n_{3C}/n, p_4^{(0)} = n_{4C}/n, p_5^{(0)} = n_{5C}/n, p_6^{(0)} = n_{6C}/n$, and $p_7^{(0)} = 1 - n_{1C}/n - \sum_{j=2}^{6} p_j^{(0)}$.

3. The Observed Information Matrix and Asymptotic Variance-Covariance Matrix Computation

The asymptotic variance-covariance matrix of parameter estimates in the EM framework is computed as the inverse of the observed incomplete-data information matrix $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y})$, where

$$\mathbf{I}(\boldsymbol{\theta}; \mathbf{y}) = -\partial^2 \log L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T,$$

and $L(\boldsymbol{\theta})$ denotes the observed data (or incomplete data) likelihood function. There are a number of ways for calculating $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y})$ in the literature. See McLachlan and Krishnan (2008) for an overview of these methods. We will use the one established in Louis (1982) to compute $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y})$.

Let $L_c(\boldsymbol{\theta})$ denote the complete data likelihood function. Let $\mathbf{S}_c(\mathbf{x}; \boldsymbol{\theta})$ denote the gradient vector of the complete-data log likelihood function, i.e.,

$$\mathbf{S}_c(\mathbf{x}; \boldsymbol{\theta}) = \partial \log L_c(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}.$$

Let $\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}; \mathbf{y})$ denote the expected conditional complete-data information matrix, i.e.,

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}; \mathbf{y}) = E_{\boldsymbol{\theta}}\{\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X}) | \mathbf{y}\}, \tag{3.1}$$

where $\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X})$ represents the complete-data information matrix, i.e.,

$$\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X}) = -\partial^2 \log L_c(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T.$$

Louis (1982) showed that

$$\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}; \mathbf{y}) - [cov_{\boldsymbol{\theta}}\{\mathbf{S}_c(\mathbf{X}; \boldsymbol{\theta}) | \mathbf{y}\}]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}. \tag{3.2}$$

We now use the example in Section 1 with the data structure given in Table 1 to illustrate how to apply (3.2) to compute the observed incomplete-data information matrix $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y})$. Let the unknown parameter vector $\boldsymbol{\theta} = (p_2, p_3, p_4)^T$. Rewriting the complete data log likelihood function given in (1.2), omitting terms that do not include unknown parameters, we have

$$\log L_c(\boldsymbol{\theta}) = (n_{2C} + n_{2NC}) \log p_2 + (n_{3C} + n_{3NC}) \log p_3 + n_{4NC} \log p_4. \tag{3.3}$$

On taking the first derivative of (3.3) with respect to $\boldsymbol{\theta}$, we have the complete-data score statistic

$$\begin{aligned}\mathbf{S}_c(\mathbf{x};\boldsymbol{\theta}) &= \partial \log L_c(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \\ &= \begin{pmatrix} \frac{n_{2C}+n_{2NC}}{p_2} \\ \frac{n_{3C}+n_{3NC}}{p_3} \\ \frac{n_{4NC}}{p_4} \end{pmatrix}. \end{aligned} \tag{3.4}$$

Taking the second derivative of (3.3) with respect to $\boldsymbol{\theta}$ yields

$$\begin{aligned}\mathbf{I}_c(\boldsymbol{\theta};\mathbf{x}) &= -\partial^2 \log L_c(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T \\ &= \begin{pmatrix} \frac{n_{2C}+n_{2NC}}{p_2^2} & 0 & 0 \\ 0 & \frac{n_{3C}+n_{3NC}}{p_3^2} & 0 \\ 0 & 0 & \frac{n_{4NC}}{p_4^2} \end{pmatrix}. \end{aligned} \tag{3.5}$$

To calculate the conditional expectation in (3.1), note that conditional on the observed data vector $\mathbf{y} = \{n_{1C}, n_{2C}, n_{3C}, n_{NC}\}$, effectively n_{NC}, the latent cell counts $(n_{2NC}, n_{3NC}, n_{4NC})$ follow a multinomial distribution with a total count of n_{NC} and cell probabilities

$$\Big(\frac{0.75p_2}{0.75p_2 + \frac{11}{12}p_3 + p_4}, \frac{\frac{11}{12}p_3}{0.75p_2 + \frac{11}{12}p_3 + p_4}, \frac{p_4}{0.75p_2 + \frac{11}{12}p_3 + p_4}\Big). \tag{3.6}$$

Assign new notations $(\tilde{p}_2, \tilde{p}_3, \tilde{p}_4)$ to the corresponding three cell probabilities in (3.6). The conditional expectation of the complete data information matrix is given by

$$\begin{aligned}\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta};\mathbf{y}) &= E_{\boldsymbol{\theta}}\Big(-\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|\mathbf{y}\Big) \\ &= \begin{pmatrix} \frac{n_{2C}+n_{NC}\tilde{p}_2}{p_2^2} & 0 & 0 \\ 0 & \frac{n_{3C}+n_{NC}\tilde{p}_3}{p_3^2} & 0 \\ 0 & 0 & \frac{n_{NC}\tilde{p}_4}{p_4^2} \end{pmatrix}. \end{aligned} \tag{3.7}$$

From (3.4), using the fact that $(n_{2NC}, n_{3NC}, n_{4NC})$ have a multinomial distribution, we find that the $[cov_{\boldsymbol{\theta}}\{\mathbf{S}_c(\mathbf{X};\boldsymbol{\theta})|\mathbf{y}\}]$ is given by

$$[cov_{\boldsymbol{\theta}}\{\mathbf{S}_c(\mathbf{X};\boldsymbol{\theta})|\mathbf{y}\}] = \begin{pmatrix} \frac{n_{NC}\tilde{p}_2(1-\tilde{p}_2)}{p_2^2} & -\frac{n_{NC}\tilde{p}_2\tilde{p}_3}{p_2p_3} & -\frac{n_{NC}\tilde{p}_2\tilde{p}_4}{p_2p_4} \\ -\frac{n_{NC}\tilde{p}_2\tilde{p}_3}{p_2p_3} & \frac{n_{NC}\tilde{p}_3(1-\tilde{p}_3)}{p_3^2} & -\frac{n_{NC}\tilde{p}_3\tilde{p}_4}{p_3p_4} \\ -\frac{n_{NC}\tilde{p}_2\tilde{p}_4}{p_2p_4} & -\frac{n_{NC}\tilde{p}_3\tilde{p}_4}{p_3p_4} & \frac{n_{NC}\tilde{p}_4(1-\tilde{p}_4)}{p_4^2} \end{pmatrix}. \tag{3.8}$$

On subtracting (3.8) from (3.7), we obtain the incomplete-data information matrix $\mathbf{I}(\boldsymbol{\theta};\mathbf{y})$ as

$$\mathbf{I}(\boldsymbol{\theta};\mathbf{y}) = \begin{pmatrix} \frac{n_{2C}+n_{NC}\tilde{p}_2^2}{p_2^2} & \frac{n_{NC}\tilde{p}_2\tilde{p}_3}{p_2p_3} & \frac{n_{NC}\tilde{p}_2\tilde{p}_4}{p_2p_4} \\ \frac{n_{NC}\tilde{p}_2\tilde{p}_3}{p_2p_3} & \frac{n_{3C}+n_{NC}\tilde{p}_3^2}{p_3^2} & \frac{n_{NC}\tilde{p}_3\tilde{p}_4}{p_3p_4} \\ \frac{n_{NC}\tilde{p}_2\tilde{p}_4}{p_2p_4} & \frac{n_{NC}\tilde{p}_3\tilde{p}_4}{p_3p_4} & \frac{n_{NC}\tilde{p}_4^2}{p_4^2} \end{pmatrix}. \tag{3.9}$$

Evaluating this last expression at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, which are the estimates obtained on the last iteration of the EM procedure, we obtain the observed incomplete-data information matrix $\mathbf{I}(\hat{\boldsymbol{\theta}};\mathbf{y})$. The asymptotic covariance matrix of the maximum likelihood estimators of $\boldsymbol{\theta} = (p_2, p_3, p_4)^T$ is computed as the inverse of the observed information matrix, $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}};\mathbf{y})$.

4. Simulation

We use a simulation study to investigate the asymptotic properties of the estimators. Denote the estimators obtained from using the EM algorithm by $\hat{\boldsymbol{\theta}}$ and the true parameter values by $\boldsymbol{\theta}$. The bias of the estimators is computed as

$$BIAS(\hat{\boldsymbol{\theta}}) = \frac{1}{r}\sum_{m=1}^{r}\hat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\theta},$$

where r is the number of simulation runs and $\hat{\boldsymbol{\theta}}^{(m)}$ is the estimate from the mth simulation run with $m = 1, \ldots, r$. The mean squared error (MSE) of the estimators is computed as

$$MSE(\hat{\boldsymbol{\theta}}) = \frac{1}{r}\sum_{m=1}^{r}(\hat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\theta})^2.$$

For the first scenario with four frequency classes, we use the true parameter values $\boldsymbol{\theta} = [p_2, p_3, p_4] = [0.30, 0.20, 0.25]$ in the simulation. In the second scenario with seven frequency classes, the true parameter values used in the simulation are $\boldsymbol{\theta} = [p_2, p_3, p_4, p_5, p_6, p_7] = [0.15, 0.20, 0.20, 0.15, 0.10, 0.10]$. We run simulation for each scenario. For each simulation we fix the sample size n from 25 to 10,000 as n = 25, 50, 100, 200, 500, 1000, 2000, 5000, and 10,000. For each sample size n, we use the following steps of simulation process.

Steps in each simulation process:

1. Create a data set of size n
   (a) Generate multinomial data
   (b) Use binomial distributions to generate total no call frequency n_2 and observed counts for each frequency class
2. Set initial values
3. Find $\hat{\boldsymbol{\theta}}$ by repeating the E- and M-steps alternatively until the differences between estimates in two consecutive iterations are all less than 10^{-5}
4. Repeat steps 1-3 1000 times and compute $BIAS(\hat{\boldsymbol{\theta}})$ and $MSE(\hat{\boldsymbol{\theta}})$.

5. Results and Discussion

For MSEs of cell probability estimates shown in Figure **??**, the pattern of convergence to zero as the sample size increases is the same for all three cell probabilities. The MSEs of the cell probability p_2 become less than 0.01 for sample sizes greater than 200. For sample sizes greater than 500, the MSEs of cell probabilities p_3 and p_4 become less than 0.01.

Table 3 shows the biases for cell probabilities p_2, p_3, and p_4 in the naïve approach of treating the absence of a sales call as a non-called on respondent in comparison with those in the EM algorithm. The biases in the naïve approach are huge and do not decrease as the sample size increases. For the EM Algorithm, the absolute value of the bias in the estimate for p_2 becomes less than 0.01 for sample sizes greater than 50 and the same achievement is made by the estimates for p_3 and p_4 for sample sizes greater than 100. In general, the bias tends to approach zero as the sample

Table 3: Comparison of biases in cell probabilities p_2, p_3, and p_4 between the EM algorithm and the naïve approach

Sample Size n	50	100	200	500	1000	2000	5000	10000
p_2								
EM Algorithm	-0.0105	-0.0016	0.0059	0.0005	0.0007	0.0002	-0.0002	-0.0007
Naïve Approach	-0.2264	-0.2251	-0.2235	-0.2249	-0.2248	-0.2249	-0.2250	-0.2252
p_3								
EM Algorithm	-0.0182	-0.0118	-0.0076	0.0054	0.0032	-0.0006	-0.0011	0.0005
Naïve Approach	-0.1832	-0.1839	-0.1839	-0.1829	-0.1831	-0.1834	-0.1834	-0.1833
p_4								
EM Algorithm	0.0243	0.0143	0.0038	-0.0057	-0.0038	0.0008	0.0010	0.0001
Naïve Approach	0.4053	0.4099	0.4095	0.4078	0.4080	0.4087	0.4082	0.4084

Table 4: Comparison of MSEs in cell probabilities p_2, p_3, and p_4 between the EM algorithm and the naïve approach

Sample Size n	50	100	200	500	1000	2000	5000	10000
p_2								
EM Algorithm	0.0194	0.0102	0.0060	0.0022	0.0012	0.0005	0.0002	0.0001
Naïve Approach	0.0526	0.0513	0.0503	0.0507	0.0506	0.0506	0.0507	0.0507
p_3								
EM Algorithm	0.0342	0.0192	0.0106	0.0049	0.0023	0.0012	0.0005	0.0002
Naïve Approach	0.0339	0.0340	0.0339	0.0335	0.0335	0.0336	0.0336	0.0336
p_4								
EM Algorithm	0.0378	0.0224	0.0127	0.0055	0.0025	0.0014	0.0005	0.0003
Naïve Approach	0.1689	0.1703	0.1689	0.1668	0.1667	0.1672	0.1667	0.1668

size increases. In other words, we have good reason to believe that the estimators we have developed are asymptotically unbiased.

Table 4 compares the MSEs in the naïve approach with those in the EM algorithm. The MSEs in the EM algorithm are much smaller than those in the naïve approach when the sample size is fixed. Moreover, the MSEs in the naïve approach stay at the similar level when the sample size increases, whereas those in the EM algorithm converge to zero as the sample size increases. Moreover, for the EM algorithm, the pattern of convergence to zero as the sample size increases is the same for all three cell probabilities. The MSEs of the cell probability p_2 become less than 0.01 for sample sizes greater than 200. For sample sizes greater than 500, the MSEs of cell probabilities p_3 and p_4 become less than 0.01.

6. Concluding Remarks

We have derived maximum likelihood estimators using the EM algorithm to estimate the multinomial cell probabilities for planned sales call frequencies with incomplete information caused by the short recording duration. The results have shown that the EM algorithm estimators are asymptotically unbiased and consistent and there-

fore more accurate and reliable than the naïve approach of treating the absence of a sales call as a non-called on respondents. Our approach can be applied to other situations where the data are collected using activity diary surveys with short recording durations. One example is the popular travel diary studies used in transportation planning. Most of these travel diaries have short durations such as a day or a week and this may cause similar missing data structures as the one discussed in this paper. One aspect of our approach that remains to be investigated in a future work is to see how different ways of specifying initial values affect the properties of the estimators.

Appendix A. Derivation of (2.11)

Rewrite the likelihood function in (1.1) in the multiparameter exponential family form of

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp[\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) T_j(\mathbf{x}) - B(\boldsymbol{\theta})], \tag{A.1}$$

where $T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))$ is a sufficient statistic as

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{n!}{n_{1C}!(n_{2C}+n_{2NC})!(n_{3C}+n_{3NC})!n_{4NC}!} \prod_{i=1}^{3} (\gamma_{1i} p_i)^{n_{iC}} \prod_{j=2}^{4} (\gamma_{2j} p_j)^{n_{jNC}} \\
&= \frac{n!W}{n_{1C}!(n_{2C}+n_{2NC})!(n_{3C}+n_{3NC})!n_{4NC}!} p_1^{n_{1C}} p_2^{n_{2C}+n_{2NC}} p_3^{n_{3C}+n_{3NC}} p_4^{n_{4NC}} \\
&= \frac{n!W}{n_{1C}!(n_{2C}+n_{2NC})!(n_{3C}+n_{3NC})!n_{4NC}!} p_1^{n} \Big(\frac{p_2}{p_1}\Big)^{n_{2C}+n_{2NC}} \Big(\frac{p_3}{p_1}\Big)^{n_{3C}+n_{3NC}} \Big(\frac{p_4}{p_1}\Big)^{n_{4NC}} \\
&= \frac{n!W}{n_{1C}!(n_{2C}+n_{2NC})!(n_{3C}+n_{3NC})!n_{4NC}!} \\
&\quad \exp[(n_{2C}+n_{2NC}) \log(p_2/p_1) + (n_{3C}+n_{3NC}) \log(p_3/p_1) + n_{4NC} \log(p_4/p_1) + n \log p_1] \\
&= \frac{n!W}{n_{1C}!(n_{2C}+n_{2NC})!(n_{3C}+n_{3NC})!n_{4NC}!} \\
&\quad \exp[(n_{2C}+n_{2NC}) \log(p_2/p_1) + (n_{3C}+n_{3NC}) \log(p_3/p_1) + n_{4NC} \log(p_4/p_1) \\
&\qquad - n \log(1 + \sum_{j=1}^{3} \exp(\log \frac{p_{j+1}}{p_1}))],
\end{aligned}$$

where $W = \prod_{i=1}^{3} \gamma_{1i}^{n_{iC}} \prod_{j=2}^{4} \gamma_{2j}^{n_{jNC}}$ is a known constant. The complete-data likelihood (1.1) is a three-parameter exponential family with $\boldsymbol{\eta} = (\log(p_2/p_1), \log(p_3/p_1), \log(p_4/p_1))$, $T(\mathbf{x}) = (n_{2C}+n_{2NC}, n_{3C}+n_{3NC}, n_{4NC})$, and $A(\boldsymbol{\eta}) = n \log(1 + \sum_{i=1}^{3} \exp(\eta_i))$.

By the Corollary 1.6.1 of Bickel and Doksum (2000), the first moments of the sufficient statistics of a k-parameter exponential family indexed by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ can be calculated as

$$E_{\boldsymbol{\eta}_0} \mathbf{T}(X) = \Big(\frac{\partial A}{\partial \eta_1}(\boldsymbol{\eta}_0), \ldots, \frac{\partial A}{\partial \eta_k}(\boldsymbol{\eta}_0)\Big)^T.$$

According to this corollary, the first moments of the sufficient statistics $T(\mathbf{x}) = (t_1, t_2, t_3) = (n_{2C}+n_{2NC}, n_{3C}+n_{3NC}, n_{4NC})$ are given by

$$E_{\boldsymbol{\theta}}(t_j) = \frac{\partial}{\partial \eta_j} n \log(1 + \sum_{j=1}^{3} e^{\eta_j}) = \frac{n e^{\eta_j}}{1 + \sum_{j=1}^{3} e^{\eta_j}} = \frac{n \frac{p_{j+1}}{p_1}}{1 + \sum_{j=1}^{3} \frac{p_{j+1}}{p_1}} = \frac{n \frac{p_{j+1}}{p_1}}{\frac{1}{p_1}} = n p_{j+1}, \text{ j=1,2,3.}$$

References

[1] Bickel, P. J. and Doksum, K. A. (2000), *Mathematical Statistics: Basic Ideas and Selected Topics, Volumn 1,* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall.

[2] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977),"Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

[3] Louis, T. A. (1982), "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 44, 226-233.

[4] McLachlan G. J. and Krishnan T. (2008), *The EM Algorithm and Extensions,* (2nd ed.), Hoboken, New Jersey: John Wiley & Sons.