Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp. 167 174

Analysis of Incomplete Data with Nonignorable Missing Values

Hyun-Jeong Kim[1)]

Abstract

In the case of "nonignorable missing data", it is necessary to assume a model dealing with the missing on each situations. In this article, for example, we sometimes meet situations where data set are income amounts in a survey of individuals and assume a model as the values are the larger, a missing data probability is the higher. The method is to maximize using the EM (Expectation and Maximization) algorithm based on the (missing data) mechanism that creates missing data of the case of exponential distribution. The method started from any initial values, and converged in a few iterations. We changed the missing data probability and the artificial data size to show the estimated accuracy. Then we discuss the properties of estimates.

Keywords : EM-algorithm, Maximum likelihood, Mechanism, Missing data

1. Introduction

When the conditional missing probability given the observed values is dependent to the missing values, it is known that the missing mechanism is nonignorable. In such cases we have to take into account the missing mechanism in the statistical analysis of the data with missing values. The aim of present paper is to show some example of statistical analysis based on the maximum likelihood (ML) method of the data with nonignaorable missing values and study the validity of such kinds of analysis.

In section 2 the likelihood function is derived for the data with nonignorable missing values. In section 3 we consider a specific model for univariate data taken from an exponential distribution with missing values. In this case the missing

1. Fulltime Lecturer, Department of General Education, Silla University, Busan 617-736, Korea.
E-mail : semikim@silla.ac.kr

probability is assumed to increase as the value increases. Then, in section 4 we explain the method for obtaining the ML estimates, i.e., the EM algorithm in general case. In section 5 we simulate changing the missing data probability and the artificial data size by three steps to show the estimated accuracy, then discuss on the properties of estimates.

From results, we know that when we estimate based on the nonignorable missing data mechanism, it is to be unbias by to assume the model including missing data mechanism.

2. Likelihood for a sample with missing values

We write $Y = (Y_{obs}, Y_{mis})$ without any loss of generality, where Y_{obs} and Y_{mis} indicate the observed and missing parts of Y, respectively.

Let R be the pattern of the missing values, that is, R is the observed vector of random variables R_i, which is defined as

$$R_i = \begin{cases} 1, & \text{if } y_i \text{ is observed,} \\ 0, & \text{if } y_i \text{ is missing.} \end{cases}$$

Then, we can formulate a model with missing data in terms of a probability distribution for Y with density $f(Y|\theta)$, indexed by unknown parameter θ, and a probability distribution $g(R|Y,\phi)$ for ϕ indicating a parameter appeared in the conditional density of R given Y. The likelihood function with missing data was defined to be any function of θ and ϕ proportional to the joint density h of R and Y as

$$\begin{aligned} L(\theta,\phi|R, Y_{obs}) &\propto h(Y_{obs}, R|\theta,\phi) \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^{n} h(y_i, r_i|\theta,\phi)\, dY_{mis} \\ &= \prod_{i=1}^{m} f(y_i|\theta) g(r_i|y_i,\phi) \prod_{j=m+1}^{n} g_1(r_j|\theta,\phi), \end{aligned}$$

where m is observed units and $n-m$ missing units, and $g_1(r|\theta,\phi)$ is the marginal density of R, i.e.,

$$g_1(r|\theta,\phi) = \int_{-\infty}^{\infty} f(y|\theta) g(r|y,\phi)\, dy.$$

Note that the joint density of Y and R can be decomposed into

$$h(y, r|\theta,\phi) = f(y|\theta) g(r|y,\phi),$$

(see, Little and Rubin(1987), and Hogg and Tains(1993)). This decomposition will be used for computing expectation of the above likelihood.

3. The Case of Exponential Distribution

In this section we consider the case where y follows an exponential distribution and the missing data mechanism is defined by

$$\Pr(R = r|y,\phi) = g(r|y,\phi) = \begin{cases} 1 - e^{-\frac{y}{\phi}}, & r = 0, \\ e^{-\frac{y}{\phi}}, & r = 1. \end{cases}$$

The missing probability is displayed in Figure 3.1. As Figure 3.1 shows the missing probability increases and tends to 1 as y becomes larger.

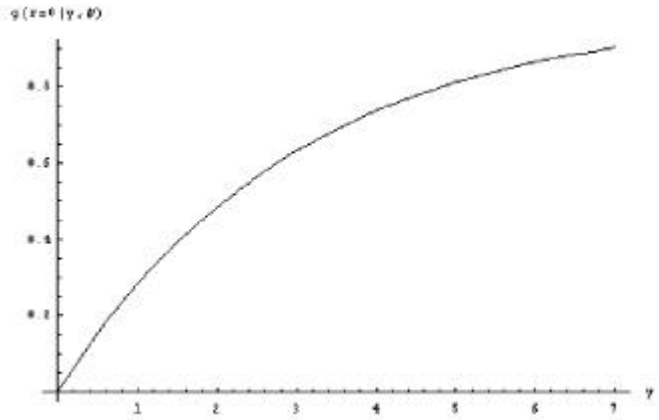

Figure 3.1 Assumed missing data mechanism $g(r = 0|y,\phi)$

Also suppose that the observations Y_n are obtained as a random sample from an exponential distribution, i.e.,

$$f(y|\theta) = \frac{1}{\theta} e^{-y/\theta}.$$

Then, the joint distribution $h(y,r|\theta,\phi)$ of y and r is obtained as

$$\begin{aligned} h(y,r|\theta,\phi) &= f(y|\theta)g(r|y,\phi) \\ &= \begin{cases} \frac{1}{\theta} e^{-y/\theta}(1 - e^{-y/\phi}), & r = 0, \\ \frac{1}{\theta} e^{-y/\theta} \cdot e^{-y/\phi}, & r = 1. \end{cases} \end{aligned}$$

and, therefore, the likelihood function with nonignorable missing data is given by

$$
\begin{aligned}
L(\theta,\psi|y_1,\cdots,y_m,r_1,\cdots,r_m,r_{m+1},\cdots,r_n) &= \prod_{i=1}^{m} h(y_i,r_i|\theta,\psi)\prod_{j=m+1}^{n} g_1(r_j|\theta,\psi) \\
&= \prod_{i=1}^{m}\frac{1}{\theta}e^{-y_i/\theta}\times e^{-y_i/\psi}\prod_{j=m+1}^{n}\frac{\theta}{\theta+\psi} \\
&= \left(\frac{1}{\theta}\right)^m\left(\frac{\theta}{\theta+\psi}\right)^{n-m}\exp\left(-\frac{\theta+\psi}{\theta\psi}\sum_{i=1}^{m}y_i\right)
\end{aligned}
$$

where

$$
\begin{aligned}
g_1(r_j|\theta,\psi) &= \int f(y_j|\theta)g(r_j=0|y_j,\psi)\,dy_j \\
&= \int_0^{\infty}\frac{1}{\theta}e^{-y_j/\theta}(1-e^{-y_j/\psi})dy_j \\
&= \frac{\theta}{\theta+\psi}.
\end{aligned}
$$

4. Estimation of Parameter Vectors θ and ψ

To obtain ML estimates $\hat{\theta}$ and $\hat{\psi}$ we can consider to apply a method. The method is the so-called EM(Expectation and Maximization) algorithm proposed by Dempster, Laird and Rubin (1977). The EM algorithm is a very general iterative algorithm for ML estimation in incomplete data problems. In the case of EM algorithm for models, missing sufficient statistics rather than individual observations need to be estimated at each iteration of the algorithm (see, McLachlan and Krishnan(1997), and Dodge(1985)). Even in the case of exponential distribution the formulation becomes complex where missing data mechanism exists.

Each iteration of EM algorithm consists of E step and M step and its construction is as follows.

Step 1. Set initial estimate $\theta^{(0)}$ and $\psi^{(0)}$ of θ, ψ, respectively.

Step 2. (E step) We compute the expected values of the (joint) sufficient statistics $\sum_{i=1}^{n} y_i$

$$
\begin{aligned}
E\left(\sum_{i=1}^{n} y_i \,|\theta^{(t)},\psi^{(t)},R,Y_{obs}\right) &= E\left(\sum_{i=1}^{m} y_i|\theta^{(t)},\psi^{(t)},R,Y_{obs}\right)+E\left(\sum_{j=m+1}^{n} y_j|\theta^{(t)},\psi^{(t)},R,Y_{obs}\right) \\
&= \sum_{i=1}^{m} y_i+\sum_{j=m+1}^{n}\hat{y}_j^{(t)}
\end{aligned}
$$

where

$$\begin{aligned}\hat{y}_j^{(t)} &= \int_0^{\infty} y_j f_1(y|r=0,\theta)\,dy_j \\ &= \int_0^{\infty} y_j \frac{1}{\hat{\theta}^{(t)}} e^{-y_j/\hat{\theta}^{(t)}} \left(1-e^{-y_j/\hat{\psi}^{(t)}}\right) \Big/ \frac{\hat{\theta}^{(t)}}{\hat{\theta}^{(t)}+\hat{\psi}^{(t)}}\,dy_j \\ &= \hat{\theta}^{(t)} + \frac{\hat{\theta}^{(t)}\hat{\psi}^{(t)}}{\hat{\theta}^{(t)}+\hat{\psi}^{(t)}} \quad (j=m+1,\cdots,n).\end{aligned}$$

Step 3. (M step) Calculate the estimate

$$\hat{\theta}^{(t+1)} = \left(\sum_{i=1}^{m} y_i + (n-m)\,\hat{y}_j^{(t)}\right) n$$

and solve the estimate $\hat{\psi}^{(t+1)}$, using the complete-data sufficient statistics $\sum_{i=1}^{n} y_i$ found in the E step.

Step 4. If converged, $\hat{\theta}^{(t)}$ and $\hat{\psi}^{(t)}$ are regarded as the ML estimates, i.e., $\hat{\theta}=\hat{\theta}^{(t)}$ and $\hat{\psi}=\hat{\psi}^{(t)}$. Otherwise go back to step 2 after putting $t := t+1$.

Numerical example. To illustrate our procedure we have generated a set of $n=100$ artificial data based on the model of exponential distribution with $\theta=1$ and on the missing data mechanism with $\psi=3.0$ about 30% missing. As shown in Table 4.1, 79 observations out of 100 are actually obtained. Values with asterisk are regarded as missing.

Table 4.2 shows the convergence of EM to this solution starting from the initial values $\theta^{(0)}=0.01$ and $\psi^{(0)}=0.01$. The EM algorithm have took 14 iterations. The iterative procedure is considered to be converged, when it holds that the Euclidean norm of the differences of successive two values of $\hat{\theta}^{(t)}$ and $\hat{\psi}^{(t)}$ is smaller than $\varepsilon=0.00001$. The obtained results are $\hat{\theta}=0.96715$ and $\hat{\psi}=2.99999$. To show the difference of converged values, we founded two parameters of standard error : $\mathrm{se}(\hat{\theta})=0.03285$ and $\mathrm{se}(\hat{\psi})=0.00001$. And the correlation coefficient of two estimated parameters was $\mathrm{corr}(\hat{\theta},\hat{\psi})=0.96703$. It is noted that in this case the estimated parameters are very close to the population parameters and that our procedures seem to work well.

Table 4.1: A set of 100 artificial data(exponential distribution with $\theta = 1$ and the missing value mechanism with $\psi = 3.0$)
Note. * : missing data

1	0.633907	0.062182	0.702323	0.134175	2.125615*
6	3.801593	0.263736	0.670557	1.096038	0.495486
11	0.208814	0.013400	0.702941	1.241645	0.253232
16	3.409731*	4.170736	1.234908	0.349487	0.062779
21	3.382418*	1.872162	0.769794	0.745341	0.040246
26	0.145197	0.582066*	0.083993	0.192054	0.858098
31	2.129653	3.625779*	2.632828*	0.623568	3.334371*
36	0.703712	1.680004	0.123132	3.007779	1.273178*
41	1.200620	1.676158*	1.258658*	0.192693	0.408614
46	0.005993	0.693403	3.443318*	0.478826	0.083737
51	2.524394	0.576381	0.741285	0.257411	0.661249
56	0.247183	0.850727	0.244805	0.096543	0.821901
61	0.954158	0.136173	0.295892	0.096641	0.922874*
66	0.372242*	0.022994	0.603037	4.252598*	0.178885
71	1.806685	0.606821	0.273151	0.657411	1.724596
76	1.227496	0.510398	0.605348	0.263653	2.014738*
81	3.636831*	0.897983	0.026007	2.898733	1.704346*
86	0.153316	1.783356	0.406625	1.981601*	1.466127
91	1.460142	0.656768	1.483078	0.857258*	1.926649
96	0.478107	2.636776	1.327901	0.010663	0.354965

5. Simulation Study

To discuss more in detail, we have simulated two cases to show the estimated accuracy for missing probability and sample size. The simulation processing is same to numerical example of previous section and we changed the missing probability and sample size. One of two is sets of $n = 100$, 400, and 1600 artificial data based on the exponential distribution with $\theta = 1$. For the other, we changed the missing data probability by three steps 10%($\psi = 12.0$), 30%($\psi = 3.0$), and 50%($\psi = 1.0$) for the generated data sets. We have found that the lower the ψ value is, the higher the missing data probability is.

And we have estimated two parameters $\hat{\theta}$ and $\hat{\psi}$ by using EM algorithm as the start points : $\theta^{(0)} = 0.01$ and $\psi^{(0)} = 0.01$. The processing method was equal to the numerical example. The obtained results are given in Table 5.1, it is same to the results of Table 4.2 for $n = 100$ and $\psi = 3.0$.

Table 4.2: Estimated parameters $\hat{\theta}$ and $\hat{\psi}$ with the EM algorithm

iteration	$\hat{\theta}$	$\lvert\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}\rvert$	$\hat{\psi}$	$\lvert\hat{\psi}^{(t+1)} - \hat{\psi}^{(t)}\rvert$
0	0.01	-	0.01	-
1	0.61465	0.60465	-0.30225	0.28775
2	0.84667	0.23202	0.41418	0.71643
3	0.92693	0.08026	1.55003	1.13585
4	0.95382	0.02689	2.38373	0.83370
5	0.96275	0.00893	2.77549	0.39176
6	0.96570	0.00295	2.92327	0.14778
7	0.96667	0.00097	2.97438	0.05111
8	0.96699	0.00032	2.99151	0.01713
9	0.96710	0.00011	2.99719	0.00568
10	0.96713	0.00002	2.99907	0.00188
11	0.96715	0.00000	2.99969	0.00062
12	0.96715	0.00000	2.99990	0.00021
13	0.96715	0.00000	2.99997	0.00007
14	0.96715	0.00000	2.99999	0.00002
15	0.96715	0.00000	2.99999	0.00000

It is shown that the convergence iterations of about 10% missing probability is fewer than those of 50% missing rate. And we can found that the lower the missing rate is, the fewer the iteration of convergency is. In the case of identical missing probability, it is notable that the larger the number of data is, the closer the estimation is to population parameters. The estimated accuracy of parameter ψ was very exact regardless of the missing rate and the data size.

In the present paper, we have shown some examples of statistical analysis based on the ML mehtods of the data with nonignorable misssing values and have studied the validity of such kinds of analysis. We can discuss on the properties of estimates.

As we known in simulation, in the case of the EM algorithm, if it find the ML estimates on the model including the missing-data mechanism, though the missing-data probability is higher, it can estimate accurately well. We have known the estimated accuracy is very close and lower missing probability, more the accuracy is higer, too.

Table 5.1: The estimate of parameters with missing probability changing

ψ	n	$\hat{\theta}$	se($\hat{\theta}$)	$\hat{\psi}$	se($\hat{\psi}$)	iterations
1.0	100	0.91198	0.08802	0.99998	0.00002	30
(50%)	400	1.05876	0.05876	0.99994	0.00006	26
	1600	1.00712	0.00712	0.99997	0.00003	26
3.0	100	0.96715	0.03285	2.9999	0.00001	15
(30%)	400	0.95617	0.04383	2.99984	0.00016	14
	1600	1.00214	0.00214	2.99982	0.00018	13
12.0	100	1.00287	0.00287	11.99997	0.00003	12
(10%)	400	0.95453	0.04547	11.99985	0.00015	7
	1600	0.98088	0.01912	11.99973	0.00027	7

References

1. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society*, B39, 1-38.
2. Dodge, Y. (1985). *Analysis of Experiments with Missing Data*, John Wiley & Sons, New York.
3. Hogg, R. V. and Tanis, E. A. (1993). *Probability and Statistical Inference 4th Edition*, Macmillan.
4. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
5. McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, John Wiley & Sons, New York.

[2002년 9월 접수, 2002년 9월 채택]