S-4

PERBANDINGAN MEKANISME DATA HILANG PADA MODEL NORMAL

[1]Zulhanif, [2]Yadi Suprijadi

[1]Department of of Statistics, FMIPA Universitas Padjadjaran, Bandung, Indonesia

[2]Department of Statistics, FMIPA Universitas Padjadjaran, Bandung, Indonesia

e-mail: [1]dzulhanif@yahoo.com, [2]yadi@bdg.centrin.net.id

Data hilang merupakan sutu fenomena yang umum terjadi dalam penelitian survei atau experimental, berdasarkan fakta tersebut berbagai metode statistika dikembangkan untuk mengatasinya. Pada makalah ini akan diteliti perbandingan nilai taksiran EM (*Expectation and Maximization*) algoritma untuk mekanisme data hilang *Missing at Random* (MAR), *Missing completely at random* (MCAR) dan *Missing Not at Random* (MNAR).

Kata kunci : Data hilang, EM Algoritma, Model normal, Mekanisme data hilang

I. Pendahuluan

Di dalam suatu penelitian survei atau eksperimental, seringkali terdapat masalah atau persoalan yang dapat menghambat suatu penelitian tersebut. Salah satunya adalah ketidaklengkapan data atau terdapat *missing value* (data hilang). Ketidaklengkapan suatu data mengakibatkan berbagai metoda statistika tidak dapat dipergunakan hal ini dikarenakan metoda statistika yang sudah ada berdasarkan pada data yang lengkap. Dalam prosedur statistika modern untuk data hilang, ketidak lengkapan suatu data di asumsikan mengikuti suatu mekanisme tertentu, Rubin(1976) mambangun tipologi mekanisme data hilang kedalam *Missing at Random* (MAR) jika mekanisme data hilang terdistribusi secara acak untuk sebagian unit observasi dan *Missing completely at random* (MCAR) jika mekanisme data hilang yang terdistribusi secara acak untuk seluruh unit observasi serta *Missing Not at Random* (MNAR) mekanisme data hilang yang tidak terdistribusi secara random. Gagasan untuk menangani data yang hilang pada penelitian ini berdasarkan asumsi distribusi normal multivariate (model normal). Distribusi normal multivariat merupakan asumsi yang secara luas dipergunakan dalam

berbagai metoda analisis data diantaranya analisis faktor, pemodelan struktural dan analisis diskriminan. Ketika suatu data multivariat tidak lengkap, hal ini akan menimbulkan kesukaran pada saat mengestimasi parameter populasinya seperti rata-rata dan matriks *varians-covarians* . Penaksiran matriks *varians-covarians* pada data yang hilang, didasarkan pada imputasi dari sekumpulan data yang lengkap, Imputasi ini didasarkan pada ekspektasi distribusi bersyarat dari data hilang terhadap data yang lengkap (Little and Rubin 2002,chapter 3.4). Metoda maksimum *likelihood* (ML) merupakan metode penaksiran parameter yang umum dipergunakan dalam menaksir parameter populasi akan tetapi metode ini akan sangat sukar jika diterapkan pada data yang tidak lengkap. Algoritma *expectation maximization* (EM) (Dempster et al. 1977) menjadi suatu metode alternatif untuk mengatasi kesukaran metode ML. Algoritma EM pada dasarnya merupakan metode pengoptimuman yang terdiri dari dua tahap yaitu: E(*expectation*)-*step* dan M(*maximum*)-*Step*. Metode EM yang dipergunakan dalam penelitian ini berdasarkan metode EM yang dikemukakan oleh Schafer(2002). Metode ini merupakan bentuk khusus dari metode EM yang mengasumsikan distribusi normal multivariat (model normal) pada distribusi bersyarat dari data hilang terhadap data yang lengkap. Pada makalah seminar ini akan diteliti perbandingan nilai taksiran algoritma EM (*Expectation and Maximization*) untuk mekanisme data hilang *Missing at Random* (MAR), *Missing completely at random* (MCAR) dan *Missing Not at Random* (MNAR) pada model normal.

II. Metodologi

Analisis mengenai *missing value* (data hilang) membantu menyelesaikan permasalahan yang disebabkan oleh data yang hilang/tidak lengkap. Data yang hilang akan memperkecil presisi dari perhitungan yang disebabkan oleh jumlah informasi yang lebih sedikit dari yang sudah ditetapkan di awal.

Missing value dapat diartikan sebagai data atau informasi yang "hilang" atau tidak tersedia mengenai subjek penelitian pada variabel tertentu akibat faktor *non sampling error*. Faktor non sampling error yang dimaksud adalah *interviewer recording error*, *respondent inability error*, dan *respondent unwillingness error*. *Interviewer recording error* terjadi akibat kealpaan petugas pengumpul data (pewawancara), misalnya ada sejumlah pertanyaan yang terlewatkan. *Respondent inability error* terjadi akibat ketidakmampuan responden dalam memberikan jawaban akurat, misalnya karena tidak memahami pertanyaan, bosan atau kelelahan (respondent fatigue) akhirnya responden mengosongkan sejumlah pertanyaan atau berhenti mengisi kuesioner di tengah jalan. *Unwillingness respondent error* tejadi karena responden tidak berkenan memberikan jawaban yang akurat, misalnya pertanyaan soal penghasilan, usia, berat badan, pengalaman melakukan pelanggaran hukum, dll. Seperti halnya pada respondent inability error, responden bisa mengosongkan jawaban atau menghentikan proses pengisian kuesioner.

Pada makalah ini peneliti mengasumsikan K buah variabel ($Y_1, Y_2, \ldots, Y_K$) yang berdistribusi normal multivariate dengan rata-rata $\mu = (\mu_1, \ldots, \mu_K)$ dan matriks *varians-covarians* $\boldsymbol{\Sigma}$. Jika suatu nilai pengamatan $Y = (Y_{obs}, Y_{mis})$, yang mana Y merupakan sampel acak berukuran n pada variabel $(Y_1, Y_2, \ldots, Y_K)$ dengan Y_{obs} adalah nilai pengamatan dari data yang lengkap dan Y_{mis} nilai pengamatan untuk data yang tidak lengkap. maka untuk keperluan membuat sebuah algoritma EM pada model normal, diberikan suatu statistik cukup S sbb:

$$S = \left(\sum_{i=1}^{n} y_{ij},\ j = 1, \ldots K \ \text{ and } \sum_{i=1}^{n} y_{ij} y_{ik} \quad j,k = 1, \ldots, K \right) \tag{1}$$

Pada langkah pertama E-*step* dari algoritma EM, menghitung nilai

$$E(\sum_{i=1}^{n} y_{ij} \left| \mathbf{Y}_{obs}, \theta^{(t)} \right.) = \sum_{i=1}^{n} y_{ij}^{(t)} \qquad j = 1, \ldots, K \tag{2}$$

$$E(\sum_{i=1}^{n} y_{ij} y_{ik} \left| \mathbf{Y}_{obs}, \theta^{(t)} \right.) = \sum_{i=1}^{n} \left(y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)} \right) \qquad j, k = 1, \ldots, K \tag{3}$$

untuk setiap nilai parameter $\theta^{(t)} = (\mu^{(t)}, \sum^{(t)})$ pada setiap iterasi ke-t .

yang mana

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{jika } y_{ij} \text{ taramati} \\ E(y_{ij} \left| y_{obs,i}, \theta^{(t)} \right.) & \text{jika } y_{ij} \text{ tidak teramati} \end{cases} \tag{4}$$

dan

$$C_{jki}^{(t)} = \begin{cases} 0 & \text{jika } y_{ij} \text{ atau } y_{ik} \text{ teramati} \\ Cov(y_{ij}, y_{ik} \left| y_{obs,i}, \theta^{(t)} \right.) & \text{jika } y_{ij} \text{ dan } y_{ik} \text{ tidak teramati} \end{cases} \tag{5}$$

Selanjutnya pada tahapan M-*step*, secara langsung nilai taksiran $\theta^{(t+1)}$ ditaksir dari statistik cukup untuk data yang lengkap sbb:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} y_{ij}^{(t)}}{n} \qquad j = 1, \ldots K \tag{6}$$

$$\sigma_{jk}^{(t+1)} = \frac{E\left(\sum_{i=1}^{n} (y_{ij} y_{ik} \left| \mathbf{Y}_{obs} \right.) \right)}{n} - \mu_j^{(t)} \mu_k^{(t)} \qquad j = 1, \ldots K \tag{7}$$

Proses iterasi pada algoritma EM ini berlangsung sampai nilai taksiran $\theta^{(t+1)}$ konvergen pada nilai tertentu. Untuk mempermudah perhitungan Algoritma EM pada data normal, pada penelitian ini akan dipergunakan makro program software R

III. Desain Simulasi

Untuk mengetahui performasi nilai taksiran dengan menggunakan algoritma EM dengan mekanisme data hilang hilang *Missing at Random* (MAR), *Missing completely at random* (MCAR) dan *Missing Not at Random* (MNAR) dilakukan prosedur simulasi sbb:

1. Bangkitkan 50 data berdistribusi normal dengan rata-rata $\mu = (125,125)$ $\sigma_X = \sigma_Y = 25$ dan $\rho = 0.6$.
2. Tetapkan jumlah data yang hilang
3. Buatlah mekanisme data hilang untuk :
   a) MCAR dengan menghapus nilai data yang dibangkitkan pada langkah 1 untuk variabel X dan Y dengan peluang p
   b) MAR dengan menghapus nilai data yang dibangkitkan pada langkah 1 pada variabel Y dengan peluang $p(Y \text{ missing}) = \frac{1}{1+\exp(0.6+0.6(X-25))}$, dan variabel X $p(X \text{ missing}) = \frac{1}{1+\exp(0.6+0.6(Y-25))}$
   c) MNAR dengan menghapus nilai data yang dibangkitkan pada langkah 1 pada variabel Y dengan peluang $p(Y \text{ missing}) = \frac{1}{1+\exp(0.6+0.6(Y-25))}$, dan variabel X dengan peluang $p(X \text{ missing}) = \frac{1}{1+\exp(0.6+0.6(X-25))}$,
4. Dari data hilang yang dibangkitkan untuk setiap mekanisme data hilang lakukan penaksiraan parameter yang berkenaan.
5. Ulangi proses ini sampai dengan 1000 kali

IV. Hasil Simulasi

Dengan mempergunakan R software, hasil simulasi untuk nilai taksiran algoritma EM (*Expectation and Maximization*) untuk mekanisme data hilang *Missing at Random* (MAR), *Missing completely at random* (MCAR) dan *Missing Not at Random* (MNAR) disajikan dalam tabel 1 sbb:Tabel 1

Nilai taksiran Algoritma EM dengan jumlah data yang hilang *m=10* untuk replikasi sebanyak 1000 dengan ukuran sampel *N=50*

Parameter	MCAR	MAR	MNAR
μ_Y =125.0	124.9	125.3	151.6
	(6.53)	(17.2)	(26.9)
$\sigma_Y = 25$	25.9	28.7	13.6
	(5.93)	(8.24)	(12.1)
$\rho = 0.6$	0.57	0.45	0.35
	(0.19)	(0.37)	(0.36)
$\beta_{Y\|X}$ = 0.6	0.61	0.59	0.21
	(0.27)	(0.52)	(0.43)
$\beta_{X\|Y}$ =0.6	0.56	0.39	0.66
	(0.22)	(0.38)	(0.56)

Dari tabel 1 dapat disimpulkan bahwa hasil taksiran yang baik didapat untuk mekanisme data hilang (MCAR) dan *Missing Not at Random*, ini ditunjukkan dari nila taksiran yang mendekati nilai parameternya dan lebar taksiran yang paling kecil

V. Kesimpulan dan Saran

Algoritma EM pada model normal merupakan salah satu metoda untuk mengatasi permasalahan data yang hilang untuk data multivariat. Metoda ini pada dasarnya merupakan metoda pengoptimuman dua tahap dari fungsi likelihood dengan cara menghitung ekspektasi bersyarat pada tahapan E-*Step* dan mencari Taksiran maksimum likelohoodnya pada thap M-*step*. Hasil simulasi menunjukkan mekanisme data hilang (MCAR) *Missing Not at Random* memiliki tingkat performasi nilai taksiran yang paling baik jika dibandingkan dengan mekanisme data hilang *Missing at Random* (MAR), dan *Missing Not at Random* (MNAR)

Ucapan Terima Kasih

Terima kasih penulis sampaikan kepada Jurusan Statistika FMIPA Universitas Padjadjaran bandung yang telah mendanai penelitian ini melalui **Program Riset Mandiri (PRM)** tahun anggaran 2009.

Daftar Pustaka

1. Dempster, A.P.,Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-38.
2. Little, R.J.A and Rubin, D.B.(2002) *Statistical Analysis with Missing Data*. J. Wily & Sons, New York.
3. Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592
4. Schafer, J.L. (1992) Analysis of Incomplete Multivariate Data. *Monographs on Statistics and Applied Probability,* 72.