ISBN : 978.602.361.002.0

METODE KEMUNGKINAN MAKSIMUM UNTUK KOREKSI SEBARAN BERSYARAT PADA ANALISIS KORELASI

Restu Arisanti
Departemen Statistika FMIPA Universitas Padjadjaran
restu.arisanti@unpad.ac.id

ABSTRAK. Masalah umum pada pendugaan validitas di bidang pendidikan adalah sebaran bersyarat pada peubah bebas. Penggunaan metode yang tepat untuk penentuan sebaran bersyarat sangat diperlukan untuk menentukan validitas instrumen yang digunakan pada penelitian. Tujuan dari penulisan ini adalah mengetahui kegunaan pendugaan kemungkinan maksimum (ML) yang diperoleh dari algoritma nilai harapan maksimum (EM) dengan membandingkan korelasi dari sampel bersyarat dan sampel tidak bersyarat dengan melibatkan metode imputasi. Sampel bersyarat pada penelitian ini adalah kondisi dimana siswa yang mengambil ujian praktek harus lulus terlebih dahulu pada ujian tertulisnya. Hasil dari penelitian ini menunjukkan metode kemungkinan maksimum yang berasal dari algoritma EM dapat digunakan dalam pendugaan korelasi suatu populasi.

Kata Kunci:*Metode kemungkinan maksimum (ML); algoritma EM; Korelasi.*

1. PENDAHULUAN

Masalah umum yang sering terjadi pada pendugaan validitas instrumen pada penelitian di bidang pendidikan adalah adanya sebaran bersyarat pada kriteria peubah. Hal ini terjadi pada saat peneliti ingin menduga korelasi antara dua peubah (dalam hal ini X dan Y) dari sebuah populasi, tetapi subjek yang dipilih pada peubah X dan data padapeubah Y hanya berasal dari sampel yang terpilih (Raju&Brand [7]). Sebagai contoh pada seleksi penerimaan siswa, nilai ujian masuk digunakan untuk menduga tingkat kesuksesan di bidang akademik dengan membandingkan hasil nilai akademik siswa tersebut. Oleh karena seleksi penerimaan siswa ini dibuat berdasarkan nilai dari berbagai instrumen yang dibuat, maka terdapat batasan nilai bersyarat pada sampel. Meskipun korelasi antara nilai ujian dan kesuksesan akademik dapat diperoleh dari sampel bersyarat namun korelasi dari populasi siswa tidak diketahui.

Pada penulisan ini, korelasi pada sampel tidak bersyarat akan diuji dengan pendekatan ML dari algoritma EM untuk koreksi sebaran bersyarat. Dengan pendekatan ini, akan dilihat data hilang pada peubah dengan sebaran bersyarat dan menduga nilai data yang hilang sebelum menduga korelasi. Untuk hal khusus pada data hilang, akan menggunakan metode statistika seperti yang dijelaskan pada Little & Rubin [6]. Asumsikan X adalah peubah yang diketahui pada semua pengujian dan Y adalah peubah yang ingin diketahui dengan data hilang pada beberapa pengujian. Terdapat tiga kondisi data hilang; MCAR, MAR dan MNAR. MCAR adalah *Missing Completely At Random yaitu* jika sebaran data hilang tidak tergantung pada objek yang diamati. Dengan kata lain, peluang hilangnya data dalam peubah Y tidak berhubungan terhadap peubah Xdan Y. MAR adalah *Missing At Random* yaitu

ISBN : 978.602.361.002.0

peluang hilangnya data dalam Y berhubungan dengan peubah X tetapi tidak berhubungan terhadap peubah Y. MNAR berarti *Missing Not At Random* yaitu peluang hilangnyaY berhubungan dengan nilai yang tidak teramati pada peubah Y itu sendiri, sehingga tidak dapat diprediksi karena tidak ada informasi tentang *missing value* tersebut. Untuk pendugaan pada kasus MCAR dan MAR dapat menggunakan metode imputasi untuk menggantikan nilai yang hilang. Pada penelitian ini, akan digunakan pendekatan MAR dengan metode imputasi sebagai alat analisis pada data pengamatan yang hilang dan dengan menggunakan sruktur dan output yang sama.

Terdapat beberapa teknik imputasi untuk menggantikan data pengamatan yang hilang. Teknik yang biasa digunakan adalah imputasi Hot-Deck, Cold-Deck, Regresi dan multiple imputasi (Sarndal et.al.,[8]). Secara umum, imputasi dapat mengakibatkan distorsi pada sebaran sebuah peubah atau hubungan antara dua atau lebih peubah. Sebagai contoh, Gustafsson& Reuterberg [3] menggunakan regresi untuk mengimputasi nilai hilang untuk mendapatkan gambaran yang realistis mengenai hubungan antara tingkat ranking dengan pencapaian ujian skolastik. Regresi imputasi menghasilkan keragaman yang rendah dari peubah X dan Y. Dengan kata lain, korelasi akan bernilai 1.0 jika dihitung dengan nilai imputasi. Oleh karena itu, pada penelitian ini akan digunakan pendugaan kemungkinan maksimum pada data hilang yang diperoleh dari algoritma nilai harapan maksimum (EM/*Expectation Maximization*).

2. METODE PENELITIAN

Metode Kemungkinan Maksimum

Metode kemungkinan maksimum adalah tehnik yang sangat luas dipakai dalam pendugaan suatu parameter distribusi data dan tetap dominan dipakai dalam pengujian uji-uji yang baru (Dempster et al.[1]).Metode ini merupakan alternatif bagi metode kuadrat terkecil dengan memaksimumkan fungsi kemungkinan (*likelihood)* atau (*log-likelihood)*. Fungsi kemungkinan model linier adalah

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2$$

Metode kemungkinan maksimum menduga parameter β dan σ^2dengan mendapatkan nilai parameter β dan σ^2 yang memaksimumkan L (β, σ^2)(Samson et al. [11]).

Algoritma EM (*Expectation Maximization*)

Algoritma EM merupakan sebuah metode optimasi iteratif yang digunakan untuk mendapatkan dugaan kemungkinan maksimum parameter model-model peluang, dimana model peluang tergantung pada peubah laten yang tidak teramati atau terdapat data hilang. Tiap iterasi dari algoritma EM terdiri dari dua proses yaitu tahap E dan tahap M. Tahap E adalah tahap pendugaan yaitu menghitung dugaan kemungkinan dengan mempertimbangkan peubah laten jika peubah tersebut diamati. Tahap E bertujuan memperoleh nilai harapan bersyarat dari data hilang dengan syarat data yang diketahui nilainya dan penduga

ISBN : 978.602.361.002.0

parameternya, kemudian mensubstitusikan nilai ekspektasi yang diperoleh terhadap data hilang. Tahap M adalah adalah tahap maksimisasi yaitu menghitung dugaan kemungkinan maksimum parameter dengan maksimisasi dugaan kemungkinan pada tahap E. Parameter didapatkan pada tahap M selanjutnya digunakan untuk tahap E berikutnya, dan proses tersebut dilakukan secara berulang.

Algoritma EM ini hampir mirip dengan pendekatan *ad hoc* untuk proses pendugaan dengan data hilang yaitu (1) mengganti *missing value* dengan nilai pendugaan, (2) menduga parameter, (3) menduga ulang *missing value* sebelumnya dengan menggunakan parameter baru yang diduga, (4) menduga ulang parameter, dan seterusnya berulang-ulang sampai dengan konvergen terhadap suatu nilai. Ide dasar dari Algoritma EM adalah mengasosiasikan suatu data lengkap dengan data yang tidak lengkap dengan tujuan agar secara komputasi menjadi lebih mudah.

Asumsikan Y adalah data tidak lengkap yang terdiri dari nilai peubah teramati, dan misalkan m menunjukkan data hilang. Y dan m secara bersama membentuk data lengkap. Misalkan p adalah fungsi kepekatan peluang bersama dari data tidak lengkap dengan parameter diberikan oleh vektor $\theta: p(y, m|\theta)$. Sebaran bersyarat data hilang jika diketahui data amatan dapat dinyatakan sebagai berikut:

$$p(m|y,\theta) = \frac{p(y,m|\theta)}{p(y|\theta)} = \frac{p(y|m,\theta)p(m|\theta)}{\int p(y|\hat{m},\theta)p(\hat{m}|\theta)d\hat{m}}$$

Algoritma EM secara iteratif akan meningkatkan dugaan awal θ_0 dengan mencari dugaan baru $\theta_1, \theta_2, \ldots, \theta_n$. Untuk setiap tahapan yang menurunkan θ_{n+1} dari θ_n dengan persamaan berikut:

$$\theta_{n+1} = \arg\max R(\theta)$$

Dimana $R(\theta)$ adalah nilai harapan *log-likelihood*. $R(\theta)$ diperoleh dari

$$R(\theta) = \sum_m p(m|y,\theta_n) \log p(y,m|\theta)$$

atau $R(\theta) = E_\theta[\log p(y,m|\theta)|y]$

sehingga dapat dikatakan θ_{n+1} adalah nilai yang memaksimumkan (M) dugaan bersyarat (E) dari log-likelihood data lengkap jika diketahui peubah teramati pada nilai parameter sebelumnya. (Deylon et al.[2]; Makowski & Lavielle [4]).

Secara ringkas Algoritma EM dapat ditulis sebagai berikut:

- Tahap E: Pendugaan statistik cukup untuk data lengkap Y dengan cara menghitung nilai harapannya.
- Tahap M: menentukan θ_{n+1} dengan metode kemungkinan maksimum dari Y.

ISBN : 978.602.361.002.0

- Iterasi sampai nilai θ_n konvergen, atau θ_{n+1}- θ_n mendekati nol.

Data

Data yang digunakan pada penelitian ini adalah data nilai ujian tertulis dan ujian praktek dari sebuah Sekolah Menengah Pertama. Nilai untuk ujian tertulis berasal dari soal pilihan berganda dan pada ujian praktek nilai menggunakan rubrik penilaian dengan skala 1-4. Sampel diambil dengan metode sampel acak. Sampel ini terbagi menjadi dua yaitu "sampel bersyarat" dan "sampel tidak bersyarat". Sampel tidak bersyarat adalah siswa yang lulus atau gagal pada ujian tertulis namun mengikuti ujian praktek. Sedangkan sampel bersyarat adalah siswa yang lulus ujian tertulis kemudian mengikuti ujian praktek. Pada sampel bersyarat terdapat 68.1% lulus ujian tertulis dan 61.3% lulus ujian praktek.

Hubungan antara hasil ujian tertulis dan praktek diasumsikan linear dimana secara analisis nilai ujian tertulis berbandinglurus dengan ujian prakteknya. Untuk mengarah pada sebaran bersyarat pada sampel bersyarat akan digunakan pendugaan korelasi dari sampel tidak bersyarat yang dinotasikan dengan $\hat{\rho}_{XY}$. Rumus korelasi sampel bersyarat dan simpangan baku peubah bebas X pada sampel bersyarat dan sampel tidak bersyarat (Sackett & Yang [9]) dituliskan sebagai berikut:

$$r_{XY} = S_X r_{xy} / (S_X^2 r_{xy}^2 + s_x^2 - s_x^2 r_{xy}^2)^{1/2} \qquad (1)$$

dimana r_{XY} adalah korelasi dugaan terkoreksi antara X dan Y pada sampel tidak bersyarat yang berasal dari sampel bersyarat; r_{xy} adalah korelasi yang diamati dari peubah X dan Y pada sampel bersyarat; s_x adalah dugaan simpangan baku peubah X pada sampel bersyarat; S_x adalah dugaan simpangan baku peubah X pada sampel tidak bersyarat.

Persamaan 1 menunjukkan pendugaan korelasi dengan asumsi regresi Y terhadap X linier dan bersifat homoskedastis. Imputasi EM digunakan untuk memperolah nilai yang hilang pada ujian praktek. Dengan kata lain, akan dibentuk sampel sebaran bersyarat dengan memindahkan nilai ujian praktek pada nilai siswa yang tidak lulus ujian tertulis sehingga pengamatan ini dilihat sebagai data hilang. Dengan menggunakan definisi data hilang, maka dapat diasumsikan menjadi MAR karena peluang hilangnya nilai ujian praktek Y dihubungkan dengan kelulusan ujian tertulis X tetapi bukan untuk nilai ujian praktek. Y. Pendugaan kemungkinan maksimum menggunakan algoritmanilai harapan maksimum (EM) yang diimputasikan pada nilai ujian praktek yang gagal pada ujian tertulis. Little [5] dan Dempster et.al. [1] menunjukkan pendugaan kemungkinan maksimum bersifat konsisten dimana konvergensi peluang mengarah pada parameter populasi. Persamaan data hilang Y yang diimputasi adalah sebagai berikut:

$$Y_{imp} = \hat{\alpha}_{EM} + \hat{\beta}_{EM} X$$

ISBN : 978.602.361.002.0

Dimana $\hat{\alpha}_{EM}$ dan $\hat{\beta}_{EM}$ adalah pendugaan yang diperoleh dari iterasi akhir dari algoritma EM. Menurut Scheffer [10] pada saat pendugaan kemungkinan maksimum diperoleh dari algoritma EM maka penggunaan imputasi EM pada pengujian data hilang akan valid.

3. HASIL PENELITIAN DAN PEMBAHASAN

Pengujian pada sampel tidak bersyarat mengindikasikan bahwa secara signifikan terdapat hubungan positif antara nilai ujian tertulis dan ujian praktek $(\hat{\rho}_{XY} = 0.63)$. Sampel bersyarat menunjukkan hubungan yang lebih kuat antara nilai ujian tertulis dan ujian praktek $(\hat{r}_{xy} = 0.46)$. Langkah berikutnya adalah menggantikan nilai data hilang dengan pendugaan kemungkinan maksimum dari algoritma EM dengan menghasilkan korelasi dugaan $\hat{r}_{EM} = 0.63$ (dapat dilihat pada tabel 1).

Tabel 1. Korelasi Dugaan Terkoreksi Antara X dan Y Pada Sampel Tidak Bersyarat yang berasal dari Sampel Bersyarat

	s_x	S_x	$\hat{r}_{xy}$	$\hat{r}_{XY}$	$\hat{r}_{EM}$	$\hat{\rho}_{XY}$
Dugaan Korelasi	2.73	3.56	0.46**	0.56	0.63**	0.63**

*p < 0.05; ** p < 0.01

Persamaan yang digunakan untuk koreksi sebaran bersyarat menghasilkan hasil korelasi yang baik pada sampel tidak bersyarat ($\hat{r}_{XY} = 0.56$). Dengan pendekatan kemungkinan maksimum dengan algoritma EM diperoleh dugaan korelasi $\hat{r}_{EM} = 0.63$.Hasil penelitian mengindikasikan bahwa korelasi dengan menggunakan metode kemungkinan maksimum yang diperoleh dari algoritma EM menghasilkan korelasi yang baik pada sampel tidak bersyarat. Algoritma EM cenderung mudah diterapkan karena bersandarkan pada perhitungan data lengkap dan cenderung lebih stabil secara numerik, dimana dalam setiap iterasi *loglikelihoodnya* meningkat. Selain itu, algoritma EM biasa digunakan untuk menduga nilai dari data hilang. Penggunaan metode yang tepat untuk koreksi sebaran bersyarat sangatlah penting pada saat menduga validitas instrumen yang digunakan seperti contohnya pada seleksi penerimaan mahasiswa baru dan penerimaan karyawan. Penggunaan metode yang tidak tepat untuk koreksi sebaran bersyarat akan menghasilkan ketidakvalidan pada instrumen penelitian.

4. SIMPULAN

Hasil perhitungan di atas mengindikasikan bahwa korelasi dengan menggunakan metode kemungkinan maksimumyang diperoleh dari algoritma EM menghasilkan korelasi yang sangat signifikan pada sampel tidak bersyarat. Masih banyak pertanyaan yang berkaitan dengan sebaran bersyarat untuk penelitian berikutnya seperti contohnya bagaimana jika hubungan antara peubah-peubah tidak linier. Berkaitan dengan imputasi EM, berapa banyak kasus yang dapat diimputasi pada waktu yang sama untuk mendapatkan dugaan korelasi populasi yang tepat.

ISBN : 978.602.361.002.0

DAFTAR PUSTAKA

[1] Dempster, A.P., Laird, N. M. & Rubin, D.B. 1997. *Maximum Likelihood from Incomplete Data Via the EM Algorithm.* Journal of the Royal Statistical Society, Series B(39), 1-38.

[2] Deylon, B., Lavielle, M., and Moulines, E. 1999. *Convergenceof a Stochastic Approximation Version of the EM Algorithm.* Annals of Statistics, 27, 94-128.

[3] Gustafsson, J.E.,& Reuterberg,S.E. 2000. *Methodological Problems Associated with the Study of the Predictive Validity of the SweSAT.* 273-283.

[4] Makowski, D., and Lavielle, M. 2006. Using SAEM to Estimate Parameters of Models of Response to Applied Fertilizer. Journal of Agricultural, Biological, and Environmental Statistics. 11: 45-60.

[5] Little,J.A. 1992. *Regression with Missing X's:* a review. Journal of American Statistical Association, 87, 1227-1237.

[6] Little, J.A. & Rubin, D.B. 2002. *Statistical Analysis with Missing Data.* Edisi 2. Hoboken, NY: Jhon Wiley and Sons.

[7] Raju, N.S.,& Brand, P.A. 2003. *Determining the Significance of Correlations Corrected for Unreability and Range Restriction.* Applied Psychological Measurement, 85, 112-118.

[8] Sarndal, C.E., Swensson,B., &Wretman,J. 1992. *Model Assisted Survey Sampling.* New York: Springer-Verlag.

[9] Sackett, P.R.,& Yang, H. 2000. *Correction for Range Restriction: an Expanded Typology.* Journal of Applied Psychology, 85, 112-118.

[10]Scheffer, J. 2002. *Dealing with Missing Data.* Res Lett Inf Math Sci,3, 153-160.

[11]Samson, A.,M. LAvielle, & F. Mentre. 2000. *The SAEM Algorithm for Group Comparison Tests in Longitudinal Data Analysis Based on Linear Mixed-Effects Model Statistics.* 1-6.