ISBN : 978.602.361.002.0

KAJIAN METODE IMPUTASI DALAM MENANGANI MISSING DATA

Triyani Hendrawati
Staf Pengajar Statistika Universitas Padjadjaran
triyani.hendrawati@gmail.com

ABSTRAK.
Pada sebuah survey, adakalanya tidak semua pertanyaan pada kuisioner dijawab atau diisi dengan lengkap oleh responden. Hal ini menyebabkan adanya *missing* data. *Missing* data akan mengakibatkan pendugaan parameter menjadi tidak tepat karena berkurangnya ukuran data. Telah dikembangkan beberapa metode untuk meminimalkan dampak negatif dari data hilang. Pada penelitian ini diambil sebuah contoh kasus penilaian mahasiswa terhadap dosen. Terhadap data ini, dilakukan analisis terhadap data set yang utuh. Kemudian dilakukan penghilangan data utuh sebesar 4,9% secara acak, sehingga diperoleh *missing* data. Kemudian *missing* data ini dianalisis menggunakan beberapa metode yaitu dengan menghapus data yang tidak lengkap, selain itu menggunakan metode imputasi. Metode imputasi yang dilakukan yaitu pertama menginput *missing* data dengan suatu nilai konstan, yang kedua dengan metode Hot Deck. Metode imputasi Hot Deck memberikan hasil yang lebih baik bila dibandingkan dengan menghapus data yang tidak lengkap maupun bila dibandingkan metode imputasi dengan nilai konstan. Besar nilai persentase kesalahan relatif berkaitan erat dengan banyaknya item yang diinput, semakin banyak item yang diinput maka semakin besar nilai persentase kesalahan relatif.

Kata Kunci: *missing* data; metode imputasi

1. PENDAHULUAN

Dalam sebuah survey atau sensus sering ditemukan adanya *missing* data /data hilang. Hal ini tidak diharapkan oleh peneliti, hal ini dikarenakan dengan adanya *missing* data maka hasil observasi tidak dapat dianalisis dengan baik. *Missing* data dapat menyebabkan pendugaan parameter menjadi tidak efisien karena berkurangnya ukuran data. Maka diperlukan metode yang dapat meminimalkan akibat negatif dari *missing* data. Banyak penelitian yang sudah dilakukan seperti Rubin DB [6] yaitu mengembangkan inferensia dari *missing* data, Dempster et al., [1] membahas Maksimum *Likelihood* dari *missing* data dengan Algoritma EM. Little & Rubin [4] telah melakukan penelitian terhadap missing data, mereka melakukan penanganan data hilang dengan beberapa prosedur, yaitu: amatan lengkap, imputasi, pembobotan, dan model.

ISBN : 978.602.361.002.0

Pada penelitian ini akan dibahas beberapa metode dalam menangani *missing* data. Pada contoh kasus penelitian ini, dilakukan analisa terhadap data set yang utuh. Kemudian dilakukan penghilangan data sehingga diperoleh *missing value*. Selanjutnya data yang mengandung *missing value* ini dianalis secara lagsung, selain itu juga dilakukan analisis dengan metode imputasi. Metode imputasi yang dilakukan yaitu pertama menginput *missing* data dengan suatu nilai konstan, yang kedua dengan metode Hot Deck.

2. METODE PENELITIAN

Missing data dapat disebabkan oleh berapa hal, diantaranya yaitu penolakan dari responden untuk menjawab beberapa pertanyaan karena pertanyaanya sangat pribadi/rahasia; kesalahan pada saat pengumpulan data, misalnya pertanyaan terlewat sehingga tidak memperoleh jawaban; selain itu *missing* data dapat juga diakibatkan oleh kesalahan pada saat *entri* data.

Little dan Rubin [4] membagi tiga tipe *missing* data berdasarkan mekanisme berikut ini:

- *Missing Completely at Random* (MCAR), terjadinya *missing* data tidak berkaitan dengan nilai semua variabel, apakah itu variabel dengan *missing* data atau dengan variabel pengamatan. Hal ini berarti *missing* data terjadi secara acak.
- *Missing at Random* (MAR), terjadinya *missing* data hanya berkaitan dengan variabel respon/pengamatan. Contohnya seseorang yang memiliki rasa waswas yang tinggi cenderung tidak akan melaporkan pendapatan mereka, rasa waswas akan berhubungan pada pelaporan pendapatan. Namun, peluang penderita rasa waswas sendiri untuk melaporkan pendapatan tidak berhubungan dengan tingkat pendapatan, maka data dapat digolongkan dengan MAR.
- *Not Missing at Random* (NMAR), terjadinya *missing* data pada suatu variabel berkaitan dengan variabel itu sendiri, sehingga ini tidak bisa diprediksi dari variabel lain pada suatu dataset.

Penanganan *missing* data dapat dilakukan dengan menghapus data yang tidak lengkap. Jika data yang tidak lengkap jumlahnya relatif kecil, dibandingkan dengan keseluruhan data, menghapus data yang tidak lengkap merupakan salah satu pendekatan yang masuk akal. Tetapi umumnya hal ini akan memberikan kesimpulan yang valid, hanya terjadi ketika *missing* data secara acak, dalam arti bahwa probabilitas respon tidak tergantung pada nilai-nilai data yang diamati atau hilang. Dengan kata lain, penghapusan secara implisit mengasumsikan bahwa kasus yang dibuang seperti subsampel acak [4].

Penghapusan data hilang ini tentunya memberikan hasil pendugaan yang kurang baik dikarenakan beberapa alasan, yaitu :

- Jika melakukan penghapusan data hilang maka ukuran contoh yang terambil akan berkurang, sehingga akan menyebabkan berkurangnya ketepatan dalam pendugaan
- Jika individu yang dikeluarkan ternyata hasilnya sangat berbeda dari data yang lain maka akan mengahasilkan penduga yang bias.

Sebagai solusi yang tepat agar ukuran sampel tidak berkurang dalam mempertahankan *statistical power*, maka jika terdapat *missing* data dalam survey adalah dengan melakukan imputasi. Imputasi, yaitu proses pengisian atau penggantian *missing values* pada dataset

ISBN : 978.602.361.002.0

dengan nilai-nilai yang mungkin berdasarkan informasi yang didapatkan pada dataset tersebut. Beberapa metode imputasi yang dapat digunakan adalah:

- Mengganti data hilang dengan suatu nilai konstan
  Metode imputasi yang paling umum dan paling mudah untuk digunakan adalah mengganti *missing* data dengan nilai rata-rata atau dengan modus tergantung dari jenis datanya. Pada data numerik digunakan cara mengganti *missing* data dengan nilai rata-rata, sedangkan untuk data kategorik maka digunakan cara mengganti *missing* data dengan nilai modus. Metode ini menghasilkan penduga rata-rata untuk peubah akan sama dengan nilai yang diimputasikan. Keunggulan metode ini adalah mengisi nilai *missing* data dengan nilai harapan yang secara relatif mempunyai tingkat kestabilan yang tinggi. Sedangkan kelemahannya adalah ragam yang diperoleh dengan metode ini tidak sesuai dengan data yang sebenarnya dan korelasi antar peubah dapat memberikan informasi yang menyesatkan. Metode ini akan menyebabkan pendugaan error yang selalu lebih rendah dari sebenarnya (underestimate error). Sehingga metode ini tidak disarankan untuk digunakan [5].

- Metode Hot Deck
  Metode ini merupakan penyempuranaan dari metode mengganti *missing* data dengan nilai rata-rata khususnya pada pendugaan standar error yang underestimate. Sebelum menggunakan metode ini, terlebih dahulu data diurutkan berdasarkan variabel yang dinilai terkait dengan variabel yang teradapat item *missing* data. Individu yang berada pada kluster yang sama maka ditempatkan pada file yang sama. Proses metode ini yaitu pertama menetapkan nilai demografi yang terpilih atau peubah lainnya. Nilai yang hilang akan diganti dengan nilai data sebelumnya setelah data disusun urut dan ditetapkan nilai pencirinya (demografinya). Kelemahan dari Hot Deck adalah jika *missing* data banyak mengakibatkan dalam pengisian nilainya akan berulang-ulang sehingga pendugaannya akan berbias [2].

- Metode Regression
  Pada metode ini, *missing* data diperoleh dengan melakukan prediksi menggunakan regresi. Banyak macam model regresi yang dapat digunakan dalam regresi imputasi misalnya regresi linear, regresi logistik.
  Misalkan variabel Y diperoleh dari data dengan *missing* data dan variabel Z diperoleh dari data lengkap. Jika Y dan Z berkaitan, dapat diprediksi nilai Y. Misalkan Y dan Z adalah terkait dengan model $Y = f(Z) + \varepsilon$, dimana f adalah fungsi, seperti $f(Z) = \beta_0 + \beta_1 z$, dan ε adalah variable random. Jika diketahi fungsi $\hat{y}_j = f(z_j)$ maka akan diperoleh nilai untuk menginput *missing* data y_j .

- Metode EM (*Expectation Maximisation*)
  Metode *Expectation Maximisation* adalah metode yang digunakan untuk memperkirakan parameter populasi yang tidak diketahui. Metode EM menggunakan prosedur iterative untuk menghitung statistik cukup dan menduga parameter. Algoritma EM dirumuskan untuk membuat model yang sesui dengan maksimum likelihood (ML). Pada algoritma EM dilakukan pengisian nilai-nilai yang hilang terlebih dahulu, kemudian menemukan kemungkinan maksimumnya. Proses ini diulang terus sampai menghasilkan pendunga parameter[1].

ISBN : 978.602.361.002.0

Sebuah fungsi likelihood $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ dimana X adalah himpunan data observasi, Z adalah himpunan *missing* value, dan $\boldsymbol{\theta}$ suatu parameter yang tidak diketahui, *Maksimum Likelihood Estimate* (MLE) dari parameter yang tidak diketahui dinyatakan dengan *marginal likelihood* data observasi

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Algoritm EM digunakan untuk mencari nilai MLE dari *marginal likelihood* dengan menggunakan iterasi dua langkah berikut:

1. Langkah Ekspektasi (E step)
   menentukan nilai ekspektasi dari fungsi log likelihood berdasarkan nilai estimasi parameter$\boldsymbol{\theta}^{(t)}$.
   $$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}} [\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$
2. Langkah Maximisasi (M step)
   Mencari parameter yang akan memaksimumkan
   $$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

- Metode *Multiple imputation*
  Pada metode ini setiap *missing* data tidak diduga melalui nilai tiruan, tetapi diperoleh dengan merepresentasikan sampel random dari nilai-nilai hilang. Pada metode ini diperoleh m dataset yang lengkap, kemudian masing-masing data set dianalisis dengan metode data lengkap. Kemudian, hasil yang diperoleh dari m dataset ini digabungkan. Proses ini menghasilkan inferensi yang valid secara statistik yang mencerminkan ketidakpastian akibat nilai-nilai yang hilang tersebut [6].

3. HASIL PENELITIAN DAN PEMBAHASAN

Contoh kasus berikut merupakan data primer yang diambil pada Februari 2013 mengenai penilaian mahasiswa terhadap seorang dosen pada mata kuliah Analisis Numerik di FIP UMJ. Kuisioner diberikan kepada mahasiswa setelah selesai ujian akhir semester (UAS). Seluruh pertanyaan pada kuisioner berjumlah 34 butir yang dikelompokkan menjadi 5 aspek. Lima aspek tersebut yaitu: Kompetensi Pedagogik (PDG) berisi 9 butir pertanyaan, Kompetensi Profesional (PRF) berisi 8 butir pertanyaan, Kompetensi Kepribadian (PBD) berisi 6 butir pertanyaan, Kompetensi Sosial (SOS) berisi 5 butir pertanyaan, Kompetensi Al-Islam dan Kemuhammadiyahan (MUH) berisi 6 butir pertanyaan. Jawaban kuisioner berupa angka 1-5, dimana

1 = sangat tidak baik/sangat rendah/tidak pernah
2 = tidak baik/ rendah/jarang
3 = biasa/cukup/kadang-kadang
4 = baik/tinggi/sering
5 = sangat baik/sangat tinggi /selalu

ISBN : 978.602.361.002.0

Banyak kuisioner ada 9, sehingga total banyak unit data ada 306 unit, yaitu berupa Kompetensi Pedagogik (PDG) ada 81unit data, Kompetensi Profesional (PRF) ada 72 unit data, Kompetensi Kepribadian (PBD) ada 54 unit data, Kompetensi Sosial (SOS) ada 45 unit data, Kompetensi Al-Islam dan Kemuhammadiyahan (MUH) ada 54 unit data. Data yang berjumlah 306 unit ini dinamakan data sebenarnya.

Sebagai simulasi, pada data sebenarnya dilakukan penghapusan data sebesar 15 unit atau 4,9%. Penghapusan ini dilakukan secara acak sehingga diperoleh data dengan *missing* data 4,9 %, dalam hal ini diberi nama data *missing value*. Sehingga pada Kompetensi Pedagogik (PDG) ada 8 unit *missing* data, Kompetensi Profesional (PRF) ada 3 unit *missing* data, Kompetensi Kepribadian (PBD) ada 1 unit *missing* data, Kompetensi Sosial (SOS) ada 2 unit *missing* data, Kompetensi Al-Islam dan Kemuhammadiyahan (MUH) ada 1 unit *missing* data.

Pada data *missing value* dilakukan metode imputasi. Pada penelitian ini dilakukan dua macam metode imputasi yaitu pertama dengan menginput *missing* data dengan suatu nilai konstan, yang kedua dengan metode Hot Deck. Metode imputasi yang pertama dengan menginput *missing* data dengan suatu nilai konstan, dalam hal ini diinput dengan nilai modus data dari tiap aspek kompetensi. Setelah selesai dilakukan imputasi, pada data yang sudah lengkap kemudian dihitung nilai rata-rata kuisioner untuk tiap aspek kompetensi. Hasil perhitungan nilai rata-rata kuisioner pada tiap aspek kompetensi disajikan pada tabel 1. Persentase kesalahan relatif dari Data *missing value*, Imputasi dengan nilai modus, Imputasi Hot Deck disajikan pada tabel 2.

Tabel 1. Nilai Rata-Rata Kuisioner pada Tiap Aspek Kompetensi

	PDG	PRF	PBD	SOS	MUH
Data sebenarnya	88.6	75.8	88.5	82.2	85.2
Data *missing value*	89.04	74.49	88.68	82.79	84.91
Imputasi dengan nilai modus	90.12	75.28	88.15	82.67	85.19
Imputasi Hot Deck	88.89	75.56	88.52	82.22	85.19

Dari tabel 1, nilai rata-rata kuisioner pada tiap aspek kompetensi akan dikonversi dalam nilai mutu, aturan rentang nilainya adalah:
A = 80.00-100
B = 68.00-79.9
C = 56.00- 67.9
D = 45.00- 59.9
Sehingga bila dikonversi dalam nilai mutu Kompetensi Pedagogik (PDG) nilainya A, Kompetensi Profesional (PRF) nilainya B, Kompetensi Kepribadian (PBD) nilainya A, Kompetensi Sosial (SOS) nilainya A, Kompetensi Al-Islam dan Kemuhammadiyahan (MUH) nilainya A. Hal ini berlaku untuk semua data yang digunakan, baik data sebenarnya, data *missing value*, data imputasi dengan nilai modus, maupun data imputasi Hot Deck.

Tabel 2. Persentase Kesalahan relatif

	PDG	PRF	PBD	SOS	MUH

ISBN : 978.602.361.002.0

Data *missing value*	0.496	1.728	0.203	0.718	0.340
Imputasi dengan nilai modus	1.716	0.686	0.395	0.572	0.012
Imputasi Hot Deck	0.327	0.317	0.023	0.024	0.012

Dari tabel 2 dapat dilihat bahwa metode imputasi Hot Deck memberikan hasil yang lebih baik bila dibandingkan dengan data *missing value* maupun bila dibandingkan imputasi dengan nilai modus. Bila diperhatikan pada tiap kompetensi, besar nilai persentase kesalahan relatif berkaitan erat dengan banyaknya item yang diinput. Dapat dilihat bahwa semakin banyak item yang diinput maka semakin besar nilai persentase kesalahan relatif. Hal ini sesuai dengan penelitian yang dilakukan oleh Little & Rubin [4].

4. SIMPULAN

Berdasarkan contoh kasus yang dibahas, menunjukkan bahwa metode imputasi Hot Deck memberikan hasil yang lebih baik bila dibandingkan dengan data *missing value,* juga bila dibandingkan dengan imputasi dengan nilai modus. Banyaknya item yang diimputasi memberikan pengaruh terhadap besar nilai persentase kesalahan relatif. Semakin banyak item yang diinput maka semakin besar nilai persentase kesalahan relatif.

5. DAFTAR PUSTAKA

[1] Dempster, A.P., Laird, N.M., & Rubin, D.B.1977. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, 39, 1-38.

[2] Ford BL. 1983. *An overview of hot-deck procedures*. In:Madow WG, OIkin I, Rubin DB (eds) Incomplete data in sample surveys, vol II: theory and bibliographies. Academic Press, New York, pp 85–207

[3] Honaker, J., King, G., and Blackwell, M. (2006), Amelia Software Web Site[accessed December 15, 2006]. Available online at http:// gking.harvard.edu/ amelia,

[4] Little, R.J.A. and Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York.

[5] T. Longford, Nicholas. 2005. *Missing Data and Small-Area*. New York: Springer.

[6] Rubin DB. 1976. *Inference and missing data*. Biometrika 63(3):581–592