JOURNAL OF FUNDAMENTAL MATHEMATICS
AND APPLICATIONS (JFMA) VOL. 2 NO. 2 (NOV 2019)
Available online at www.jfma.math.fsm.undip.ac.id

METODE *MULTIPLE IMPUTATION* UNTUK MENGATASI KOVARIAT TAK LENGKAP PADA DATA KEJADIAN BERULANG

Rianti Siswi Utami[1], Danardono[2]

[1,2]*Departemen Matematika, Universitas Gadjah Mada, Yogyakarta*
Email : [1]riantisiswi.u@ugm.ac.id, [2]danardono@ugm.ac.id

Abstract. Multiple imputation is one of estimation method used to impute missing observations. This method imputes missing observation several times then it is more possible to get the right estimate than just one time imputation. In this research, the method will be applied to estimate missing observations in covariates of recurrent event data. Some multiple imputation methods will be considered including combination of the event indicator, the event times, the logarithm of event times, and the cumulative baseline hazard. To compare these methods, Monte Carlo simulation will be used based on relative bias and Mean Squared Error (MSE). The recurrent events will be modelled using Cox proportional hazard model. Furthermore, real data application will be presented.
Keywords: multiple imputation, recurrent events, Cox regression, Monte Carlo Simulation

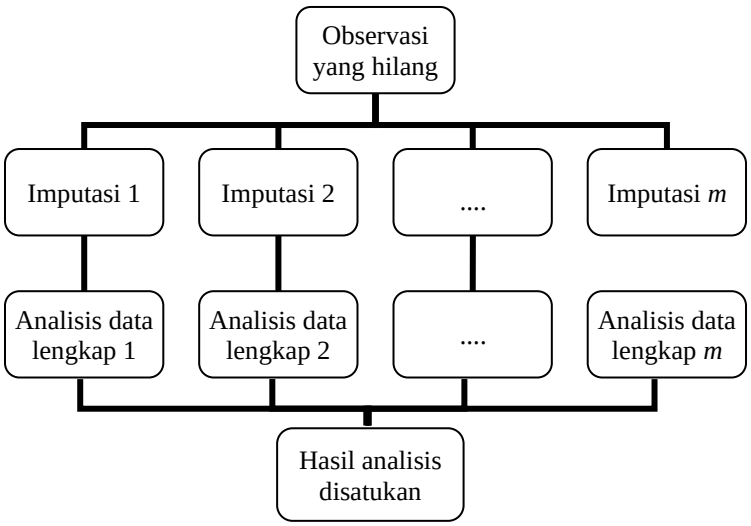
Abstrak. Metode *multiple imputation* merupakan salah satu metode untuk mengestimasi observasi yang hilang. Metode ini melakukan imputasi beberapa kali pada observasi yang hilang sehingga kemungkinan untuk memperoleh estimasi yang tepat lebih besar dari pada hanya melakukan imputasi satu kali. Pada penelitian ini metode tersebut akan diaplikasikan pada data kejadian berulang untuk mengestimasi observasi yang hilang pada kovariat. Akan dicoba beberapa metode *multiple imputation* yang melibatkan pengamatan terhadap kejadian, waktu kejadian, logaritma dari waktu kejadian, dan *baseline* hazard kumulatif. Kejadian berulang akan dimodelkan dengan model hazard proporsional Cox. Simulasi Monte Carlo akan digunakan untuk membandingkan beberapa metode *multiple imputation* berdasarkan nilai bias relatif dan *Mean Square Error* (MSE). Lebih lanjut, aplikasi pada data real akan diberikan.
Keywords: *multiple imputation*, kejadian berulang, regresi Cox, simulasi Monte Carlo

I. PENDAHULUAN

Dalam analisis survival, observasi yang hilang sering ditemui pada kovariat (variabel independen) [1]. Salah satu analisis survival yang banyak digunakan pada bidang kesehatan adalah analisis kejadian berulang (*recurrent events*). Kejadian berulang merupakan kejadian sama yang berulang pada suatu individu seperti serangan jantung, stroke, epilepsi, menstruasi pada wanita, diare, dan lain sebagainya. Analisis pada kejadian berulang sering melibatkan kovariat, tetapi pada beberapa kasus ada observasi yang hilang pada kovariat tersebut.

Salah satu metode untuk mengatasi observasi yang hilang adalah dengan menggunakan *multiple imputation*. Metode ini mengestimasi observasi yang hilang dengan dua atau lebih nilai berdasarkan suatu distribusi probabilitas [2]. Masing-masing nilai estimasi digunakan untuk melengkapi data kemudian data dianalisis menggunakan analisis yang telah ditentukan. Hasil analisis berdasarkan beberapa nilai estimasi tersebut kemudian dikombinasikan [3].

Dalam [2] diusulkan metode *multiple imputation* untuk inferensi statistika, dan sekarang metode tersebut banyak digunakan untuk menangani observasi yang hilang. Prinsip dari *multiple imputation* adalah mengestimasi observasi yang hilang dengan beberapa nilai berdasarkan distribusi data yang teramati. Nilai-nilai tersebut digunakan untuk melengkapi data sehingga diperoleh beberapa data dengan observasi yang lengkap. Masing-masing data dianalisis menggunakan metode yang telah ditentukan, kemudian hasil analisis dari beberapa data tersebut disatukan. Ilustrasi dari metode *multiple imputation* diberikan pada Gambar 1.



Gambar 1. Metode *multiple imputation*

Dalam [4] ditunjukkan bahwa dalam kasus observasi yang hilang pada kovariat, lebih baik menyertakan variabel respon (*outcome*) dalam prosedur imputasinya. Dalam analisis survival yang menyertakan kovariat, prosedur imputasi melibatkan *outcome* yang terdiri dari pengamatan kejadian, waktu kejadian, dan atau logaritma dari waktu kejadian. Perbandingan beberapa model imputasi dalam kasus satu kovariat tak lengkap pada data survival dilakukan oleh [5]. Diperoleh hasil bahwa imputasi berdasarkan estimator Nelson-Aalen lebih baik dari pada model-model yang lain.

Beberapa model kejadian berulang diusulkan antara lain oleh [6] dan [7]. Berikut model dari [7]

$$h_k\left(t|\mathbf{Z}_k\right)=h_{0k}\left(t\right)\exp\left(\boldsymbol{\beta}_k^T\mathbf{Z}_k\left(t\right)\right) \tag{1}$$

dengan $h\left(t|\mathbf{Z}\right)$ adalah hazard (resiko) mendapatkan kejadian pada saat t bersyarat pada kovariat $\mathbf{Z}$, $h_0\left(t\right)$ adalah *baseline* hazard pada saat t, $\boldsymbol{\beta}^T=\left(\beta_1,\beta_2,\ldots,\beta_p\right)$ adalah vektor parameter, dan $\mathbf{Z}$ adalah n x p kovariat dengan $\mathbf{Z}_i=\left(Z_{i1},Z_{i2},\ldots,Z_{ip}\right)^T$ adalah p vektor kovariat k = 1, 2, ..., K menunjukkan ulangan kejadian ke-k, dan $\mathbf{Z}_k\left(t\right)$ menunjukkan kovariat untuk kejadian ke-k dan dapat berupa kovariat bergantung waktu. Dalam [8]

dijelaskan bahwa terdapat kovariat yang tidak berubah-ubah dari kejadian satu ke yang lain misalnya jenis kelamin, sehingga modelnya menjadi

$$h_k\left(t|\mathbf{Z}\right) = h_{0k}\left(t\right)\exp\left(\boldsymbol{\beta}^T\mathbf{Z}\left(t\right)\right). \quad (2)$$

Pada penelitian ini akan dibahas metode *multiple imputation* untuk menangani kovariat tak lengkap pada data kejadian berulang. Dengan simulasi Monte Carlo akan dibandingkan beberapa metode *multiple imputation* yang merupakan kombinasi antara pengamatan kejadian dengan waktu kejadian, logaritma waktu kejadian, dan *baseline* hazard kumulatif. Perbandingan dilakukan pada nilai bias relatif dan *Mean Square Error* (MSE) dari hasil estimasi parameter model kejadian berulang. Lebih lanjut, metode ini akan diaplikasikan pada data real.

II. METODE *MULTIPLE IMPUTATION*

Observasi yang hilang, atau dikenal dengan *missing data*, banyak ditemui di berbagai bidang penelitian, terutama penelitian yang melibatkan data berskala besar. Berdasarkan [9], data hilang dapat diklasifikasikan menjadi tiga macam sebagai berikut:

a. *Missing Completely at Random* (MCAR) yaitu apabila probabilitas data hilang tidak bergantung pada data yang teramati maupun data yang tidak teramati;
b. *Missing at Random* (MAR) yaitu apabila probabilitas data hilang tidak bergantung pada data yang tidak teramati bersyarat pada data yang teramati;
c. *Missing not at Random* (MNAR) yaitu apabila probabilitas data hilang bergantung pada data yang tidak teramati bersyarat pada data yang teramati.

Berbagai metode telah dikembangkan untuk mengatasi data hilang. Metode yang paling sederhana adalah menghapus observasi yang memuat data hilang, atau dikenal dengan *Complete Case* (CC) *analysis*. Metode ini sesuai untuk tipe MCAR tetapi apabila data yang hilang cukup banyak maka data yang dianalisis akan berkurang cukup besar juga. Selain dengan menghapus data hilang, terdapat beberapa metode lain sebagai berikut:

a. *mean imputation* yaitu mengisi data hilang dengan rata-rata observasi teramati pada variabel yang memuat data hilang;
b. *last observation carried forward* yaitu mengisi data hilang dengan observasi sebelumnya;
c. *random hot deck imputation* yaitu mengisi data hilang dengan observasi yang memiliki karakteristik mirip dengan data hilang tersebut.

Ketiga metode di atas mengestimasi data hilang satu kali dengan satu nilai saja, sehingga cukup besar kemungkinan bahwa estimasi yang diperoleh kurang tepat. Metode lain yang cukup banyak digunakan adalah *multiple imputation*. Metode ini sangat fleksibel untuk berbagai teknik analisis data dan sesuai untuk tipe MAR. Inti dari *multiple imputation* adalah menggunakan distribusi dari observasi yang teramati untuk mengestimasi himpunan nilai yang mungkin untuk data yang hilang. Lebih lanjut, komponen random juga dipertimbangkan dalam proses estimasi data hilang untuk mewakili ketidakpastian nilai yang disetimasi.

Secara umum terdapat tiga tahapan dalam *multiple imputation*.

a. Tahap 1: membangkitkan *m* data hasil imputasi
   Data hilang diisi *m* kali dengan observasi yang diambil dari distribusi prediktif posterior data hilang bersyarat pada data yang teramati. Sebagai contoh, dalam suatu data terdapat variabel *Z* yang memuat data hilang dan terdapat variabel *X* yang lengkap (tidak memuat data hilang). Dibentuk model regresi antara variabel *Z* dan *X* untuk individu yang

teramati pada variabel Z. Misalkan $\hat{\boldsymbol{\beta}}$ dan $\mathbf{V}$ berturut-turut adalah estimasi parameter regresi dan matriks kovariannya. Selanjutnya kedua langkah berikut diulangi *m* kali. Diambil $\boldsymbol{\beta}^*$ secara random dari distribusi prediktif posterior yang umumnya didekati dengan $\boldsymbol{\beta}^* \sim \mathbf{MVN}\left(\hat{\boldsymbol{\beta}}, \mathbf{V}\right)$. Imputasi untuk *Z* diambil dari distribusi prediktif posterior dari *Z* menggunakan $\boldsymbol{\beta}^*$ dan distribusi probabilitas yang sesuai.

b. Tahap 2: menganalisis data bangkitan hasil imputasi
Setelah diperoleh *m* data lengkap hasil imputasi, masing-masing data dianalisis secara terpisah menggunakan teknik analisis yang telah ditentukan sehingga diperoleh *m* hasil analisis yang umumnya berupa estimasi parameter dan matriks kovariansinya.

c. Tahap 3: menggabungkan estimasi hasil analisis data bangkitan
Estimasi parameter dari *m* hasil analisis digabungkan dengan metode [2] yaitu berdasarkan teori asimtotik pada kerangka Bayesian. Misalnya $\hat{\boldsymbol{\theta}}_j$ adalah estimasi parameter pada analisis data hasil imputasi ke *j* dan $\mathbf{W}_j$ adalah estimasi variansi dari $\hat{\boldsymbol{\theta}}_j$ maka estimasi gabungan adalah rata-rata dari estimasi masing-masing data hasil imputasi sebagai berikut

$$\hat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j \,. \tag{3}$$

Variansi gabungan dari $\hat{\boldsymbol{\theta}}$ disusun berdasarkan variansi masing-masing imputasi

$$\mathbf{W} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{W}_j \tag{4}$$

dan variansi antar imputasi

$$\mathbf{B} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}\right)^2 \tag{5}$$

menjadi sebagai berikut

$$\text{var}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{W} + \left(1 + \frac{1}{m}\right)\mathbf{B} \,. \tag{6}$$

III. MENGATASI KOVARIAT TAK LENGKAP PADA DATA KEJADIAN BERULANG

3.1 Notasi Data Kejadian Berulang

Kejadian berulang atau dikenal dengan *recurrent event* banyak ditemui pada studi longitudinal biomedis seperti kekambuhan tumor, serangan epilepsi, asma, migrain, dan kunjungan ke rumah sakit. Seorang individu dapat mengalami kejadian yang sama pada beberapa waktu yang berbeda, sehingga kejadian yang sama terulang beberapa kali.

Didefinisikan terdapat *n* individu dalam penelitian, masing-masing individu dapat mengalami *K* kejadian. Waktu saat individu ke *i* memperoleh kejaian ke *k* dinotasikan dengan T_{ik}. Waktu kejadian tersebut diasumsikan dapat tersensor kanan pada saat C_{ik}. Didefinisikan notasi X_{ik} yang merupakan min(T_{ik}, C_{ik}), sehingga X_{ik} akan bernilai T_{ik} apabila kejadian ke *k* pada individu *i* terobservasi, dan akan bernilai C_{ik} apabila tersensor. Didefinisikan fungsi indikator sebagai berikut

$$\delta_{ik} = I\left(T_{ik} \le C_{ik}\right) = \begin{cases} 1 & T_{ik} \le C_{ik} \\ 0 & \text{lainnya} \end{cases}.$$

Kejadian pada suatu individu dapat dipengaruhi oleh satu atau lebih kovariat yang dinotasikan dengan $\mathbf{Z}_{ik}^T = \left[Z_{i1k}, Z_{i2k}, \ldots, Z_{ipk} \right]$. Apabila pengaruh kovariat diasumsikan sama antara kejadian yang satu dengan yang lain maka $\mathbf{Z}_i^T = \left[Z_{i1}, Z_{i2}, \ldots, Z_{ip} \right]$. Fungsi hazard untuk individu ke *i* pada kejadian ke *k* dinotasikan dengan $h_{ik}\left(t\right)$.

3.2 Model dan Estimasi Parameter pada Data Kejadian Berulang

Dalam model regresi hazard untuk data kejadian berulang, akan digunakan *gap time*, yaitu selang waktu antara dua kejadian yang berurutan [6]. Asumsi untuk model ini sama dengan model hazard proporsional Cox, yaitu rasio hazard antara dua individu yang memiliki kovariat berbeda konstan sepanjang waktu [10], tetapi dalam model ini *baseline* hazard dapat berubah-ubah untuk masing-masing kejadian.

Didefinsikan model hazard untuk *gap time* kejadian berulang adalah sebagai berikut

$$h_{ik}\left(t \middle| \mathbf{Z}_{ik}\left(t\right)\right) = h_{0k}\left(t - t_{k-1}\right) \exp\left(\sum_{j=1}^{p} \beta_{jk} Z_{ijk}\left(t\right) \right) \tag{7}$$

dengan *t* menunjukkan waktu sejak individu masuk dalam pengamatan, t_{k-1} adalah waktu pengamatan untuk kejadian ke *k*-1, dan $h_{0k}\left(t\right), k = 1, 2, \ldots, K$ adalah *baseline* hazard untuk kejadian ke *k*. Apabila pengaruh kovariat diasumsikan sama untuk semua kejadian maka modelnya menjadi sebagai berikut

$$h_{ik}\left(t \middle| \mathbf{Z}_i\left(t\right)\right) = h_{0k}\left(t - t_{k-1}\right) \exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\left(t\right) \right). \tag{8}$$

Estimasi parameter untuk $\boldsymbol{\beta}$ dilakukan dengan metode *partial* likelihood. Berikut *partial* likelihood menurut Wei, dkk. (1989)

$$L\left(\boldsymbol{\beta}\right) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{\exp\left(\sum_{j=1}^{p} \beta_{jk} Z_{ijk}\left(X_{i,k-1} + G_{ik} \right) \right)}{\sum_{l=1}^{n} Y_{lk}\left(G_{ik}\right) \exp\left(\sum_{j=1}^{p} \beta_{jk} Z_{ljk}\left(X_{l,k-1} + G_{ik} \right) \right)} \right\}^{\delta_{ik}} \tag{9}$$

dengan $G_{ik} = X_{ik} - X_{i,k-1}$ adalah interval *gap time* dan $Y_{lk}\left(G_{ik}\right) = I\left(G_{lk} \ge G_{ik}\right)$ adalah indikator himpunan individu yang beresiko. Estimator maksimum likelihood dari $\boldsymbol{\beta}$ adalah solusi dari persamaan skor berikut

$$U\left(\boldsymbol{\beta}\right) = \frac{\partial \log L\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}}. \tag{10}$$

Apabila waktu kejadian saling independen maka matriks varian kovarian dari estimator $\hat{\boldsymbol{\beta}}$ dapat diperoleh dari invers matriks informasi, yaitu $I^{-1}\left(\hat{\boldsymbol{\beta}}\right)$ dengan

$$I\left(\hat{\boldsymbol{\beta}}\right)=\frac{\partial U\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}. \tag{11}$$

Matriks $I^{-1}\left(\hat{\boldsymbol{\beta}}\right)$ sering dinamakan matriks varian kovarian "*naive*". Akan tetapi apabila waktu kejadian tidak saling independen maka digunakan matriks varian kovarian "*sandwich*" atau "*robust*" $\hat{Q}\left(\hat{\boldsymbol{\beta}}\right)$. Penjelasan mengenai perhitungan $\hat{Q}\left(\hat{\boldsymbol{\beta}}\right)$ dapat dilihat di [11], serta [7].

3.3 Multiple Imputation pada Data Kejadian Berulang

Seringkali terdapat observasi yang hilang pada kovariat, misalnya salah satu kovariat pada beberapa individu tidak teramati. Untuk mengatasinya, dapat digunakan metode *multiple imputation* yang telah dijelaskan pada pembahasan sebelumnya. Secara lebih khusus, diperhatikan bahwa hubungan antara kovariat dengan waktu kejadian bersifat non linear sehingga metode PMM lebih sesuai digunakan. Pada data kejadian berulang, kovariat yang memuat data hilang diimputasi dengan memodelkan kovariat tersebut dengan kovariat yang lengkap, waktu kejadian X_{ik} (dapat berupa X_{ik}, $\log X_{ik}$, atau $\hat{H}\left(X_{ik}\right)$ yang merupakan estimator Nelson-Aalen untuk hazard kumulatif) dan δ_{ik} [1].

IV. SIMULASI

Tujuan dari dilakukannya simulasi data adalah untuk mengetahui cara terbaik dalam memodelkan kovariat dengan variabel respon dalam proses imputasi data.

4.1 Skema Simulasi

Skema simulasi yang digunakan mengikuti skema dari [1]:

1. data terdiri dari dua kejadian berulang dan dua kovariat;
2. pengaruh kovariat diasumsikan sama untuk setiap kejadian berulang sehingga variabel-variabel yang ada dalam data yaitu X_{i1}, δ_{i1}, X_{i2}, δ_{i2}, Z_{i1}, dan Z_{i2};
3. ukuran sampel yang digunakan adalah 2500;
4. waktu kejadian yang tersensor sebanyak 40%;
5. terdapat 20% pengamatan yang hilang pada Z_{i1} dengan mengikuti konsep MAR, sedangkan pengamatan pada Z_{i2} lengkap;
6. korelasi antara X_{i1} dan X_{i2} sebesar $\rho=0$ dan $\rho=0{,}5$;
7. korelasi antara Z_{i1} dan Z_{i2} sebesar $\rho_c=0$ dan $\rho_c=0{,}5$;
8. model imputasi yang akan digunakan adalah
   a. M1: regresi linear antara Z_{i1} dengan Z_{i2}, X_{i1}, dan δ_{i1};
   b. M2: regresi linear antara Z_{i1} dengan Z_{i2}, $\hat{H}\left(X_{i1}\right)$, dan δ_{i1};
   c. M3: regresi linear antara Z_{i1} dengan Z_{i2}, X_{i1}, δ_{i1}, X_{i2}, dan δ_{i2};
   d. M4: regresi linear antara Z_{i1} dengan Z_{i2}, $\hat{H}\left(X_{i1}\right)$, δ_{i1}, $\hat{H}\left(X_{i2}\right)$, dan δ_{i2};
9. karena keterbatasan komputer, maka imputasi hanya dilakukan 5 kali dan direplikasi 10 kali untuk masing-masing skema dan masing-masing model imputasi;
10. keempat model imputasi dibandingkan berdasarkan nilai bias relatif (RB) dan *Mean Square Error* (MSE) sebagai berikut

$$\mathrm{RB}\left(\hat{\beta}\right) = E\left(\frac{\hat{\beta}-\beta}{\beta}\right) \tag{12}$$

dan

$$\mathrm{MSE}\left(\hat{\beta}\right) = E\left[\left(\hat{\beta}-\beta\right)^2\right]; \tag{13}$$

11. imputasi dilakukan dengan fungsi "mice" dari *package* "mice" pada *software* R;
12. estimasi parameter pada data kejadian berulang menggunakan iterasi Newton Raphson dengan fungsi "nlm" dari *package* "stats" pada *software* R.

4.2 Algoritma Simulasi

Langkah-langkah dalam melakukan simulasi mengikuti algoritma dalam [1] sebagai berikut:

1. membangkitkan kovariat Z_{i1} dan Z_{i2} dari distribusi normal standar multivariat dengan korelasi sebesar ρ_c;
2. untuk suatu nilai ρ, dihitung ρ_0 dengan cara
   $$\rho_0 = \frac{-w+\sqrt{w^2+2\rho(1-w)}}{1-w}$$
   dengan $w \approx 0,941$, sehingga untuk $\rho = 0$ diperoleh $\rho_0 = 0$ dan untuk $\rho = 0,5$ diperoleh $\rho_0 = 0,522782$;
3. untuk individu *i* dibangkitkan y_{i1} dari distribusi $N(0,1)$, kemudian dibangkitkan y_{i2} dari distribusi $N\left(\rho_0 y_{i1}, 1-\rho_0^2\right)$. Selanjutnya y_{ik} ditransformasi menjadi variabel random uniform menggunakan fungsi distribusi kumulatif normal standar, yaitu $U_{ik} = \Phi\left(y_{ik}\right)$;
4. membangkitkan waktu kejadian dari distribusi Weibull dengan parameter $\lambda_T = 0,0001$ dan $\kappa = 1$ yang memiliki model regresi hazard
   $h_T(t) = \lambda_T \kappa t^{\kappa-1} \exp\left(\beta_1 Z_1 + \beta_2 Z_2\right)$.
   Nilai dari β_1 dan β_2 ditentukan sebesar $\beta_1 = \beta_2 = 0,5$. *Gap time* dihitung dari
   $$g_{ik} = \left(-\frac{\log\left(U_{ik}\right)}{\lambda_T \exp\left(\beta_1 Z_1 + \beta_2 Z_2\right)}\right)$$
   sehingga diperoleh waktu kejadian $T_{ik} = g_{i1} + \ldots + g_{ik}$;
5. waktu sensor C_i dibangkitkan dari distribusi Weibull dengan parameter $\lambda_C = 0,00002$ dan $\kappa = 1$ sehingga waktu kejadian tersensor sekitar 40%;
6. diperoleh waktu pengamatan $X_{ik} = \min\left(T_{ik}, C_i\right)$ dengan indikator δ_{ik}.

Pada skema simulasi yang telah dijelaskan sebelumnya, variabel Z_{i1} memuat pengamatan yang tidak lengkap sebesar 20% mengikuti konsep MAR. Misalnya I_1 adalah indikator untuk data lengkap, yaitu $I_1 = 1$ jika pada individu *i* variabel Z_{i1} teramati, dan

$I_1 = 0$ untuk Z_{i1} tidak teramati. Pada asumsi MAR, data yang hilang hanya bergantung pada data yang teramati

$$f\left(I_1 | Z_1, Z_2, \phi\right) = f\left(I_1 | Z_2, \phi\right)$$

dengan ϕ adalah parameter yang tidak diketahui. Dengan menggunakan fungsi logit sebagai fungsi penghubung, diperoleh invers dari fungsi di atas adalah

$$P\left(I_1 | Z_2\right) = \frac{\exp\left(f\left(Z_2\right)\right)}{1+\exp\left(f\left(Z_2\right)\right)}.$$

Dengan mengambil $f\left(Z_2\right) = 0,85 + Z_2$ maka data yang hilang pada Z_1 akan mendekati 20%.

4.3 Hasil Simulasi

Berdasarkan skema dan algoritma simulasi yang telah dijelaskan, diperoleh hasil pada Tabel 1 dan 2.

Tabel 1. Nilai RB dan MSE untuk estimasi parameter β_1

ρ	ρ_c	M1	M2	M3	M4
0	0	-0,06695676 0,0009545472	-0,03998479 0,0003642703	0,02981098 0,0002668108	**0,02002715** **0,0001775215**
	0,5	-0,012722211 0,0002438346	**-0,010796747** **0,00009883719**	0,01858696 0,0001068952	0,02610620 0,0001970214
0,5	0	-0,06280729 0,0008267094	-0,07502112 0,0010603295	**0,02471283** **0,0002005020**	0,03195774 0,0004278032
	0,5	-0,03939937 0,0003771950	-0,026296810 0,0002153968	**0,01978984** 0,0002143411	0,02149180 **0,0001382935**

Tabel 2. Nilai RB dan MSE untuk estimasi parameter β_2

ρ	ρ_c	M1	M2	M3	M4
0	0	-0,03049896 0,0001942493	**-0,01654418** **0,0001500337**	-0,04969301 0,0004933905	-0,04678363 0,0005035408
	0,5	-0,004800234 0,00008440405	**0,003753121** **0,00008000834**	-0,02501397 0,0001536567	-0,03614256 0,0002848717
0,5	0	**-0,01215959** 0,0001536168	-0,01673472 **0,0001009035**	-0,03737973 0,0002966461	-0,03369040 0,0003300400
	0,5	0,01587085 0,0001090048	**0,008175508** **0,0001059670**	-0,04045980 0,0004254317	-0,02953247 0,0002623304

Pada Tabel 1 dan 2, nilai RB dan MSE terkecil mendekati nol di setiap baris dicetak tebal. Dari Tabel 1 terlihat bahwa M2, M3, dan M4 sama-sama memiliki nilai RB dan MSE terkecil, tetapi lebih banyak di M3 dan M4. Hal ini berarti memasukkan X_{i2} atau $\hat{H}\left(X_{i2}\right)$ ke dalam model imputasi dapat memperkecil bias pada estimasi kovariat yang mengandung data hilang. Hasil yang berbeda diperoleh untuk kovariat lengkap, pada Tabel 2 terlihat bahwa hampir di setiap baris nilai RB dan MSE terkecil terletak di kolom M2, berarti hasil imputasi dengan memasukkan $\hat{H}\left(X_{i1}\right)$ sudah cukup baik untuk kovariat lengkap.

V. STUDI KASUS

Akan diaplikasikan metode *multiple imputation* pada data kejadian berulang yaitu data haid. Haid pada wanita terjadi hampir di setiap bulan, sehingga kejadian ini dapat dipandang sebagai kejadian berulang. Siklus haid adalah lamanya atau jarak waktu mulai haid sampai mulai haid berikutnya. Siklus haid normalnya antara 21-35 hari, rata-rata 28 hari dan jika siklus haid kurang dari 21 hari atau lebih dari 35 hari kemungkinan bukan darah haid [12]. Faktor-faktor yang dapat mempengaruhi siklus haid antara lain faktor genetik, status gizi, psikis dan fisik, hormon, dan sosial-ekonomi [13].

Dilakukan survei pada 200 wanita mengenai mengenai siklus haidnya di tiga bulan terakhir. Selain itu ditanyakan pula beberapa variabel yang diperkirakan mempengaruhi siklus haid antara lain usia, status pekerjaan (mahasiswi, bekerja, ibu rumah tangga), status (menikah, belum menikan), dan lama tidur dalam sehari. Variabel-variabel tersebut diambil karena diperkirakan berkaitan dengan faktor psikis, fisik, hormon, dan sosial-ekonomi. Berikut ringkasan dari data.

Tabel 3. Statistika deskriptif dari data

Variabel kontinu				
	Usia (tahun)	Tidur (jam)	Siklus haid pertama (hari)	Siklus haid ke dua (hari)
Rata-rata	25,12	7,00	31,84	31,10
Sd	4,65	1,197	8,85	6,88
Min	19	4	8	15
Max	35	12	77	61

Variabel kategorik			
	Status (belum menikah)	Bekerja (ya)	Ibu rumah tangga (ya)
Frekuensi	120	98	41

Data yang diperoleh adalah data lengkap, tidak ada variabel yang memuat data hilang. Untuk keperluan studi kasus, sebanyak 16,5% pengamatan pada variabel tidur dihilangkan dengan konsep MAR, yaitu hilangnya data pada variabel tidur tidak bergantung pada variabel tersebut bersyarat pada variabel yang lain. Pada kasus ini dipilih variabel usia sebagai variabel bersyaratnya karena memiliki korelasi paling besar dengan variabel tidur dibandingkan variabel-variabel yang lain.

Model imputasi yang digunakan sama seperti pada simulasi data dengan tambahan variabel independen. Misalnya nama-nama variabel disingkat sebagai berikut:

- usia = Usia
- tidur = Tidur
- status = Status (1 = belum menikan, 0 = menikah)
- bekerja = Bekerja (1 = ya, 0 = lainnya)
- irt = Ibu rumah tangga (1 = ya, 0 = lainnya)
- t1 = Siklus haid pertama
- gap = Siklus haid ke dua
- t2 = t1 + gap
- d1 = status pengamatan pada t1 (1 = teramati, 0 tersensor)
- d2 = status pengamatan pada gap (1 = teramati, 0 tersensor).

Model imputasi yang akan digunakan adalah:

1. M1: regresi linear antara tidur dengan usia, status, bekerja, irt, t1, d1;
2. M2: regresi linear antara tidur dengan usia, status, bekerja, irt, H(t1), d1;
3. M3: regresi linear antara tidur dengan usia, status, bekerja, irt, t1, d1, t2, d2;
4. M4: regresi linear antara tidur dengan usia, status, bekerja, irt, H(t1), d1, H(t2), d2.

Karena terdapat lima variabel independen dalam data, maka model hazard untuk kejadian berulang untuk orang ke *i* adalah:

$$h_{ik}\left(t\middle|\mathbf{Z}_i\right)=h_{0k}\left(t-t_{k-1}\right)\exp\left(\beta_1\text{usia}_i+\beta_2\text{status}_i+\beta_3\text{bekerja}_i+\beta_4\text{irt}_i+\beta_5\text{tidur}_i\right)$$

Berikut estimasi parameter untuk data lengkap dan data hasil imputasi beserta selisih absolutnya dengan estimasi parameter data lengkap.

Tabel 4. Estimasi parameter dan selisih absolut (dalam kurung)

	Data lengkap	M1	M2	M3	M4
$\hat{\beta}_1$	0,01764524	0,01762787 (**0,00001737**)	0,01724719 (0,00039805)	0,01754032 (0,00010492)	0,0173681 (0,00027714)
$\hat{\beta}_2$	0,22783158	0,22924078 (0,0014092)	0,22653232 (0,00129926)	0,2278092 (**0,00002238**)	0,2275762 (0,00025538)
$\hat{\beta}_3$	0,15153362	0,15171067 (0,00017705)	0,1404767 (0,01105692)	0,15143448 (**0,00009914**)	0,15180976 (0,00027614)
$\hat{\beta}_4$	0,32493494	0,33166634 (0,0067314)	0,31879814 (0,0061368)	0,3258225 (0,00088756)	0,32521802 (**0,00028308**)
$\hat{\beta}_5$	-0,0186241	-0,0307041 (0,01207999)	-0,0263177 (0,00769352)	-0,0202639 (0,00163977)	-0,0181865 (**0,00043765**)

Pada Tabel 4 selisih absolut terkecil pada setiap baris dicetak tebal. Hasil yang diperoleh pada studi kasus hampir sama dengan hasil pada simulasi data, yaitu model imputasi yang menyertakan kedua waktu kejadian (model M3) atau estimasi hazard kumulatif dari kedua waktu kejadian (model M4) memberikan estimasi parameter yang lebih dekat dengan data lengkap dibandingkan model imputasi yang lain.

VI. KESIMPULAN

Berdasarkan uraian pada pembahasan sebelumnya dapat ditarik beberapa kesimpulan sebagai berikut:

1. pada data kejadian berulang dengan dua waktu pengamatan, memasukkan kedua waktu kejadian atau estimasi hazard kumulatif dari kedua waktu kejadian pada model imputasi dapat memperkecil bias dalam estimasi parameter;
2. pada simulasi data, diperoleh hasil yang berbeda antara model imputasi terbaik untuk kovariat tidak lengkap dan kovariat lengkap. Perbedaan hasil ini dapat disebabkan oleh kurangnya replikasi data.kesimpulan disini.

REFERENSI

[1] Z. Huo, A Comparison of Multiple Imputation Methods for Missing Covariate Values in Recurrent Event Data, *Tesis*, Uppsala University, 2015.
[2] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc, New York, 1987.
[3] Y.C. Yuan, Multiple Imputation for Missing Data: Concepts and New Development, *Artikel*, SAS Institute Inc., Rocville, MD, 2005.
[4] K.G. Moons, R.A. Donders, T. Stijnen, dan F.E. Harrel, Using the Outcome for Imputation of Missing Predictor Values was Preferred, *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1092-1101, 2006.
[5] I.R. White, dan P. Royston, Imputing Missing CovariateValues for The Cox Model, *Statistics in Medicine*, vol. 28, no. 15, pp. 1982-1998, 2009.
[6] R.L. Prentice, B.J. Williams, dan A.V. Peterson, On the Regression Analysis of Multivariate Failure Time Data, Biometrika, vol. 68, pp. 373—379, 1981.
[7] L.J. Wei, D.Y. Lin, dan L. Weissfeld, Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribution, *Journal of the American Statistical Association*, vol. 84, pp. 1065—1073, 1989.
[8] X. Liu, *Survival Analysis*: *Models and Applications*, John Wiley & Sons, Inc, New York, 2012.
[9] R.J.A. Little dan D.B. Rubin, *Statistical Analysis with Missing Data*, edisi 2, Wiley: Hoboken, N.J., 2002.
[10] J.P. Klein dan M.L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data*, edisi 2, Springer-Verlag, New York, 2003.
[11] H.J. Lim dan X. Zhang, Additive and Multiplicative Hazards Modeling for Recurrent Event Data Analysis, *BMC Medical Research Methodology*, vol. 11, pp. 101 - 121, 2011.
[12] Z.A. Baso dan J. Raharjo, *Kesehatan Reproduksi, Panduan Bagi Perempuan*, Pustaka Pelajar, Yogyakarta, 1999.
[13] T. Suwarni, Faktor Determinan yang Mempengaruhi Siklus Menstruasi, *Indonesian Journal On Medical Science*, vol. 2, no. 1, pp. 33 - 38, 2015.