Estimasi Parameter Data Tersensor Tipe I Berdistribusi Log-logistik Menggunakan *Maximum Likelihood Estimate* dan Iterasi Newton-Rhapson

ALFENSI FARUK
Fakultas MIPA, Universitas Sriwijaya; email: alfensifaruk@unsri.ac.id

Abstract: Survival analysis is one of the topics in the field of mathematics which deals with statistical analysis of the time until the occurrence of one or more a particular event. The purposes of this study are to obtain the survival model and to estimate the parameters of the type I censored data which log-logistic distributed. Maximum likelihood estimate was used to estimate the unknown parameters. Based on the results and discussion, if the exact values of the parameters γ and β are analytically difficult to obtain, then the numerical approach by the Newton-Rhapson iteration can be used to approach the values of both parameters.

Keywords: censored data, log-logistic distribution, maximum likelihood estimate, Newton-Rhapson iteration

1 PENDAHULUAN

Analisis *survival* adalah analisis mengenai lamanya waktu hidup suatu subjek pada suatu keadaan tertentu. Data *survival* dapat dikatakan sebagai data yang berupa waktu hingga terjadinya suatu kejadian. Data tersensor (*censored data*) merupakan data yang sebagian informasi dari data tersebut tidak lengkap, yang diakibatkan oleh berbagai alasan seperti subjek pengamatan memilih keluar atau kejadian yang diamati tidak terjadi selama waktu penelitian (Lee, 2003). Salah satu jenis data tersensor adalah data tersensor kanan (*right censored*), yang terbagi menjadi data tersensor kanan tipe I, II, dan III. Berbagai mekanisme penyensoran untuk data tersensor kanan acak dan model regresi untuk data tersensor serta estimasi model *proportional hazards* dibahas oleh Gijbels (2010).

Penyensoran tipe I dapat dilakukan apabila subjek-subjek penelitian diamati pada suatu jangka waktu yang tetap sejak awal waktu pengamatan. Contoh penggunaan sensor kanan tipe I pada data *survival* adalah ketika seorang peneliti ingin mengetahui bagaimana pengaruh suatu terapi terhadap proses kesembuhan pasien penyakit kanker. Pemberian terapi dilakukan kepada semua subjek penelitian di awal pengamatan, sehingga jika dalam jangka waktu pengamatan ada yang sembuh akibat terapi maka diperoleh data survival dari subjek yang bersangkutan, sedangkan jika dalam jangka waktu pengamatan tersebut terdapat subjek yang sembuh bukan karena terapi yang diberikan atau terdapat subjek yang masih belum sembuh maka subjek-subjek ini dikategorikan sebagai data tersensor kanan tipe I.

Metode parametrik digunakan jika suatu distribusi yang sesuai dicocokkan dengan data atau ketika suatu distribusi dapat diasumsikan kepada populasi dari sampel yang diambil. Beberapa distribusi statistik yang sering digunakan dalam analisis *survival* antara lain distribusi weibul, eksponensial, dan log-normal. Penggunaan distribusi-distribusi parametrik pada data *survival* dilakukan oleh Zhao (2008), yang dalam penelitian tersebut dilakukan simulasi terhadap berbagai bentuk data *survival* dan dibandingkan hasilnya sehingga terlihat metode mana yang paling cocok untuk mengestimasi fungsi *survival*.

Distribusi log-logistik merupakan salah satu distibusi statistik yang dapat digunakan pada data *survival*. Berdasarkan penelitian Bennet (1983), diketahui bahwa distribusi log-logistik memiliki fungsi *hazard* yang tidak monoton, sehingga distribusi ini cocok digunakan dalam pemodelan waktu *survival* penyakit. Estimasi parameter fungsi *survival* berdistribusi log-logistik dapat dilakukan dengan metode *Maximum Likelihood Estimate* (MLE). Sari (2011) telah memperlihatkan bagaimana mengestimasi fungsi *survival*, fungsi *hazard*, dan fungsi kepadatan peluang dengan MLE dari data tersensor

kanan tipe II berdistribusi log-logistik, namun belum dibahas untuk data tersensor kanan tipe I serta pendekatan secara numeriknya. Oleh karena itu, dalam penelitian ini dibahas bagaimana estimasi parameter dari data tersensor kanan tipe I menggunakan metode MLE, kemudian dilanjutkan dengan penyelesaian secara numerik menggunakan iterasi Newton-Rhapson.

2 METODE PENELITIAN

Secara garis besar, tahapan-tahapan yang dilakukan dalam penelitian ini adalah:

1. Menentukan fungsi kepadatan peluang dan fungsi *hazard* dari data *survival* tersensor tipe I yang berdistribusi log-logistik.
2. Menentukan $\hat{\gamma}$ dan $\hat{\beta}$ menggunakan metode MLE dan iterasi Newton-Rhapson.
3. Memberikan contoh penerapan penentuan estimator MLE $\hat{\gamma}$ dan $\hat{\beta}$ dari suatu data *survival* tersensor tipe I berdistribusi log-logistik menggunakan program berbasis Maple 18.

3 HASIL DAN PEMBAHASAN

Model Survival Berdistribusi Log-Logistik

Bentuk umum fungsi kepadatan peluang berdis-tribusi log-logistik adalah

$$f(t,\gamma,\beta) = \frac{\gamma\beta\ t^{\beta-1}}{\left[1+\gamma t^{\beta}\right]^2}, \tag{1}$$

dengan parameter $\gamma, \beta > 0$, dan variabel waktu $t \geq 0$.

Berdasarkan definisi fungsi distribusi kumulatif, pers.(1) diintegralkan dari 0 sampai t diperoleh

$$F(t,\gamma,\beta) = \int_0^t f((t,\gamma,\beta)dt = \int_0^t \frac{\gamma\ \beta\ t^{\beta-1}}{\left[1+\ \gamma t^{\beta}\right]^2}\ dt = \frac{\gamma\ t^{\beta}}{1+\gamma\ t^{\beta}}. \tag{2}$$

Pers.(2) di atas adalah fungsi distribusi kumulatif berdistribusi log-logistik, yang selanjutnya berdasarkan hubungan antara $F(x)$ dengan fungsi *survival* $S(x)$ maka dapat ditentukan bentuk umum dari fungsi *survival* berdistribusi log-logistik yaitu

$$S(t,\gamma,\beta) = 1 - F(t,\gamma,\beta) = 1 - \frac{\gamma\ t^{\beta}}{1+\gamma\ t^{\beta}} = \frac{1}{1+\gamma\ t^{\beta}} \tag{3}$$

Selain fungsi *survival* dan kepadatan peluang, terdapat fungsi penting lainnya dalam analisis *survival*, yaitu fungsi *hazard* $h(x)$. Fungsi *hazard* diinterpretasikan sebagai kecepatan terjadinya kejadian yang diamati dari subjek pengamatan pada jangka waktu pengamatan. Fungsi *hazard* dari data berdistribusi log-logistik dengan parameter γ dan β dapat diperoleh dengan membagi fungsi kepadatan peluang (1) dengan fungsi *survival* (3), sehingga didapat

$$h(t,\gamma,\beta) = \frac{f(t,\gamma,\beta)}{S(t,\gamma,\beta)} = \left(\frac{\gamma\beta\ t^{\beta-1}}{\left[1+\gamma t^{\beta}\right]^2}\right)\Big/\left(\frac{1}{1+\gamma\ t^{\beta}}\right) = \frac{\gamma\beta\ t^{\beta-1}}{1+\gamma\ t^{\beta}}. \tag{4}$$

Fungsi Likelihood Data Tersensor Tipe I Berdistribusi Log-Logistik

Data tersensor kanan (*right censored*) terjadi apabila semua data tersensor diperoleh setelah waktu pengamatan dimulai, dengan kata lain data tersensor terjadi ketika $t \geq 0$. Apabila pengamatan dari semua subjek penelitian dimulai pada waktu yang sama, misalkan pada saat $t_1 = 0$ dan berakhir pada waktu $t_2 = a$, maka pada kasus-kasus tertentu tidak semua waktu *survival* dari semua subjek penelitian dapat diperoleh. Hal ini dikarenakan pada selang waktu $[0, a]$ kejadian yang diinginkan tidak terjadi disebabkan berbagai alasan, yaitu subjek memilih keluar dari penelitian, penyebab terjadinya kejadian tidak sesuai dengan penelitian, atau kejadian tersebut terjadi pada waktu $t > a$. Data yang sebagian informasinya tidak lengkap seperti dalam kasus ini disebut sebagai data tersensor tipe I.

Apabila terdapat n buah subjek penelitian yang diamati pada selang waktu $[0, a]$, dan dari n data tersebut terdapat sebanyak r data tersensor tipe I, dengan $0 \leq r \leq n$ dan $n, r \in Z^+$, maka waktu *survival* dari n subjek tersebut dapat dinotasikan menjadi $t_1, t_2, \ldots, t_{n-r}, t^*_{n-r+1}, t^*_{n-r+2} \ldots t^*_n$, di mana

tanda "$*$" adalah simbol bagi data yang tersensor tipe I. Data-data tersensor yang kejadiannya melebihi waktu a, nilai waktu *survival*nya adalah $t = a$, sehingga nilai-nilai $t_{n-r+1}^*,\ t_{n-r+2}^*, \dots, t_n^*$ juga terletak dalam selang $[0, a]$. Diasumsikan bahwa setiap waktu *survival* adalah kejadian yang saling bebas, sehingga fungsi kepadatan peluang bersama berdistribusi log-logistik dari n buah data dengan $(n - r)$ data tersensor tipe I diperoleh dengan mengalikan peluang bersama dari data tak tersensor $\prod_{i=1}^{r} f(t_i, \gamma, \beta)$ dengan peluang bersama dari data tersensor $\prod_{i=r+1}^{n} S(t_i^*, \gamma, \beta)$. Jika dilihat dari sudut pandang bahwa γ dan β adalah variabel serta t_i adalah parameter, maka fungsi kepadatan peluang bersama ini disebut sebagai fungsi *likelihood*, yang dilambangkan dengan $L(\gamma, \beta)$, sehingga diperoleh

$$L(\gamma, \beta) = \prod_{i=1}^{r} f(t_i, \gamma, \beta) \prod_{i=r+1}^{n} S(t_i^*, \gamma, \beta) = \prod_{i=1}^{r} \frac{\gamma\beta\, t_i^{\beta-1}}{\left[1+\gamma\, t_i^{\beta}\right]^2} \cdot \prod_{i=r+1}^{n} \frac{1}{1+\gamma\, t_i^{*\beta}}. \quad (5)$$

Fungsi likelihood (5) dapat diinterpretasikan sebagai ukuran kemungkinan untuk memperoleh suatu himpunan spesifik dari waktu-waktu survival $t_1, t_2, \dots, t_r, t_{r+1}^*,\ t_{r+2}^*, \dots, t_n^*$, dengan diberikan parameter-parameter γ dan β.

Penentuan Estimator Parameter γ dan β

Metode dalam MLE adalah menemukan estimator dari parameter-parameter yang memaksimumkan fungsi *likelihood*nya, dengan kata lain menemukan parameter-parameter yang memiliki kemungkinan terbesar untuk mendapatkan waktu-waktu *survival* $t_1, t_2, \dots, t_r, t_{r+1}^*,\ t_{r+2}^*, \dots, t_n^*$.

Apabila diambil logaritma natural (ln) dari fungsi likelihood (5) dan dilambangkan dengan $L_L(\gamma, \beta)$, maka didapatkan

$$L_L(\gamma, \beta) = \ln L(\gamma, \beta) = \prod_{i=1}^{r} \ln f(t_i, \gamma, \beta) + \prod_{i=r+1}^{n} \ln S(t_i^*, \gamma, \beta) = \prod_{i=1}^{r} \ln \frac{\gamma\beta\, t_i^{\beta-1}}{\left[1+\gamma\, t_i^{\beta}\right]^2} + \prod_{i=r+1}^{n} \ln \frac{1}{1+\gamma\, t_i^{*\beta}}$$

$$= \sum_{i=1}^{r} \ln \gamma + \sum_{i=1}^{r} \ln \beta + \sum_{i=1}^{r} \ln(t_i)^{\beta-1} - 2\sum_{i=1}^{r} \ln\left(1+\gamma\, t_i^{\beta}\right) + \sum_{i=r+1}^{n} \ln 1 - \sum_{i=r+1}^{n}\left(1+\gamma\, t_i^{*\beta}\right)$$

$$= r\,(\ln \gamma + \ln \beta) + (\beta - 1)\sum_{i=1}^{r} \ln(t_i) - 2\sum_{i=1}^{r} \ln\left(1+\gamma\, t_i^{\beta}\right) - \sum_{i=r+1}^{n} \ln\left(1+\gamma\, t_i^{*\beta}\right). \quad (6)$$

Jika nilai estimator dari parameter memaksimalkan fungsi log-likelihood (6), maka estimator tersebut juga memaksimalkan fungsi likelihood (5). Langkah pertama mendapatkan nilai estimator maksimum dari persamaan log-likelihood (6) adalah menurunkan pers.(6) terhadap γ dan β, kemudian hasilnya disamakan dengan nol sehingga didapatkan

$$\frac{\partial L_L(\gamma,\beta)}{\partial \gamma} = \frac{r}{\gamma} - 2\sum_{i=1}^{r} \frac{t_i^{\beta}}{1+\gamma\, t_i^{\beta}} - \sum_{i=r+1}^{n} \frac{t_i^{*\beta}}{1+\gamma\, t_i^{*\beta}} = 0, \quad (7)$$

dan

$$\frac{\partial L_L(\gamma,\beta)}{\partial \beta} = \frac{r}{\beta} + \sum_{i=1}^{r} \ln(t_i) - 2\gamma\sum_{i=1}^{r} \frac{t_i^{\beta} \ln(t_i)}{1+\gamma\, t_i^{\beta}} - \gamma\sum_{i=r+1}^{n} \frac{t_i^{*\beta} \ln(t_i^*)}{1+\gamma\, t_i^{*\beta}}. \quad (8)$$

Nilai estimator dari γ dan β, yaitu $\hat{\gamma}$ dan $\hat{\beta}$, dapat diperoleh dengan menyelesaikan sistem pers.(7) dan (8) secara simultan. Pemeriksaan apakah estimator $\hat{\gamma}$ dan $\hat{\beta}$ yang diperoleh dari pers.(7) dan (8) merupakan estimator yang memaksimumkan persamaan log-likelihood (6) adalah dengan memeriksa turunan kedua dari pers.(6) kurang dari nol atau tidak setelah disubstitusikan dengan $\hat{\gamma}$ dan $\hat{\beta}$. Jika nilainya kurang dari nol maka estimator-estimator yang diperoleh merupakan estimator yang maksimum atau sering disebut sebagai MLE, sebaliknya estimator tersebut bukan MLE. Bentuk umum turunan kedua dari pers.(6) terhadap γ dan β berturut-turut adalah

$$\frac{\partial^2 L_L(\gamma,\beta)}{\partial \gamma^2} = -\frac{r}{\gamma^2} - 2\sum_{i=1}^{r} \frac{-t_i^{2\beta}}{\left(1+\gamma\, t_i^{\beta}\right)^2} - \sum_{i=r+1}^{n} \frac{t_i^{*2\beta}}{\left(1+\gamma\, t_i^{*\beta}\right)^2} \quad (9)$$

dan

$$\frac{\partial^2 L_L(\gamma,\beta)}{\partial \beta^2} = -\frac{r}{\beta^2} - 2\gamma\left(\sum_{i=1}^{r}\left(\frac{t_i^{\beta} \ln t_i \ln t_i}{1+\gamma\, t_i^{\beta}} \frac{\left(t_i^{\beta}\right)^2 \gamma \ln t_i \ln t_i}{\left(1+\gamma\, t_i^{\beta}\right)^2}\right)\right) - \gamma\left(\sum_{i=r+1}^{n}\left(\frac{t_i^{\beta} \ln t_i^* \ln t_i^*}{1+\gamma\, t_i^{*\beta}} \frac{\left(t_i^{\beta}\right)^2 \gamma \ln t_i^* \ln t_i^*}{\left(1+\gamma\, t_i^{*\beta}\right)^2}\right)\right). \quad (10)$$

Penyelesaian Secara Numerik

Penentuan estimator parameter $\hat{\gamma}$ dan $\hat{\beta}$ dengan menyelesaikan sistem pers.(7) dan (8) secara simultan sulit dilakukan dengan cara analitik, oleh karena itu pendekatan numerik dapat dijadikan sebagai cara alternatif. Metode numerik yang digunakan disini adalah iterasi Newton-Rhapson, yaitu suatu prosedur iterasi numerik yang dapat digunakan untuk menyelesaikan baik persamaan maupun sistem persamaan nonlinear. Langkah pertama iterasi Newton-Rhapson untuk menyelesaikan sistem persamaan nonlinear (7) dan (8) adalah dengan memisalkan pers.(7) dengan $f_1(\gamma,\beta)$ dan pers.(8) dengan $f_2(\gamma,\beta)$. Selanjutnya, ditentukan matriks Jacobian J dari sistem pers.(7) dan (8) yang berbentuk

$$\mathrm{J} = \begin{bmatrix} \frac{\partial f_1(\gamma,\beta)}{\partial \gamma} & \frac{\partial f_1(\gamma,\beta)}{\partial \beta} \\ \frac{\partial f_2(\gamma,\beta)}{\partial \gamma} & \frac{\partial f_2(\gamma,\beta)}{\partial \beta} \end{bmatrix}. \tag{11}$$

Langkah berikutnya adalah menentukan nilai estimasi awal dari parameter γ dan β, yaitu $\hat{\gamma}^o$ dan $\hat{\beta}^o$, sedemikan sehingga $f_1\left(\hat{\gamma}^o,\hat{\beta}^o\right)$ dan $f_2\left(\hat{\gamma}^o,\hat{\beta}^o\right)$ nilainya mendekati nol. Dimisalkan invers dari matriks Jacobian (11) adalah J^{-1} yang berbentuk

$$\mathrm{J}^{-1} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}, \tag{12}$$

dan nilai estimasi parameter γ dan β pada iterasi ke-k dimisalkan γ^k dan β^k dengan $k = 1,2,\ldots,m$, serta $f_1^k = f_1^k(\gamma^k,\beta^k)$ dan $f_2^k = f_2^k(\gamma^k,\beta^k)$ adalah nilai fungsi f_1 dan f_2 pada iterasi ke-k, maka aproksimasi dari parameter γ dan β pada iterasi ke-$k+1$ diberikan oleh

$$\gamma^{k+1} = \gamma^k - (b_{11}^k f_1^k + b_{12}^k f_2^k) \quad \text{dan} \tag{13}$$

$$\beta^{k+1} = \beta^k - (b_{21}^k f_1^k + b_{22}^k f_2^k). \tag{14}$$

Iterasi yang dilakukan dimulai dari nilai estimasi awal $\hat{\gamma}^o$ dan $\hat{\beta}^o$, selanjutnya iterasi berjalan meng-ikuti pers.(13) dan (14) yang pada kemudian iterasi dapat berhenti ketika f_1 dan f_2 sangat dekat dengan nol atau pada saat selisih antara dua iterasi yang berurutan hampir sama dengan nol, bahkan penghentian iterasi terkadang cukup subjektif karena dapat dihentikan pada iterasi ke-m yang mana nilai m ini telah ditentukan di awal.

Contoh Penerapan

Model yang dikembangkan disini adalah model *survival* dari data tersensor tipe I berdistribusi log-logistik. Misalkan terdapat 30 data *survival* yang diasumsikan mengikuti distribusi log-logistik, yaitu: 50, 56, 65, 66, 73, 77, 84, 86, 87, 119, 140, 140*, 153, 177, 181, 191, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*, 200*. Simbol "*" data tersebut adalah data tersensor kanan tipe I.

Menggunakan program yang telah dibuat dengan *software* Maple 18, nilai estimator MLE dari data tersebut ditentukan dengan iterasi Newton-Rhap-son. Berdasarkan perhitungan tersebut, nilai estimator MLE untuk kedua parameter γ dan β adalah $\hat{\gamma} = 0.000025484$ dan $\hat{\beta} = 2.01866$.

4 KESIMPULAN

Penentuan estimator $\hat{\gamma}$ dan $\hat{\beta}$ dari data tersensor kanan tipe I berdistribusi log-logistik dengan MLE dilakukan dengan menyelesaikan sistem persamaan nonlinear (7) dan (8) secara simultan. Kedua persamaan tersebut memuat aritmatika dan fungsi-fungsi yang rumit, seperti notasi sigma, polinomial, dan fungsi logaritma natural (ln) sehingga penyelesaian secara analitik sulit dilakukan. Iterasi Newton-Rhapson dapat digunakan sebagai alternatif penyelesaiannya. Berdasarkan contoh penerapan yang telah dilakukan, metode iterasi Newton-Rhapson cukup efektif dan cepat dalam penentuan nilai $\hat{\gamma}$ dan $\hat{\beta}$.

REFERENSI

[1] Bennet, S., 1983, Log-Logistic Regression Models for Survival Data, *Journal of the Royal Statistical Society*, vol. 32 (2): 165-171

[2] Gijbels, I., 2010, Censored Data, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2 (2): 178-188

[3] Lee, T., Elisa, dan Wang, W., John, 2003, *Statistical Methods for Survival Data Analysis 3rd ed.*, John Wiley & Sons, Hoboken

[4] Sari, D.R., 2011, Analisis Survival untuk Data Tersensor Tipe II Menggunakan Model Distribusi Log-Logistik, *Tugas Akhir*, FMIPA, Universitas Negeri Yogyakarta, Yogyakarta

[5] Zhao, G., 2008, Nonparametric and Parametric Survival Analysis of Censored Data, *Thesis*, Faculty of The Graduate School, University of North Carolina, Greensboro