METODE BUCKLEY-JAMES UNTUK ESTIMASI MODEL REGRESI LINIER PADA DATA TERSENSOR KANAN

Muhammad Bayu Nirwana
Sekolah Tinggi Ilmu Kesehatan Muhammadiyah Kudus
Email : mbnirwana@stikesmuhkudus.ac.id

ABSTRAK

Data tersensor merupakan permasalahan yang sering dihadapi pada penelitian yang berhubugan dengan lama waktu terjadinya suatu kejadian. Penggunaan analisis statistika tanpa memperhatikan komponen tersensor dapat mengakibatkan bias pada hasil analisis data yang diperoleh. Dalam analisis regresi linier, di mana variabel dependen adalah variabel yang mengandung data tersensor, analisis data tanpa memperhatikan komponen tersensor dapat mengakibatkan koefisien model regresi yang diperoleh tidak tepat. Hal tersebut akan menjadikan model regresi yang diperoleh tidak dapat digunakan untuk memodelkan dan memprediksi data dengan baik. Metode Buckley-James adalah salah satu metode yang dapat digunakan untuk mengatasi permasalahan tersebut. Estimasi dari metode Buckley-James mengubah titik tersensor pada data tersensor ke nilai ekspektasinya. Selanjutnya model regresi linier diestimasi dengan memberikan bobot pada metode *least square* menggunakan estimator Kaplan-Meier.

Kata kunci : Regresi Linear, Data Tersensor, Kaplan-Meier, *Least Square*.

PENDAHULUAN

Karakteristik dari penelitian yang berhubungan dengan lama waktu terjadinya suatu kejadian adalah munculnya data dengan pengamatan yang tidak lengkap. Salah satu dari jenis data dengan pengamatan yang tidak lengkap adalah data tersensor. Data tersensor banyak muncul dalam berbagai bidang penelitian. Dalam bidang kesehatan, analisis untuk data tersensor sering disebut dengan analisis survival atau analisis tahan hidup. Di bidang ekonomi disebut dengan analisis durasi, dan pada bidang teknik disebut dengan analsis reliabilitas.

Sebagai contoh, terdapat beberapa pasien kanker darah yang diamati selama periode waktu tertentu sampai pasien tersebut meninggal dunia karena kanker darah. Jika pasien masih hidup sampai akhir periode penelitian, maka waktu tahan hidup dari pasien tersebut merupakan data tersensor, karena waktu tahan hidup sampai pasien meninggal tidak diperoleh secara lengkap. Demikian pula jika ternyata pasien pasien meninggal dunia setelah periode penelitian berakhir, maka data yang diperoleh juga tidak lengkap karena periode penelitian sudah berakhir. Selain itu, apabila dalam periode penelitian pasien tidak melanjutkan penelitian, maka data yang diperoleh oleh peneliti untuk pasien tersebut juga tidak lengkap. Waktu tahan hidup pasien kanker darah baru dikatakan lengkap atau teramati jika pasien meninggal dunia di waktu antara awal periode penelitian sampai akhir periode penelitian.

Penentuan suatu pengamatan termasuk pengamatan tidak lengkap berupa data tersensor, perlu memperhatian waktu awal penelitan *(origin)* dan akhir penelitian *(end point)*, serta definisi kejadian *(event)* yang jelas. Kriteria suatu pengamatan termasuk data tersensor kanan yaitu penelitian berakhir, sehingga kemungkinan terdapat subjek yang belum mengalami kejadian *(study ends, no event)*. Selain itu, data tersensor juga dapat terjadi karena subjek tidak melanjutkan mengikuti penelitian ketika periode penelitian masih berjalan *(lost)*, dan subjek dikeluarkan dari penelitian *(withdraw)*.

Selain data tersensor kanan, terdapat pula data terssensor kiri dan tersensor interval. Data tersensor kiri terjadi apabila subjek telah mengalami kejadian sebelum periode waktu tertentu atau sebelum penelitian dimulai. Data tersensor interval jika informasi mengenai terjadinya suatu kejadian tidak diketahui secara pasti, hanya diketahui terjadi pada suatu interval waktu tertentu.

Analisis statistik untuk data tersensor harus memperhatikan juga informasi mengenai tersensornya data. Data yang tersensor tidak bisa diabaikan dalam analisis data. Oleh karena itu diperlukan metode statistik yang dapat mengakomodaisi informasi dari data tersensor. Estimasi fungsi survival merupakan salah satu metode agar dapat menangani informasi tersensornya data dengan baik.

Fungsi survival dinotasikan dengan $S(t)$ yaitu probabilitas suatu individu atau subjek akan bertahan lebih lama dari waktu t. Bertahan artinya adalah subjek tersebut dalam kondisi belum mengalami suatu kejadian *(event)*. Fungsi survival dapat diestimasi secara parametrik maupun nonparametrik. Dari estimasi fungsi survival yang diperoleh, selanjutnya dapat dihitung rata-rata tahan hidup *(mean time to failure)* subjek-subjek yang diamati.

Hubungan sebab akibat juga tidak lepas dari munculnya data tersensor. Dalam analisis survival dikenal model regresi Cox dan model uji hidup dipercepat *(accelerated failure time, AFT)* untuk menangani data tersensor pada variabel dependen. Namun, pemodelan dalam model regresi Cox dan AFT adalah memodelkan fungsi survival atau fungsi hazard dari variabel dependen, bukan nilai dari variabel dependen itu sendiri. Akibatnya, tidak diperoleh hubungan yang linier antara variabel dependen dengan variabel independen. Hubungan linier antara variabel dependen dan variabel independen dapat diperoleh menggunakan analisis regresi linier. Namun analisis regresi linier tidak dapat menangani masalah ketika data tersensor muncul. Penelitian ini membahas tentang metode Buckley-James [1] untuk estimasi model regresi linier dengan variabel dependen mengandung data tersensor.

Fungsi survival merupakan peluang suatu subjek bertahan lebih lama dari waktu t [4], yaitu

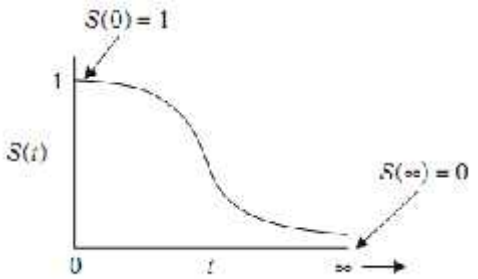
$$S(t) = P(T > t) \qquad (1)$$

Dengan $S(t)$ bernilai antara nol sampai satu, sedangkan t bernilai antara nol sampai tak hingga. Fungsi survival $S(t)$ merupakan fungsi *non-increasing* terhadap waktu t dengan sifat karakteristik-karakteristik yaitu:

1. Semakin besar t maka $S(t)$ semakin kecil.
2. Untuk $t = 0$ maka $S(t) = S(0) = 1$.
3. Untuk $t = \infty$ maka $S(t) = S(\infty) = 0$.

Fungsi survival dapat diestimasi secara parametrik maupun nonparametrik. Estimasi fungsi survival secara parametrik dapat dilakukan melalui hubungan fungsi survival dengan fungsi distribusi kumulatif yaitu

$$S(t) = 1 - F(t) \qquad (2)$$



Gambar 1. Grafik fungsi survial.

Rata-rata waktu tahan hidup (mean time to failure, MTTF) dari fungsi survival adalah luas area di bawah kurva $S(t)$ yang dapat dihitung dengan rumus

$$MTTF = E[T] = \int_0^\infty S(t)dt \qquad (3)$$

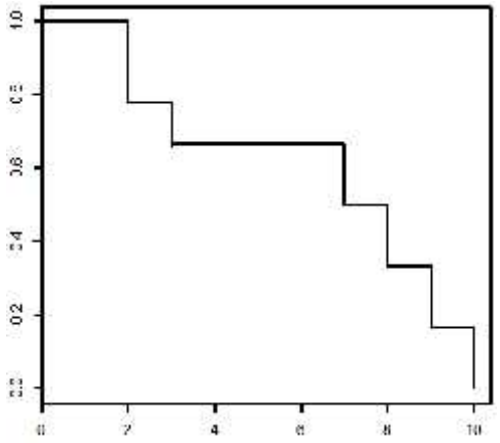
Salah satu metode estimasi fungsi survival secara nonparametrik yaitu dengan menggunakan metode Kaplan-Meier. Estimator Kaplan-Meier atau sering disebut sebagai *Product-Limit Estimator*, merupakan salah satu metode nonparametrik yang dapat digunakan untuk fungsi survival $S(t)$. Estimator Kaplan-Meier didefinisikan untuk semua nilai dalam rentang waktu t, ditunjukkan dengan:

$$\hat{S}(t) = \begin{cases} 1 & ;t < t_1 \\ \prod_{t_i \le t} \left(1 - \frac{d_i}{Y_i}\right) & ;t_i \le t \end{cases} \qquad (4)$$

dimana d_i adalah banyaknya *event* dan Y_i adalah banyaknya individu yang beresiko mengalami *event* [4]. Variansi dari estimator Kaplan-Meier dihitung menggunakan *Greenwood's Formula*

$$Var\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{Y_i(Y_i - d_i)} \qquad (5)$$

Estimasi fungsi survival dengan estimator Kaplan-Meier bernilai konstan pada setiap interval waktu di mana data lengkap diperoleh *(piecewise-constant)*. Hal ini membuat grafik dari estimator Kaplan-Meier membentuk kurva anak tangga *(step function)*.



Gambar 2. Grafik estimasi Kaplan-Meier.

METODE PENELITIAN

Sumber Data dan Variabel Penelitian

Data yang digunakan adalah data sekunder yang diambil dari penelitian dengan data penderita *Karsinoma Nesofaring* (KNF) yang terdaftar di bagian Telinga Hidung Tenggorok – Kepala Leher RSUP dr. Sardjito pada tahun 2008-2010 yang dilakukan oleh [3].

Analisis dimulai pada saat pasien pertama kali terdiagnosis KNF (tercatat di bagian rekam medis) dan diakhiri pada tanggal 1 Januari 2014. Pasien dinyatakan hidup apabila tanggal kematian pasien lebih dari atau sama dengan 36 bulan setelah tanggal pertama kali terdiagnosis, atau masih hidup sampai penelitian berakhir yaitu tanggal 1 Januari 2014.

Metode Analisis

Simulasi bertujuan untuk membandingkan performa antara regresi linier sederhana dengan regresi Buckley-James dalam menangani data tersensor. Simulasi dilakukan dengan membangkitkan data berdasarkan model regresi linear dengan satu variabel independen. Variabel independen *x* dan

error e dibangkitkan secara random, selanjutnya ditentukan koefisien regresi a dan b sehingga diperoleh nilai dari variabel dependen y.

Selanjutnya dilakukan sensor secara random sehingga variabel dependen y menjadi variabel yang mengandung data tersensor. Kemudian diestimasi model regresi untuk variabel dependen y dan variabel independen x untuk memperoleh koefisien a dan b untuk model regresi linier dan model regresi Buckley-James dengan beberapa replikasi. Koefisien a dan b yang diperoleh pada masing-masing model dibandingkan untuk mengetahui estimator yang memberikan koefisen regresi a dan b yang dekat dengan nilai a dan b yang ditentukan di awal simulasi.

HASIL PENELITIAN

Estimasi model Regresi Buckley-James

Metode Buckley-James [1] digunakan untuk mengestimasi model regresi linier pada data tersensor kanan. Estimasi dengan metode Buckley-James berdasarkan pada penyelesaian iteratif pada persamaan least square yang telah dimodifikasi untuk mengakomodasi data tersensor kanan pada variabel dependen.

Estimasi metode Buckley-James untuk suatu sampel sejumlah n individu atau subjek pada persamaan regresi linier

$$Y_i = \sim + \quad X_i + £_i$$

Selanjutnya dinotasikan $¢_i$ adalah indikator tersensor pada subjek i, di mana jika $¢_i = 1$ menunjukkan bahwa data pengamatan lengkap atau mendapatkan kejadian, dan $¢_i = 0$ menunjukkan bahwa data tersensor kanan pada suatu nilai T_i, diperoleh hubungan antara Y_i, T_i, dan $¢_i$

$$¢_i = \begin{cases} 0 & ;Y_i > T_i \\ 1 & ;Y_i \le T_i \end{cases}$$

Kemudian dibentuk Z_i dengan rumus

$$Z_i = \min\left(Y_i, T_i\right) \qquad (6)$$

Distribusi residual $£_i$ pada (4) tidak ditentukan mengikuti suatu distribusi parametrik tertentu. Estimasi distribusi dari residual $£_i$ dilakukan secara nonparametrik menggunakan estimator Kaplan-Meier (2).

Untuk mengakomodasi data tersensor, variabel dependen dimodifikasi dengan mengubah titik tersensor pada data tersensor dengan nilai ekspektasinya [5], yaitu

$$E\left[Y_i \mid Y_i > T_i\right] \qquad (7)$$

Sehingga dengan memberikan bobot pada variabel dependen Y_i berdasarkan status tersensornya dan menggunakan (7), diperoleh nilai baru dari Y_i yang sudah mengakomodasi data tersensor, $Y_i^*(b)$, yaitu

$$Y_i^*(b) = Y_i ¢_i + E\left[Y_i \mid Y_i > T_i\right]\left(1 - ¢_i\right) \qquad (8)$$

Notasi (b) pada $Y_i^*(b)$ di ruas kiri dari (8), menunjukkan bahwa nilai Y_i^* bergantung pada koefisien regresi b. Nilai Y_i^* akan terus diganti sesuai dengan koefisien regresi b, hingga b mencapai kekonvergenan. Dapat ditunjukkan bahwa.

$$E\left[Y_i^*(b)\right] = E\left[Y_i\right] = \sim + \quad X_i \qquad (9)$$

Untuk $¢_i = 0$ [2], bagian ekspektasi pada ruas kanan dari (9) menjadi

$$\begin{aligned} E\left[Y_i \mid Y_i > T_i\right] = {} & \sim + \quad X_i + \\ & E\left[£_i \mid £_i > T_i - \sim - \quad X_i\right] \end{aligned} \qquad (10)$$

Berdasarkan estimasi $E\left[Y_i \mid Y_i > T_i\right]$ pada (10), nilai $Y_i^*(b)$ pada (8) menjadi

$$Y_i^*(b) = \sim + \quad X_i + \left(£_i(b)¢_i + E\left[£_i(b) \mid £_i(b) > c_i(b)\right](1-¢_i)\right) \quad (1)$$

dengan $£_i(b) = Y_i - \sim - \quad X_i$ dan $c_i(b) = T_i - \sim - \quad X_i$.

Estimasi $E\left[£_i(b) \mid £_i(b) > c_i(b)\right]$ merupakan jumlahan terbobot dari residual data tidak tersensor yang lebih besar dari $c_i(b)$. Bobot yang diberikan berasal dari estimator Kaplan-Meier $\widehat{S}(t)$ yang dihitung dari residual $e_i(b) = Z_i - \sim - \quad X_i$ dengan Z_i seperti pada (6). Pembobotan dilakukan dengan terlebih dahulu mengurutkan nilai e_i dari nilai terkecil ke nilai terbesar. Selanjutnya diperoleh estimasi dari $E\left[£_i(b) \mid £_i(b) > c_i(b)\right]$ yaitu

$$\widehat{E}\left[£_i(b) \mid £_i(b) > c_i(b)\right] = \sum_{k=1}^{n} w_{ik}(b)\, e_k(b) \quad (2)$$

Di mana

$$w_{ik}(b) = \begin{cases} \dfrac{f\left(e_k(b)\right)¢_k\left(1-¢_i\right)}{\hat{S}\left(e_i(b)\right)} & k > i \\ 0 & k \le i \end{cases} \quad (3)$$

dengan $f\left(e_k(b)\right)$ merupakan massa probabilitas pada e_k dari fungsi $F = 1 - \hat{S}$ dan $\hat{S}\left(e_i(b)\right)$ merupakan estimasi Kaplan-Meier pada residual $e_i(b)$.

Setelah nilai $Y_i^*(b)$ untuk masing-masing subjek diperoleh, dilakukan estimasi *least square* untuk memperoleh estimasi dari $= b$. Estimasi slope b merupakan solusi iterasi, karena variabel $Y_i^*(b)$ bergantung pada nilai estimasi b [5]. Nilai awal *(initial value)* dari slope b, dinotasikan dengan $b^{(0)}$. Hasil $b^{(0)}$ yang telah diperoleh digunakan kembali untuk mengestimasi nilai $Y_i^*(b)$. Lebih lanjut, estimasi dari b yaitu

$$\hat{b}^{(m+1)} = \frac{n\sum_{i=1}^{n} x_i y_i^*\left(b^{(m)}\right) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i^*\left(b^{(m)}\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (11)$$

Di mana $m = 0,1,2,3,....$ dan $b^{(m)}$ adalah estimasi untuk iterasi ke-m. Iterasi dilakukan sampai konvergensi diperoleh, yaitu $\left|b^{(m+1)} - b^{(m)}\right|$ cukup kecil. Estimasi b didapatkan ketika (11) telah mencapai kekonvergen, dan dinotasikan dengan b^*. Selanjutnya, dari b^* dapat dihitung estimasi $\sim = a$ yaitu

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i^*\left(b^*\right) - b^* x_i}{n} \quad (12)$$

Dengan $y_i^*\left(b^*\right)$ adalah nilai akhir dari ekspektasi $y_i^*(b)$ setelah melalui proses iterasi.

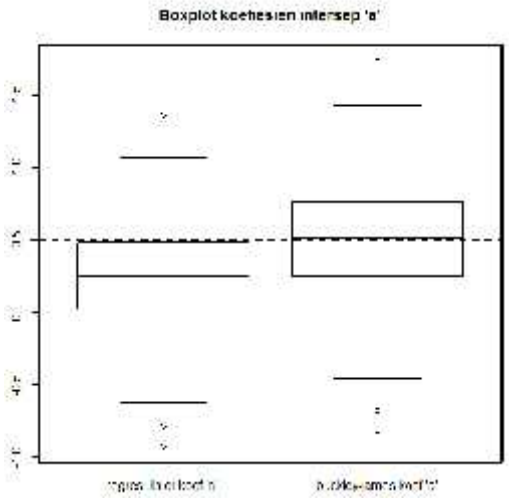
Simulasi model Regresi Buckley-James

Langkah-langkah dalam simulasi data untuk membandingkan performa regresi linier dengan regresi Buckley-James dalam mengakomodasi data tersensor:

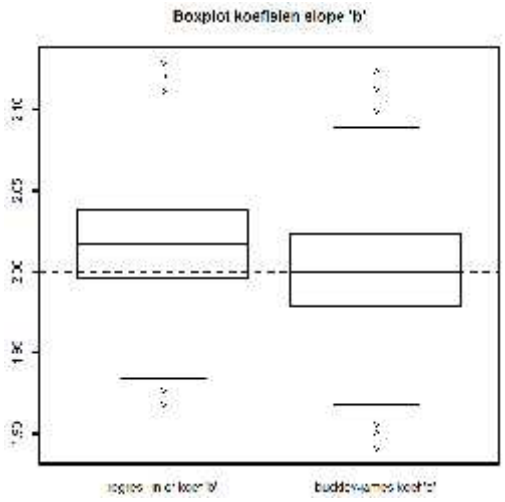
1. Bangkitkan x dengan $x \sqcup Unif(5,15)$.
2. Bangkitkan e_1 dengan $e_1 \sqcup N(0,2)$.
3. Hitung y berdasarkan model $y_i = \sim + \quad x_i + e_{1i}$, dengan $\sim = 0,5$ dan $= 2$.
4. Bangkitkan e_2 dengan $e_2 \sqcup N(0,2)$.
5. Hitung d berdasarkan model $d_i = \sim + \quad x_i + e_{2i}$, dengan $\sim = 0,5$ dan $= 2$.
6. Ambil nilai z dengan syarat $z_i = \min\left(y_i, d_i\right)$ z_i= min(y_i; d_i).

7. Bentuk indikator tersensor $¢_i = 1$ jika $z_i = y_i$ dan $¢_i = 0$ jika $z_i = d_i$.
8. Estimasi nilai a dan b dari:
   a. RegresiLinear x mempengaruhi z.
   b. Regresi Buckley-James x mempengaruhi z.
9. Lakukan replikasi untuk langkah ke-2 sampai ke-8.
10. Buat boxplot dari hasil langkah ke-9 untuk membandingkan hasil ketiga estimator.

Hasil boxplot estimasi dengan menggunakan regresi linear (linear model) dengan estimasi dengan regresi Buckley-James.



Gambar 3. Boxplot koefisien intersep model regresi linier dan regresi Buckley-James.



Gambar 4. Boxplot koefisien slope model regresi linier dan regresi Buckley-James.

Gambar 3 dan Gambar 4 di atas menunjukan boxplot estimasi koefisien a dan b dari simulasi data antara regresi linier klasik dengan regresi Buckley-James.

Regresi linear klasik kurang baik untuk mengestimasi data jika terdapat data tersensor. Terlihat dari boxplot untuk a dan b di atas, nilai median koefisien a dan b yang diestimasi regresi linier, jauh dari nilai parameter yang telah ditetapkan sebelumnya yaitu $a = 0,5$ dan $b = 2$.

Sedangkan pada boxplot hasil estimasi dengan menggunakan regresi Buckley-James, sangat baik digunakan untuk memodelkan data tersensor karena nilai median hasil estimasi koefisien regresi sangat dekat dengan nilai parameter yang ditetapkan sebelumnya yaitu $a = 0,5$ dan $b = 2$.

Tabel 1. Rata-rata dan variansi dari replikasi koefisien regresi hasil simulasi

	Rata-Rata	Variansi
Koef a regresi linier	0,2588	0,1145
Koef b regresi linier	2,0173	0,0010
Koef a regresi BJ	0,4967	0,1430
Koef b regresi BJ	2,0002	0,0012

Tabel 1 menyajikan rata-rata dan variansi dari replikasi hasil simulasi koefisien regresi linier dan regresi Buckley-James. Dari tabel 1 dapat dilihat bahwa rata-rata hasil simulasi untuk koefisien intersep pada regresi linier sebesar 0,2588 relatif jauh dari nilai intersep yang ditentukan di awal simulasi yaitu 0,5. Namun untuk koefisien slope, sebesar 2,0173 sudah mendekati nilai simulasi yang ditentukan di awal yaitu 2. Sedangkan untuk rata-rata koefisien hasil simulasi dari regresi Buckley-James, masing-masing untuk intersep sebesar 0,4967 dan slope sebesar 2,0002 sudah mendekati nilai intersep dan slope yang ditentukan di awal simulasi. Dari rata-rata hasil simulasi juga diperoleh bahwa regresi Buckley-James dapat

mengakomodasi adanya data tersensor dengan baik.

Studi Kasus dan Analisis Data

Studi kasus dan analisa data pada bagian ini menggunakan data dari penelitian [3]. Dalam studi kasus ini diambil variabel waktu (dalam bulan) sebagai variabel dependen, dan usia (dalam tahun) sebagai variabel independen. Diperoleh hasil sebagai berikut

Tabel 2. Estimasi koefisien regresi Buckley-James

	Koef	S.E. Koef	Wald Z	Pr (>|Z|)
Intercep	72,579	8,990	8,07	<0,0001
Usia	-0,901	0,177	-5,09	<0,0001

Tabel 2 menyajikan hasil estimasi model regresi Buckley-James. Sehingga model regresi buckley-james untuk data penderita KNF yaitu:

$$time = 72,579 - 0,901 * Usia$$

Nilai signifikansi untuk koefisien intersep dan usia berada di bawah nilai 0,05 sehingga koefisien intersep dan usia dapat masuk ke dalam model regresi.
Pada model tersebut, dapat disimpulkan bahwa waktu sampai seseorang mengalami kematian karena serangan jantung, dipengaruhi oleh usia. Bertambahnya 1 satuan usia akan semakin meningkatkan resiko mengalami event (meninggal), yaitu kenaikan 1 satuan usia pada pasien KNF, akan mengurangi angka harapan hidup sebesar 0,901 bulan. Singkatnya semakin tua seseorang semakin cepat pula seorang pasien akan mengalami event (meninggal).

KESIMPULAN

Berdasarkan hasil analisis dan pembahasan, dapat disimpulkan sebagai berikut :

1. Regresi linier kurang baik jika digunakan untuk memodelkan data tersensor.
2. Regresi Buckley-James dapat memodelkan data tersensor dengan baik. Selain itu regresi Buckley-James memodelkan variabel depeden dan variabel independen secara linier Sehingga interpretasi regresi Buckley-James lebih mudah dipahami.
3. Dari hasil simulasi terlihat bahwa model regresi Buckley-James dapat mengestimasi data tersensor dengan baik.

DAFTAR PUSTAKA

[1] Buckley, J., dan James, I., 1979, Linear regression with censored data. *Biometrika*, Vol. 66, 429-436.

[2] Cui, J., 2005, Buckley-James method for analyzing censored data, with an application to a cardiovascular disease and an HIV/AIDS study, *The Stata Journal*, Vol. 5, 517-526.

[3] Haroen, R. F., 2014, Angka Ketahanan Hidup 3 Tahun Pasien Karsinoma Nasofarings yang Mendapat Nonconcurrent Chemoradiotherapy di Bagian Telinga Hidung Tenggorok – Kepala Leher RSUP dr. Sardjito pada Tahun 2008-2010, *Skripsi*, Fakultas Kedokteran, Universitas Gadjah Mada, Yogyakarta.

[4] Kleinbaum, D. G., dan Klein, M., 2005, *Survival Analysis : A Self-Learning Text*, Springer, New York.

[5] Smith, P. J. 2002. *Analysis of Failure and Survival Data*. Boca Raton, Florida: Chapman & Hall/CRC.