

Pendugaan Model Survival Menggunakan Program SAS

Julio Adisantoso, julioipb@gmail.com

30 Januari 2010

Ringkasan

Model untuk waktu survival Y dapat menggunakan sebaran exponential, Weibull, log-logistic, log-normal, generalized gamma, logistic, dan normal. Disamping analisis deskripsi data, penentuan model terbaik merupakan salah satu aspek penting di dalam analisis pengambilan kesimpulan. Dengan banyak model yang mungkin, kurang efisien jika analisis dilakukan dengan menuliskan program SAS satu per satu. Makalah ini memberikan program SAS macro yang dapat digunakan untuk melakukan pendugaan model survival tahap demi tahap.

Dari data yang dicobakan yaitu hasil observasi terhadap 686 pasien kanker payudara yang dilakukan oleh *German Breast Cancer Study Group*, sebaran gamma memberikan model terbaik dan variabel kovariat yang dilibatkan dapat ditentukan pengaruhnya terhadap model. Dengan taraf nyata 5% hanya variabel ukuran tumor (*tsize*) dan kandungan estrogen (*estrec*) yang tidak memberikan pengaruh nyata pada waktu survival.

1 Pendahuluan

Analisis survival merupakan alat statistik yang tujuan utamanya adalah menganalisis data yang selalu positif dalam skala pengukuran dengan jarak interval data awal dan akhir yang panjang (McCullagh & Nelder, 1983). Data dengan karakteristik tidak lengkap atau tersensor dan fokus pada pendugaan parameter populasi *life data*, sehingga analisis yang digunakan adalah *life data analysis* (Nelson, 1982). Metode analisis survival yang menghubungkan antara waktu survival dengan variabel lain adalah model *hazard proporsional* dimana formulanya memungkinkan untuk interpretasi pengaruh dari masing-masing variabel bebas akan lebih mudah.

Model untuk waktu survival Y dapat menggunakan sebaran *exponential*, *Weibull*, *gamma*, *logistic*, *normal*, dan lainnya. Pendugaan model survival dapat dilakukan

melalui beberapa tahapan, diantaranya adalah melihat grafik data survival dan mengukur kesesuaian model. Ada beberapa ukuran statistik yang dapat digunakan dalam penentuan kebaikan model, di antaranya Pearson Chi-square, Deviance, Uji Rasio likelihood, dan uji lainnya. Makalah ini membahas tahapan pendugaan model survival dengan menggunakan program SAS dan menyusun program SAS macro untuk menganalisis data survival. Data yang dianalisis adalah hasil observasi terhadap 686 pasien kanker payudara yang dilakukan oleh *German Breast Cancer Study Group* dalam paket data program **R** (ipred) dengan sumber dari <http://www.blackwellpublishers.com/rss/Volumes/A162p1.htm>.

2 Analisis Survival

Analisis survival adalah analisis mengenai data yang diperoleh dari catatan waktu yang dicapai suatu obyek sampai terjadinya peristiwa gagal (*failure event*). Dalam menentukan waktu survival, Y , terdapat tiga elemen yang harus diperhatikan yaitu waktu awal (*time origin*), definisi *failure time* yang harus jelas, dan skala waktu sebagai satuan pengukuran. Perbedaan antara analisis survival dengan analisis statistik lainnya adalah adanya data tersensor. Menurut Pyke & Thompson (1986) data dikatakan tersensor jika pengamatan waktu survival hanya sebagian, tidak sampai *failure event*. Penyebab terjadinya data tersensor antara lain:

1. *Loss to follow up*, terjadi bila obyek pindah, meninggal atau menolak untuk berpartisipasi.
2. *Drop out*, terjadi bila perlakuan dihentikan karena alasan tertentu.
3. *Termination*, terjadi bila masa penelitian berakhir sementara obyek yang diobservasi belum mencapai *failure event*.

Jika Y melambangkan waktu survival dan mempunyai fungsi kepekatan peluang $f(y)$, maka fungsi sebaran kumulatif dinyatakan sebagai

$$F(y) = P(Y \leq y) = \int_0^y f(t) dt$$

yang merupakan peluang kejadian gagal sebelum waktu y .

Fungsi survival, $S(y)$, didefinisikan sebagai peluang suatu obyek bertahan setelah waktu ke- y , yaitu

$$S(y) = P(Y \geq y) = 1 - F(y) \quad (1)$$

Fungsi hazard merupakan laju kegagalan sesaat antara selang waktu yang sempit y dan $(y + \delta y)$ dengan asumsi obyek telah bertahan sampai waktu ke- y , yang

didefinisikan sebagai

$$\begin{aligned} h(y) &= \lim_{\delta y \rightarrow 0} \frac{P(y \leq Y < y + \delta y \mid Y > y)}{\delta y} \\ &= \lim_{\delta y \rightarrow 0} \frac{F(y + \delta y) - F(y)}{\delta y} \times \frac{1}{S(y)} \end{aligned}$$

Karena $f_Y(y)$ merupakan turunan pertama dari $F_Y(y)$ atau

$$\lim_{\delta y \rightarrow 0} \frac{F(y + \delta y) - F(y)}{\delta y} = f(y)$$

maka diperoleh

$$h(y) = \frac{f(y)}{S(y)} \quad (2)$$

Dari fungsi survival pada persamaan 1 diperoleh $F(y) = 1 - S(y)$ yang dapat dituliskan sebagai

$$\int f(y) dy = 1 - S(y)$$

dan jika diturunkan terhadap y maka diperoleh

$$f(y) = \frac{d(1 - S(y))}{dy} = -\frac{d}{dy} S(y)$$

Dengan demikian,

$$h(y) = \frac{-\frac{d}{dy} S(y)}{S(y)} \Leftrightarrow -h(y) dy = \frac{d(S(y))}{S(y)}$$

Dengan mengintegrasikan $h(y)$ diperoleh

$$\begin{aligned} -\int_0^y h(t) dt &= \int_0^y \frac{1}{S(t)} d(S(t)) \\ -H(y) &= \log[S(y)] \end{aligned}$$

atau

$$H(y) = -\log[S(y)] \quad (3)$$

yang disebut sebagai **fungsi kumulatif hazard**.

Nilai tengah waktu survival umumnya diduga dengan median dari sebaran karena karakteristik data yang memiliki kemiringan. Median waktu survival, $y(50)$, diperoleh dari jawaban persamaan $F(y) = \frac{1}{2}$.

2.1 Model Hazard Proporsional

Jika resiko gagal pada waktu tertentu bergantung pada nilai $x_1 \dots x_p$ dari p variabel kovariat, $X_1 \dots X_p$, maka nilai variabel tersebut diasumsikan telah tercatat sebagai *time origin*. Misalkan $h_0(y)$ sebagai fungsi hazard untuk setiap obyek dengan nilai dari semua variabel X adalah nol maka fungsi $h_0(y)$ dikatakan sebagai fungsi *baseline hazard* (Shuo-Jye Wu, 2002). **Model hazard proporsional** atau lebih dikenal dengan **regresi cox** adalah

$$h_1(y; \beta) = h_0(y) \exp \left(\sum_{i=1}^p x_i \beta_i \right) \quad (4)$$

dan fungsi kumulatif hazard diberikan oleh

$$H_1(y) = \int_0^y h_1(t) dt = \int_0^y h_0(t) e^{\mathbf{x}^T \beta} dt = H_0(y) e^{\mathbf{x}^T \beta}$$

sehingga

$$\log H_1(y) = \log H_0(y) + \sum_{i=1}^p x_i \beta_i \quad (5)$$

Fungsi hazard proporsional pada persamaan (5) menunjukkan suatu model linier terampat dengan *link function log*. McCullagh & Nelder (1989) menunjukkan bahwa ada beberapa *link function* yang umum digunakan, tergantung pada asumsi sebaran variabel respon y . Jika sebaran y adalah keluarga eksponensial seperti Normal, Gamma, Inverse Normal, dan Poisson maka *link function* yang dapat digunakan antara lain adalah:

- *Identity link* : $f(z) = z$
- *Log link* : $f(z) = \log(z)$
- *Power link* : $f(z) = z^a$ untuk nilai tertentu.

Sedangkan jika diasumsikan sebaran y adalah binomial atau multinomial maka dapat digunakan *link function*

- *Logit link* : $f(z) = \log(z/(1 - z))$
- *Probit link* : $f(z) = \phi^{-1}(z)$
- *Complementary log-log link* : $f(z) = \log(-\log(1 - z))$
- *Log-log link* : $f(z) = -\log(-\log(z))$

Jika variabel kovariat x bernilai biner, yaitu $x_k = 0$ untuk tanpa perlakuan dan $x_k = 1$ untuk perlakuan, maka **rasio Hazard** atau **Hazard relatif** antara ada dan tidak ada perlakuan adalah

$$\frac{h_1(y; \beta)}{h_0(y; \beta)} = e^{\beta_k}$$

menunjukkan bahwa nilai $\sum_{i \neq k} x_i \beta_i$ adalah konstan. Bentuk umum dari model tersebut seperti yang dituliskan pada persamaan (4).

2.2 Fungsi Survival Empirik

Fungsi survival empirik adalah penduga dari peluang survival lebih dari y , diberikan oleh persamaan

$$\hat{S}(y) = \frac{\text{banyaknya subyek dengan waktu survival} \geq y}{\text{total banyaknya subyek}}$$

Cara yang umum untuk menghitung fungsi survival empirik tersebut menggunakan penduga **Kaplan Meier**, yang juga disebut sebagai penduga **product limit**. Pertama dilakukan pengurutan secara menaik dari data waktu survival sehingga $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(k)}$. Jika n_j melambangkan banyaknya subyek yang hidup sebelum $y_{(j)}$ dan d_j melambangkan banyaknya kematian terjadi selama selang waktu yang kecil, $y_{(j)} - \delta$ sampai $y_{(j)}$, maka penduga Kaplan Meier untuk fungsi survival pada waktu y adalah

$$\hat{S}(y) = \prod_{j=1}^k \left(\frac{n_j - d_j}{n_j} \right) \quad (6)$$

untuk y antara $y_{(j)}$ dan $y_{(j+1)}$.

2.3 Pendugaan Model

Dalam analisis survival, model regresi parametrik dapat dituliskan dengan bentuk

$$Y = \beta_0 + \sum_{j=1}^m \beta_j x_j + \sigma \epsilon$$

dimana Y adalah waktu survival/gagal T atau $\log(T)$, x_j adalah variabel kovariat, ϵ adalah galat acak, dan β_j dan σ adalah parameter yang akan diduga. Pada program SAS, penduga maksimum likelihood dari parameter dapat dihitung dengan menggunakan PROC LIFEREG jika diketahui fungsi sebaran survival T (menggunakan pilihan **dist=** atau **d=** dalam pernyataan MODEL). Sebaran yang dapat digunakan adalah exponential (d=EXPONENTIAL), Weibull (d=WEIBULL),

log-logistic (d=LLOGISTIC), log-normal (d=LNORMAL), generalized gamma (d=GAMMA), logistic (d=LOGISTIC), dan normal (d=NORMAL). Jika tidak diberi pernyataan apapun, PROC LIFEREG menggunakan model $Y = \log(T)$. Pilihan NOLOG diberikan ke dalam model jika tidak ingin menggunakan transformasi logaritme.

Dengan demikian, semua kombinasi antara pilihan model dengan dan tanpa nolog menghasilkan 10 model parametrik yang berbeda, yaitu exponential, Weibull, gamma, normal, lnormal, logistic, llogistic, exponential nolog, Weibull nolog, dan gamma nolog. Gharibvand (2008) mengemukakan bahwa ada dua pendekatan yang dapat dilakukan untuk menentukan satu model terbaik, yaitu:

1. Model terbaik untuk fit data dipilih berdasarkan nilai terbesar dari maksimum log likelihood. Walaupun demikian, pemilihan model tidak dapat dilakukan secara sederhana seperti ini karena setiap model memiliki sebaran dengan jumlah parameter yang berbeda, mulai dari hanya satu parameter untuk exponential dan nilai ekstrim satu parameter hingga tiga parameter untuk generalized gamma dan log-gamma. Oleh karena itu, kriteria penentuan model terbaik perlu ditambah dengan informasi lainnya, antara lain yang umum digunakan adalah AIC (*Akaike Information Criterion*) dan SBC (*Schwarz's Bayesian Criterion*) yang juga dikenal dengan istilah BIC (*Bayesian Information Criterion*) dengan formula

$$-2 * (\text{Log-Likelihood}) + k * (p)$$

dimana p adalah banyaknya parameter, dan $k = 2$ untuk AIC atau $k = \log(n)$, n adalah banyaknya pengamatan, untuk SBC.

2. Pendekatan kedua berdasarkan pada uji rasio likelihood, misalnya exponential dengan Weibull karena exponential merupakan kasus khusus dari sebaran Weibull dengan parameter skala=1. Sebaran Weibull juga dapat dibandingkan dengan gamma karena Weibull merupakan kasus khusus dari gamma dimana parameter $shape=1$. Sebaran log-normal juga merupakan kasus khusus dari gamma dimana parameter $shape=0$.

Secara umum, jika $k_1 > k_2$ dan $LL(k_1)$ dan $LL(k_2)$ adalah maksimum likelihood dari dua model dengan semua k_1 parameter dan k_2 parameter tidak memenuhi hipotesis nol, maka $2[LL(k_1) - LL(k_2)]$ menyebar Chi-square dengan derajat bebas $k_1 - k_2$.

Setelah menentukan model terbaik dan menduga parameter model, fungsi sebaran survival (FSS) $S(t) = P(T > t)$ dapat ditentukan untuk setiap t .

3 Analisis Data

3.1 Bahan dan Metode

Sebagai uji coba, dilakukan analisis data terhadap 686 pasien kanker payudara yang dilakukan oleh *German Breast Cancer Study Group* dalam paket data program **R** (ipred). Terdapat enam variabel yang diamati, yaitu (1) **horTh** yaitu terapi hormonal yang terdiri atas dua level yaitu 'no' dan 'yes', (2) **age** adalah usia pasien (tahun), (3) **menostat** adalah status menopausal yang terdiri atas dua level yaitu 'pre' (premenopausal) dan 'post' (postmenopausal), (4) **tsize** adalah ukuran tumor (mm), (5) **tgrade** adalah stadium tumor (I, II, III), (6) **pnodes** adalah banyaknya node positif, (7) **progrec** adalah kandungan progesteron (fmol), (8) **estrec** adalah kandungan estrogen (fmol), (9) **time** adalah waktu survival (hari), dan (10) **sensor** adalah indikator sensor (0:tersensor, 1:tidak tersensor).

Tahapan analisis adalah dimulai dengan plot data survival untuk setiap strata untuk melihat perbedaan waktu survival antar strata. Tahapan selanjutnya adalah pemilihan model terbaik berdasarkan nilai maksimum log-likelihood, AIC dan SBC, dan selanjutnya adalah pendugaan fungsi sebaran survival. Seluruh tahapan analisis ini disusun dalam bentuk program SAS macro sehingga dapat digunakan dengan mudah untuk analisis data survival lainnya.

3.2 Hasil Analisis

Untuk keperluan ujicoba, data disusun dan dimasukkan sebagai dataset SAS sebagai berikut:

```
data gbsg;
  length horTh $6 menostat $9 tgrade $7;
  input no horTh $ age menostat $ tsize tgrade $ pnodes progrec estrec time censor;
  datalines;
    1   no  70   Post   21   II    3    48    66 1814   1
    2   yes 56   Post   12   II    7    61    77 2018   1
    3   yes 58   Post   35   II    9    52   271  712   1
    ...
 685  no  52   Post   23   II    3    15    34  727   1
 686  no  55   Post   23   II    9   116   15 1701   1 ;
run;
```

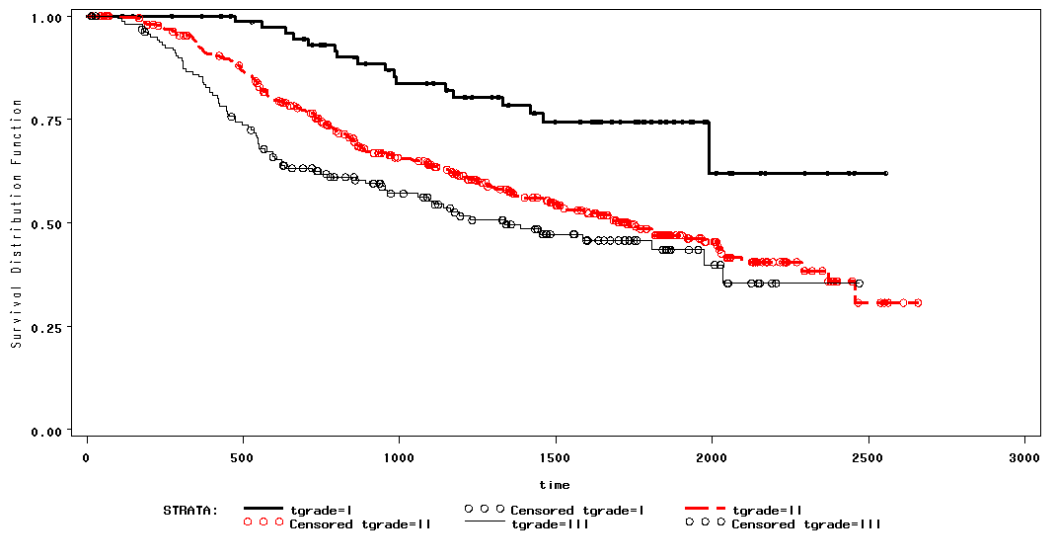
Untuk memudahkan pengolahan, setiap variabel kovariat harus berbentuk numerik sehingga perlu dilakukan proses transformasi seperti contoh berikut:

```
data gbsg_1;
set gbsg;
  if horTh = 'yes' then ghorTh = 1; else ghorTh = 0;
  if menostat = 'Post' then gmenostat = 1; else gmenostat = 0;
  if tgrade = 'III' then gtgrade = 3;
    else if tgrade='II' then gtgrade = 2; else gtgrade = 1;
run;
```

Tahapan awal analisis data dilakukan dengan menghitung deskripsi data dan membuat plot data survival. Program SAS macro disusun dan hasilnya disajikan pada Lampiran 1 dengan nama PLOT_DATA. SAS macro PLOT_DATA memiliki empat variabel masukan, yaitu (1) **dataset** adalah nama dataset yang akan dianalisis, (2) **time** adalah data waktu survival, (3) **ensor** adalah indikator data sensor (0 berarti data tersensor, dan 1 berarti data tidak tersensor), dan (4) **strata** adalah variabel strata perlakuan terhadap pasien yang merupakan data kategori. Sebagai contoh, dilakukan analisis plot data survival untuk setiap variabel **tgrade** (stadium tumor) dengan menuliskan program SAS sebagai berikut:

```
%PLOT_DATA(dataset=gbsg_1,time=time,censor=censor,strata=tgrade);
```

sehingga dihasilkan frekuensi data survival untuk setiap stadium tumor beserta plot frekuensi dan waktu survival (textbftime), dan plot survival untuk setiap stadium tumor seperti yang tercantum pada Gambar 1.

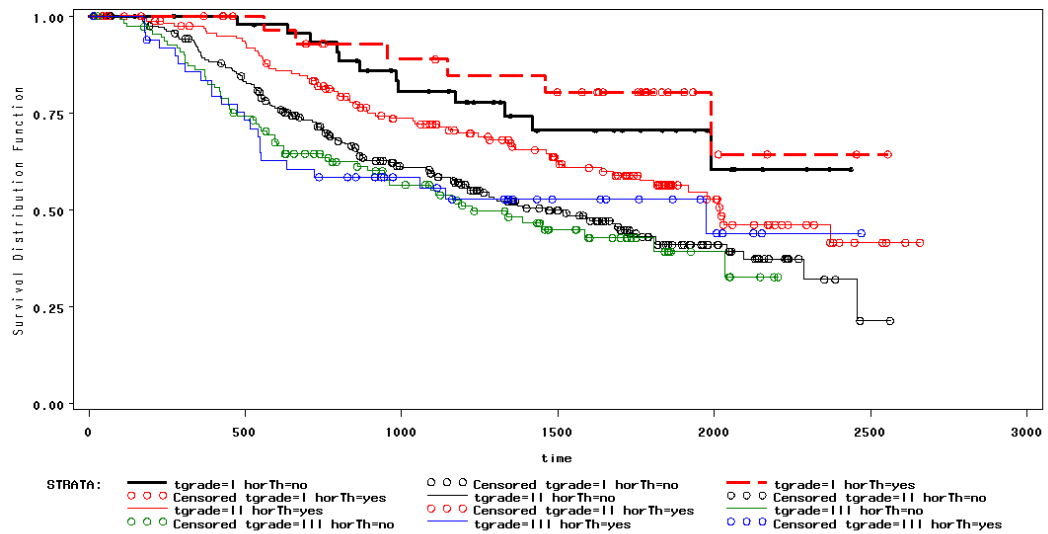


Gambar 1: Plot data waktu survival pasien kanker payudara untuk tiap stadium

Gambar 1 menunjukkan bahwa waktu hidup pasien kanker payudara dengan stadium I lebih tinggi dibanding stadium lanjut (II dan III). Hal ini menunjukkan

bahwa semakin dini deteksi penyakit kanker payudara maka pengaruh pengobatan semakin baik dan peluang untuk sembuh semakin tinggi. Sebagai contoh, jika dilihat plot waktu survival untuk setiap stadium tumor dan pemberian terapi hormonal maka semakin kuat dugaan bahwa deteksi dini dan pemberian terapi akan semakin menambah peluang pasien untuk sembuh (Gambar 2). Hal ini mudah dilakukan dengan menuliskan program SAS untuk memanggil macro yang telah disusun sebagai berikut:

```
%PLOT_DATA(dataset=gbsg_1,time=time,censor=censor,strata=tgrade horTh);
```



Gambar 2: Plot data waktu survival pasien kanker payudara untuk tiap stadium dan terapi hormonal

Untuk menentukan model survival terbaik, disusun program SAS macro untuk memudahkan perhitungan nilai AIC dan SBC untuk 10 model yang mungkin (Lampiran 2). Pada tahap awal dilakukan analisis dengan melibatkan semua variabel kovariat dengan memberikan program SAS sebagai berikut:

```
%AIC_SBC(dataset=gbsg_1,time=time,censor=censor,
covariates=ghorTh age gmenostat tsize gtgrade pnodes progrec estrec,mi=50);
```

Hasil analisis dengan melibatkan semua variabel untuk semua model yang mungkin menghasilkan nilai AIC dan SBC yang dihasilkan oleh program SAS macro %AIC_SBC sebagai berikut.

Model	Maximum log-likelihood	AIC	SBC
WEIBULL	-643.64	1307.28	1352.59
NORMAL	-2662.19	5344.39	5389.70
LNORMAL	-622.40	1264.80	1310.11
LOGISTIC	-2675.48	5370.97	5416.28
LLOGISTIC	-629.44	1278.88	1324.19
EXPONENTIAL	-663.47	1344.94	1385.72
GAMMA	-619.26	1260.53	1310.37
TWO_PARAM_EXTREME_VALUE	-2702.92	5425.84	5471.15
LOG_GAMMA	-2593.69	5209.39	5259.23

Satu model dengan nilai satu parameter ekstrim eksponensial tidak dapat dihasilkan karena terdapat kesalahan *floating point* pada fungsi penduga. Hal ini terjadi karena tidak dapat dihitung fungsi penduga untuk model eksponensial tanpa menggunakan transformasi log. Dari output yang dihasilkan terlihat bahwa model survival dengan sebaran gamma merupakan model terbaik jika melibatkan semua variabel kovariat ke dalam model.

Jika ingin melihat pengaruh variabel kovariat terhadap kesesuaian model dapat dilakukan dengan mudah, yaitu menjalankan program SAS macro dengan menuliskan variabel kovariat yang diinginkan, misalnya hanya melibatkan variabel terapi hormonal (ghorTh), usia (age), ukuran tumor (size), dan stadium tumor (gtgrade) sebagai berikut:

```
%AIC_SBC(dataset=gbsg_1,time=time,censor=censor,
covariates=ghorTh age tsize gtgrade ,mi=50);
```

dan berdasarkan nilai statistik kesesuaian model yang dihasilkan, model survival dengan sebaran gamma tetap merupakan model terbaik seperti berikut:

Model	Maximum log-likelihood	AIC	SBC
WEIBULL	-677.20	1366.40	1393.58
NORMAL	-2695.47	5402.94	5430.13
LNORMAL	-655.38	1322.75	1349.94
LOGISTIC	-2711.75	5435.49	5462.68
LLOGISTIC	-664.25	1340.50	1367.69
EXPONENTIAL	-691.21	1392.43	1415.08
GAMMA	-650.07	1314.14	1345.85
TWO_PARAM_EXTREME_VALUE	-2739.50	5491.00	5518.19
LOG_GAMMA	-2666.89	5347.79	5379.50

Disamping nilai-nilai statistik kesesuaian model, program SAS macro yang telah disusun juga menghasilkan nilai pendugaan parameter dari setiap model. Penggunaan dapat mengambil nilai penduga parameter untuk model terbaik saja seperti berikut:

Type III Analysis of Effects							
Wald							
Effect	DF	Chi-Square	Pr > ChiSq				
ghorTh	1	10.5666	0.0012				
age	1	4.1026	0.0428				
gmenostat	1	3.0985	0.0784				
tsize	1	2.6384	0.1043				
gtgrade	1	11.0567	0.0009				
pnodes	1	33.9322	<.0001				
progrec	1	15.8450	<.0001				
estrec	1	0.0006	0.9804				

Analysis of Parameter Estimates							
Parameter	DF	Estimate	Standard	95% Confidence		Chi-Square	Pr > ChiSq
			Error	Limits	Limits		
Intercept	1	7.3835	0.3815	6.6358	8.1312	374.58	<.0001
ghorTh	1	0.3174	0.0976	0.1260	0.5087	10.57	0.0012
age	1	0.0140	0.0069	0.0005	0.0276	4.10	0.0428
gmenostat	1	-0.2581	0.1466	-0.5455	0.0293	3.10	0.0784
tsize	1	-0.0053	0.0032	-0.0116	0.0011	2.64	0.1043
gtgrade	1	-0.2802	0.0843	-0.4454	-0.1150	11.06	0.0009
pnodes	1	-0.0554	0.0095	-0.0740	-0.0367	33.93	<.0001
progrec	1	0.0013	0.0003	0.0007	0.0020	15.84	<.0001
estrec	1	-0.0000	0.0003	-0.0007	0.0006	0.00	0.9804
Scale	1	1.0663	0.0519	0.9693	1.1730		
Shape	1	-0.5684	0.2297	-1.0186	-0.1183		

Dengan taraf nyata 5% hanya variabel ukuran tumor (tsize) dan kandungan estrogen (estrec) yang tidak memberikan pengaruh nyata pada waktu survival.

Analisis model survival dapat dilanjutkan dengan melakukan pendugaan model logistic-hazard dengan menggunakan program SAS macro LOGHAZ yang telah disusun (Lampiran 4). Contoh hasil untuk pendugaan model logistic-hazard dari data yang dicobakan dengan melibatkan semua variabel kovariat adalah seperti berikut:

```
%LOGHAZ(dataset=gbsg_1, knots=10 14 22 26 35 37, t=time, censor=censor,
covariates=ghorTh age gmenostat tsize gtgrade pnodes progrec estrec,
ps=0.2, seed=9, modeltxt=loghaz.txt);
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	410.984	364.821
SC	414.938	408.307
-2 Log L	408.984	342.821

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	66.1629	10	<.0001
Score	53.9344	10	<.0001
Wald	.	9	.

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

```

b1 = -999.991 * Intercept
b2 = -2743.91 * Intercept
b4 = 44523.1 * Intercept + 5.832 * b3
b6 = - 445134 * Intercept - 2.21E-6 * ghorTh + 2.16E-6 * gmenostat
      + 3.69E-6 * gtgrade - 46.5919 * b3 + 2.37037 * b5

```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15860.3	2015.0	61.9544	<.0001
ghorTh	1	-0.7436	0.2916	6.5006	0.0108
age	1	-0.0201	0.0217	0.8587	0.3541
gmenostat	1	0.3795	0.4529	0.7023	0.4020
tsize	1	0.0218	0.0129	2.8648	0.0905
gtgrade	1	0.2640	0.2564	1.0603	0.3031
pnodes	1	0.1980	0.0475	17.3768	<.0001
progrec	1	-0.00247	0.000901	7.5318	0.0061
estrec	1	0.00109	0.00101	1.1522	0.2831
b5	1	-0.0373	0.0470	0.6307	0.4271

4 Kesimpulan

Berdasarkan analisis yang telah dilakukan telah diperoleh beberapa kesimpulan, yaitu model untuk waktu survival dapat menggunakan beberapa model yaitu exponential, Weibull, log-logistic, log-normal, generalized gamma, logistic, dan normal. Program SAS macro yang telah disusun untuk plot data survival, menghitung statistik kesesuaian model, dan pendugaan parameter model sangat memudahkan pengguna dalam melakukan analisis model survival.

Dari data yang dicobakan, sebaran gamma memberikan model terbaik dan variabel kovariat yang dilibatkan dapat ditentukan pengaruhnya terhadap model. Dengan

taraf nyata 5% hanya variabel ukuran tumor (tsize) dan kandungan estrogen (estrec) yang tidak memberikan pengaruh nyata pada waktu survival.

5 Daftar Pustaka

- Al-Fawzan, M.A. 2000. *Methods for Estimating the Parameters of the Weibull Distribution*. King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia.
- Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. 2nd Ed. John Wiley and Sons, Inc.
- Dobson, A.J. 2001. *An Introduction to Generalized Linear Models*. Chapman Hall/CRC Texts in Statistical Science Series.
- Everitt, B.S. & T. Hothorn. 2006. *Survival Analysis: Glioma Treatment and Breast Cancer Survival*. A Handbook of Statistical Analysis Using R. CRAN-Project, R Packages Programming.
- First S. & K. Ronk. 2006. *SAS Macro Variables and Simple Macro Programs*. Systems Seminar Consultants, Madison, WI.
- Gharibvand, L. 2008. *A Step-by-Step Guide to Survival Analysis*. University of California, Riverside.
- Gharibvand, L; D.R.Jeske, & S.Liao. 2000. *Evaluation of a Hospice Care Referral Program Using Cox Proportional Hazards Model*. SAS Institute Inc.
- McCullagh,P. and Nelder,J.A. 1983. *Generalized Linear Models*. 2nd Ed. Chapman and Hall.
- Peters, A. & T. Hothorn. 2009. *Improved Predictors (Package ipred)*. CRAN-Project, R Packages Programming.

Southey, B.R.; S.L.Rodriguez-Zas; & K.A.Leymaster. 2003. *Discrete Time Survival Analysis of Lamb Mortality in a Terminal Sire Composite Population*. Journal of Animal Sciences 2003. 81:1399-1405.

Wei-Wang. 2004. *Proportional Hazards Regression Models with Unknown Link Function and Time-Dependent Covariates*. Statistica Sinica 14(2004), 885-905. Harvard University.

6 Lampiran

Lampiran 1. SAS macro untuk plot data survival

```

Deskripsi data dan plot data survival
DATASET - nama dataset
TIME - waktu survival
CENSOR - indikator data tersensor (0)
STRATA - variabel strata perlakuan terhadap pasien

%macro PLOT_DATA(dataset=,time=,censor=,strata=);
  proc freq data = &dataset;
    tables &strata*&censor*&time / list nocum nopercnt;
    ods output list = mylist (keep= &strata &censor &time frequency);
  run;
  goptions reset = all; /* hsize = 6 vsize = 2; */
  axis1 offset = (1,1) minor = none label=( ' ');
  symbol1 i = needle value = P font = marker h = .5 c=black w=3;
  symbol2 i = needle value = circle c = red w = 2 l=4;
  proc gplot data = mylist;
    by &strata;
    plot frequency*&time = &censor /vaxis= axis1;
  run;
  proc lifetest data = &dataset plot=(s);
    &time &time*&censor(0);
    strata &strata;
  run;
%mend PLOT_DATA;

```

Lampiran 2. SAS macro untuk menghitung AIC dan SBC

```

DATASET nama dataset
TIME nama variabel respon (survival)
CENSOR nama variabel untuk indikator sensor (0:tersensor, 1:tidak tersensor)
COVARIATES daftar nama variabel kovariat (masing-masing dipisahkan satu spasi)
MI - maksimum iterasi untuk menduga model

%macro AIC_SBC(dataset=,time=,censor=,covariates=,mi=50);
  * mi=maksimum iterasi, default 50;
  %if %length(&mi) = 0 %then %let mi=50;

  * PROC LIFEREG tanpa option NOLOG;
  %LET distr = exponential weibull gamma normal lnormal
              logistic llogistic;

  %do i=1 %to 7;
    proc lifereg data=&dataset outest=%scan(&distr,&i)_;
      %scan(&distr,&i): model &time*&censor(0)=&covariates /
      dist=%scan(&distr,&i) maxiter=&mi;
      title "Macro AIC_SBC";
    run;
  %end;

  * PROC LIFEREG dengan option NOLOG;
  proc lifereg data=&dataset outest=_EXPONENTIAL_NL_;
    one_param_EXTREME_VALUE: model &time*&censor(0)=
    &covariates / dist=exponential NOLOG maxiter=&mi;
  run;
  proc lifereg data=&dataset outest=_WEIBULL_NL_;
    two_param_EXTREME_VALUE: model &time*&censor(0)=
    &covariates / dist=weibull NOLOG maxiter=&mi;

```

```

run;
proc lifereg data=&dataset outest=_GAMMA_NL_;
  LOG_GAMMA: model &time*&cursor(0)=
    &covariates / dist=gamma NOLOG maxiter=&mi;
run;

* hitung banyaknya kovariat;
%LET nvar=0;
%do %while (%length(%scan(&covariates,&nvar+1))>0);
  %LET nvar=%eval(&nvar+1);
%end;

* hitung banyaknya pengamatan;
proc sql noprint;
  select count(*) into :n
  from &dataset
  %if %length(&covariates) = 0 %then
  where &time>.; %else %do;
  where
    nmiss(%do i=1 %to &nvar; %scan(&covariates,&i),
      %end;
      &time)=0
  %end;;
quit;

* hitung AIC dan SBC tiap model;
data _models_;
set _weibull_ _normal_ _lnormal_ _logistic_ _llogistic_
  _exponential_ (in=e) _gamma_ (in=g)
  _weibull_nl_ _exponential_nl_ (in=enl) _gamma_nl_ (in=gnl);
if e or enl then k=1; else if g or gnl then k=3; else k=2;
k=k+&nvar;
AIC=-2*_LNLIKE_ + 2*k;
SBC=-2*_LNLIKE_ + log(&n)*k;
label k='Number of parameters';
drop &time _type_ _name_;
run;

* cetak hasil;
proc print data=_models_ noobs label;
  label _LNLIKE_='Maximum log-likelihood'
  _MODEL_ ='Model';
  var _model_ _LNLIKE_ AIC SBC;
run;
%mend AIC_SBC;

```

Lampiran 3. SAS macro untuk menduga fungsi sebaran survival

```

DATASET nama dataset
PARAM nama datase yang dihasilkan oleh macro %AIC_SBC (default: _MODELS_)
MODEL model yang digunakan (LNORMAL, NORMAL, WEIBULL, LLOGISTIC, LOGISTIC,
  EXPONENTIAL, GAMMA, TWO_PARAM_EXTREME_VALUE,
  ONE_PARAM_EXTREME_VALUE, and LOG_GAMMA)
START, STEP, FINISH titik untuk menghitung fungsi sebaran survival
COVARIATES daftar nama variabel kovariat (masing-masing dipisahkan satu spasi)
OUT nama output dataset

%macro FSS (dataset=, param=_models_, model=, start=, step=, finish=, covariates=, out=);
  * hitung banyaknya variabel kovariat;
  %LET nvar=0;
  %do %while (%length(%scan(&covariates,&nvar+1))>0);
    %LET nvar=%eval(&nvar+1);
  %end;

```



```

* parameter model;
data _NULL_;
  set &param;
  if _MODEL_=upcase("&model");
  %if &nvar>0 %then array b (&nvar);;
  call symput('intercept', intercept);
  %do i=1 %to &nvar;
  call symput("B&i", %scan(&covariates,&i));
  %end;
  call symput('scale', _scale_);
  if _DIST_='Gamma' then call symput('shape', _shape1_);
  else do; %let shape=1; end;
run;

data &out;
  %if %length(&dataset) > 0 %then set &dataset;;
  %if &nvar>0 %then array covs (*) &covariates;;
  bx=&intercept;
  %do i=1 %to &nvar;
  bx=bx+&b&i*covs(&i);
  %end;
  length mod $ 23;
  call vname(&model,mod);
  %do t=&start %to &finish %by &step;
  select(upcase(mod));
  when("LNORMAL")
    survp&t=1-probnorm((log(&t)-bx)/&scale);
  when("NORMAL")
    survp&t=1-probnorm((&t-bx)/&scale);
  when("WEIBULL")
    survp&t=exp(-exp((log(&t)-bx)/&scale));
  when("LLOGISTIC")
    survp&t=1/(1+exp((log(&t)-bx)/&scale));
  when("LOGISTIC")
    survp&t=1/(1+exp((&t-bx)/&scale));
  when("EXPONENTIAL")
    survp&t=exp(-exp(log(&t)-bx));
  when("GAMMA")
    do; w=(log(&t)-bx)/&scale; ss=(&shape)**2;
    if &shape > 0 then
      survp&t=1-probgam(exp(&shape*w)/ss,1/ss);
    else
      survp&t=probgam(exp(&shape*w)/ss,1/ss);
    end;
  when("TWO_PARAM_EXTREME_VALUE")
    survp&t=exp(-exp((&t-bx)/&scale));
  when("ONE_PARAM_EXTREME_VALUE")
    survp&t=exp(-exp(&t-bx));
  when("LOG_GAMMA")
    do; w=(&t-bx)/&scale; ss=(&shape)**2;
    if &shape > 0 then
      survp&t=1-probgam(exp(&shape*w)/ss,1/ss);
    else
      survp&t=probgam(exp(&shape*w)/ss,1/ss);
    end;
  otherwise put "Macro %FSS: spesifikasi tidak sesuai";
  end;
  label survp&t="Probability of survival to &t";
%end;
drop bx w ss mod &model;
run;
%mend FSS;

```

Lampiran 4. SAS macro untuk menduga fungsi logistic-hazard

```

DATASET  nama dataset
KNOTS   knots pada cubic spline: tiap bilangan dipisahkan oleh spasi
T       nama variabel respon (survival)
CENSOR  nama variabel untuk indikator sensor (0:tersensor, 1:tidak tersensor)
COVARIATES  daftar nama variabel kovariat (masing-masing dipisahkan satu spasi)
PS      peluang terambil sensor (default:1)
SEED    bilangan inisialisasi awal untuk pembangkitan bilangan acak pemilihan sensor
MODELTXT  file untuk menyimpan hasil fungsi hazard

%macro LOGHAZ(dataset=, knots=, t=, censor=, covariates=, ps=1, seed=0,
modeltxt=);
  * setting the default value of 1 to ps;
  %if %length(&ps) = 0 %then %let ps=1;
  * setting the default value of 0 to seed;
  %if %length(&seed) = 0 %then %let seed=0;
  * counting #knots;
  %LET nk=0;
  %do %while (%length(%scan(&knots,&nk+1))>0);
    %LET nk=%eval(&nk+1);
  %end;

  data _sample_;
    set &dataset;
    array k{&nk} _temporary_ (&knots);
    if &ps=1 or
    (&censor=1 or (&censor=0 and ranuni(&seed)<=&ps));
    %do i=1 %to &nk;
      b&i=(&t>k[&i])*(&t-k[&i])**3-&t**3+3*k[&i]*&t**2-
      3*k[&i]**2*&t;
    %end;
  run;

  ods output parameterestimates=_pe_;
  proc logistic data=_sample_;
    model &censor(ref='0')= &covariates b1-b&nk;
    title "Macro LOGHAZ";
  run;

  * writing scoring code;
  data _null_;
    file "&modeltxt";
    length plus $4;
    set _pe_ end=last;
    if variable='Intercept' then variable='eta' ||'=1';
    plus='+'; if last then plus='';
    put variable '*' estimate best16. plus;
    if last then do;
      if &ps<1 then put "eta=eta+log(&ps)";
      put 'haz_function=exp(eta)/(1+exp(eta));';
    end;
  run;
%mend LOGHAZ

```