PERKAWINAN MODEL REGRESI DAN MODEL BERSTRUKTUR POHON BAGI PEMODELAN VARIABEL RESPON KETAHANAN HIDUP DENGAN RISIKO BERSAINGAN

[1] Abdul Kudus

[1] *Fakultas MIPA, Jurusan Statistika, Universitas Islam Bandung,*
Jl. Purnawarman No. 63 Bandung 40116

e-mail: [1] kudus@unisba.ac.id

Abstrak. *Pemodelan variabel respon ketahanan hidup biasanya menggunakan model regresi Cox. Adapun jika variabel respon ketahanan hidupnya disertai dengan keterangan lebih dari satu jenis penyebab kegagalan, maka ia disebut data ketahanan hidup dengan risiko-risiko bersaingan (competing risks) yang pemodelannya biasanya menggunakan model regresi subdistribusi kegagalan. Pemilihan model terbaik untuk model-model regresi subdistribusi kegagalan dilakukan dengan memilih model yang mempunyai nilai kriteria informasi Akaike (AIC) terkecil. Dalam makalah ini diajukan metode untuk meningkatkan kemampuan model regresi terbaik tersebut dengan cara mengawinkannya dengan model berstruktur pohon. Langkah ini dilatarbelakangi bahwasannya model regresi subdistribusi kegagalan dan model berstruktur pohon adalah bersifat saling melengkapi satu sama lain. Langkah pertama adalah mencari model regresi subdistribusi kegagalan yang terbaik, lalu menggunakan model berstruktur pohon untuk memodelkan komponen sistematik yang masih tertinggal oleh model regresi. Penerapan model baru ini pada data berhentinya pemakaian alat kontrasepsi menunjukkan bahwa model yang menjelaskan lama waktu bertahannya pengguna alat kontrasepsi sampai dengan mengganti alat tersebut dengan alat kontrasepsi lain ternyata dapat ditingkatkan kemampuannya setelah dikawinkan dengan model berstruktur pohon yang ditunjukkan dengan makin kecilnya nilai AIC.*

Kata kunci : AIC, berhenti memakai alat kontrasepsi, model berstruktur pohon, perkawinan model, model regresi subdistribusi.

1. Pendahuluan

Kajian waktu ketahanan hidup dengan risiko-risiko bersaingan (*competing risks*) merupakan kajian yang umum dilakukan dalam bidang kedokteran dan rekayasa. Dalam kajian ini, berhenti bertahannya suatu individu yang diteliti dapat disebabkan oleh beberapa penyebab, akan tetapi hanya satu penyebab saja yang benar-benar menjadi penyebab kegagalan obyek tersebut. Oleh karena itu, hal ini dikenal dengan istilah risiko-risiko yang bersaingan. Dalam pelaksanaan penelitiannya dilakukan pencatatan waktu ketahanan hidup beserta penyebab kegagalan bertahan dari setiap individu.

Dalam melakukan kajian terhadap data risiko-risiko bersaingan, biasanya menggunakan fungsi subdistribusi sebagai ukuran kemungkinan terjadinya penyebab tertentu. Dimana pada titik waktu *t* tertentu, fungsi subdistribusi ini memberikan nilai peluang terjadinya kegagalan oleh karena penyebab tertentu sampai titik waktu tersebut.

Dalam makalah ini, sebuah model baru dikemukakan. Model baru ini merupakan perkawinan dari model regresi subdistribusi kegagalan (Fine dan Gray, 1999) dengan model berstruktur pohon (Ibrahim *et al.*, 2008). Hal ini terinspirasi oleh Su dan Tsai (2005) yang mengemukakan model serupa untuk data ketahanan hidup dengan satu jenis penyebab kegagalan.

Motivasi untuk melakukan penelitian ini berasal dari munculnya pandangan bahwa model regresi subdistribusi kegagalan dan model berstruktur pohon bersifat saling melengkapi satu sama lain dalam beberapa aspek, antara lain: 1) model regresi subdistribusi kegagalan merupakan metode statistika untuk memodelkan hubungan linier antara transformasi log-log komplementer dari fungsi subdistribusi dengan variabel-variabel penjelasnya, sementara di lain pihak model berstruktur pohon tidak efisien untuk memodelkan linieritas. 2) model berstruktur pohon sangat baik dalam menangani variabel penjelas kategorik, sedangkan model regresi subdistribusi kegagalan hanya mampu menangani variabel dummy dari variabel penjelas kategorik tersebut, yang pada akhirnya akan membuat model regresinya menjadi kompleks, apalagi jika jumlah variabel penjelas kategoriknya banyak. 3) model regresi subdistribusi seringkali gagal untuk memodelkan nonlinieritas, sedangkan metode berstruktur pohon dengan kemampuan fungsi tangganya merupakan pendekatan yang baik untuk memodelkan nonlinieritas. 4) cara untuk mendeteksi interaksi antarvariabel penjelas dalam regresi subdistribusi merupakan pekerjaan yang sangat sulit, sedangkan regresi berstruktur pohon akan menangani masalah pendeteksian interaksi tersebut secara otomatis. Oleh karena itu, maka sangat masuk akal jika kedua model statistika ini dikawinkan untuk mendapatkan model yang lebih baik.

Metode untuk mengawinkan dua model statistika yang sudah ada ini dilakukan dengan cara menambahkan komponen regresi berstruktur pohon terhadap model regresi subdistribusi. Pertama-tama dilakukan pemodelan regresi subdistribusi, lalu cari model terbaiknya. Kemudian gunakan model berstruktur pohon untuk memodelkan komponen sistematik yang tersisa yang masih luput dimodelkan oleh model regresi subdistribusi. Makalah ini disusun dengan sistematika sebagai berikut; bagian 2 membahas mengenai metode pendugaan model hasil perkawinan. Di bagian 3 dibahas tentang penerapan model yang sudah dibangun pada data berhentinya pemakaian alat kontrasepsi.

2. Perkawinan Model

2.1 Struktur Model

Misalkan terdapat J jenis penyebab kegagalan, dengan waktu kegagalan potensialnya (X_{1i}, X_{2i},…, X_{Ji}) untuk setiap individu i. Kegagalan bagi individu ke-i terjadi pada waktu $T_i^* = \min(X_{1i}, X_{2i},\ldots, X_{Ji})$, dimana kegagalan itu terjadi karena penyebab $\delta_i^* = j$ jika $T_i^* = X_{ji}$, $j = 1,...,J$, yakni yang menjadi penyebab paling awal. Misalkan pula bahwa pemantauan terhadap individu i berhenti pada titik waktu C_i, sehingga data hasil pengamatannya adalah lamanya waktu ketahanan individu tersebut atau lamanya waktu sampai penelitian berhenti $T_i = \min\left(T_i^*, C_i\right)$ beserta indikator jenis kegagalan δ_i, dimana $\delta_i = \delta_i^*$ jika $T_i^* \le C_i$, dan $\delta_i = 0$ selainnya. Diasumsikan pula bahwa antarindividu saling bebas dan juga C_i saling bebas dengan T_i^*. Di samping itu

data dilakukan pencatatan terhadap variabel-variabel penjelas $Z_i = (Z_{1i}, Z_{2i},\ldots, Z_{pi},) \in \Re^p$, yang mungkin terdiri atas variabel-variabel acak kontinu dan diskrit. Sehingga data yang dipunyai terdiri atas n individu (T_i,δ_i,Z_i) yang saling bebas. Misalkan F_j adalah fungsi subdistribusi bagi penyebab kegagalan j, yakni peluang kumulatif terjadinya kegagalan oleh karena penyebab j sampai titik waktu tertentu dan model yang sebenarnya dimisalkan sebagai

$$\log\left[-\log\left(1-F_j\left(t_i;\boldsymbol{Z}_i\right)\right)\right]=\tilde{\lambda}_{j0}^{*}\left(t_i\right)+g\left(\boldsymbol{Z}_i\right) \qquad ...(1)$$

dan model regresi subdistribusi terbaiknya adalah

$$\log\left[-\log\left(1-F_j\left(t_i;\boldsymbol{Z}_i\right)\right)\right]=\tilde{\lambda}_{j0}^{*}\left(t_i\right)+\boldsymbol{\beta}_j^T\boldsymbol{Z}_i^0 \qquad ...(2)$$

dimana $\boldsymbol{z}_i^0 \in \Re^q$, $q < p$, merupakan vektor berdimensi- q yang berasal dari $\boldsymbol{Z}_i$. Misalkan parameter model regresi tersebut diduga dengan metode kemungkinan maksimum, yang fungsi kemungkinannya

$$l\left(\boldsymbol{\beta}_j\right)=\sum_{i=1}^{n}I\left(\delta_i=j\right)\left(\boldsymbol{\beta}_j^T\boldsymbol{Z}_i^0-\log\sum_{i'\in R(t_{i'})}\exp\left(\boldsymbol{\beta}_j^T\boldsymbol{Z}_{i'}^0\right)\right) \qquad ...(3)$$

Model terbaik dicari dari sekumpulan model dengan variabel penjelasnya $\boldsymbol{Z}_i^0 \subseteq \boldsymbol{Z}_i$ berdasarkan kriteria informasi Akaike, *AIC* (Sakamoto *et al.*, 1986):

$$AIC=-2l\left(\hat{\boldsymbol{\beta}}_j\right)+2q \qquad ...(4)$$

dimana model terbaik adalah yang mempunyai *AIC* terkecil.

Akan tetapi, model (2) mungkin masih belum memuaskan. Untuk meningkatkan kemampuan model regresi, dilakukan pembentukan model berikut

$$\log\left[-\log\left(1-F_j\left(t_i;\boldsymbol{Z}_i\right)\right)\right]=\tilde{\lambda}_{j0}^{*}\left(t_i\right)+\boldsymbol{\beta}_j^T\boldsymbol{Z}_i^0+\alpha_j^T\boldsymbol{Z}_i^{(\Gamma)} \qquad ...(5)$$

dimana vektor $\boldsymbol{Z}_i^{(\Gamma)}$ berasal dari regresi berstruktur pohon Γ. Struktur pohon Γ memberikan pendekatan fungsi tangga bagi komponen $g\left(\boldsymbol{Z}_i\right)-\beta_j^T\boldsymbol{Z}_i^0$, yakni sisa komponen sistematik yang luput dimodelkan oleh model subdistribusi. Keuntungan dengan mengawinkan kedua model ini adalah:

1. Model subdistribusi (2) menangkap pola global dari data, sedangkan model berstruktur pohon mendeteksi sifat lokal dari data yang luput dimodelkan dengan model (2). Sifat lokal tersebut seperti misalnya pola nonlinier dan interaksi yang kompleks.
2. Struktur pohon Γ tidak hanya menjadi alat diagnostik bagi model subdistribusi (2), tapi juga memberikan masukan mengenai bagaimana cara untuk memperbaiki model tersebut.
3. Model (5) meningkatkan kemampuan dari model subdistribusi.

2.2 Algoritme Perkawinan Model

Untuk membentuk struktur pohon Γ, kita mulai dengan model subditribusi terbaik (2) dan lakukan langkah pembentukan regresi berstruktur pohon (Breiman *et al*., 1984), yang terdiri atas tiga tahap: (i) pembentukan pohon awal yang besar, Γ_0, (ii) pemangkasan terhadap Γ_0 yang menghasilkan sekumpulan subpohon yang tersarang, dan (iii) pemilihan pohon dengan ukuran yang optimal dari sekumpulan subpohon yang dihasilkan pada tahap (ii). Tahapan tersebut secara rinci dijelaskan di bawah ini.

A. Pembentukan Pohon Awal yang Besar

Pemilahan data didasarkan pada model berikut:

$$\log\left[-\log\left(1-F_j\left(t_i;\mathbf{Z}_i\right)\right)\right]=\tilde{\lambda}^{*}_{j0}\left(t_i\right)+\boldsymbol{\beta}_j^T\mathbf{Z}_i^0+\alpha I\left(Z_{ki}<\gamma\right) \qquad ...(6)$$

Fungsi indikator $I(Z_{ki} < \gamma)$ berkaitan dengan pemilahan data berdasarkan variabel penjelas kontinu Z_k. Jika variabel penjelasnya berjenis diskrit dengan nilai-nilainya merupakan anggota himpunan $D = \{d_1,\ldots, d_r\}$, maka kita akan melakukan pemilahan bagi sebarang bentuk $I(Z_{ki} \in A)$ dengan $A \subset D$.

Pemilahan terbaik s^* adalah yang berpadanan dengan nilai devians terkecil dari hasil pendugaan model (6) dengan variabel-variabel penjelas yang ada. Setelah data dipilah menjadi dua bagian, kemudian kita ulangi pencarian pemilahan terbaik pada setiap bagian data tadi. Lakukan hal tersebut secara berulang dan akhirnya akan didapatkan pohon awal Γ_0 yang besar.

B. Pemangkasan

Setelah dibuat pohon awal Γ_0 yang besar, kemudian lakukan pemangkasan sesuai dengan algoritme Segal (Segal, 1988). Langkah ini akan menghasilkan sekumpulan subpohon yang tersarang. Sebuah pohon terbaik akan dipilih dari kumpulan subpohon tersebut. Algoritmenya sebagai berikut:

- Mulanya dibentuk pohon awal yang besar Γ_0.
- Terhadap setiap bagian data dari pohon tersebut (yang dinyatakan dengan simpul dalam), padankan statistik devians terbesar yang dikandung oleh dahannya. Statistik tersebut menggambarkan kekuatan dahan itu.
- Diantara simpul-simpul dalam tersebut, pilihlah simpul yang mempunyai padanan statistik devians terkecil. Artinya, pilihlah dahan yang paling lemah untuk kemudian dipangkas. Langkah ini akan menghasilkan pohon dengan satu dahan terpangkas.
- Pohon terpangkas selanjutnya dapat diperoleh dengan cara yang sama dengan menerapkan dua tahap di atas terhadap pohon dengan satu dahan terpangkas tadi..
- Ulangi proses ini sampai dengan diperoleh pohon terpangkas yang hanya terdiri atas simpul utama. Hasilnya adalah sekumpulan subpohon yang tersarang..

Pohon terbaik diperoleh dengan melakukan pemplotan antara jumlah simpul akhir yang dipunyai pohon dengan statistik devians terkecil yang dikandungnya. Posisi segmen patah dari plot tersebut menunjukkan letak pohon terbaik.

Bagi struktur pohon Γ dalam model (5), misalkan $\tilde{\Gamma}$ menyatakan himpunan semua simpul akhir dari pohon Γ dan $|\cdot|$ menyatakan banyaknya anggota himpunan. Lebih lanjut, didefinisikan matriks $\boldsymbol{Z}^{(\Gamma)}$ yang berukuran $n \times |\tilde{\Gamma}|$ sebagai berikut

$$Z_{hi}^{(\Gamma)} = \begin{cases} 1, & \text{jika observasi ke-}i \in \text{simpul luar ke-}h \\ 0, & \text{lainnya} \end{cases}$$

Model terbaik (2) digandengkan dengan pohon terbaik Γ untuk membentuk model hasil perkawinan (5).

3. Contoh Penerapan: Data Berhenti Memakai Alat Kontrasepsi

Metode yang dibangun akan diterapkan pada data berhentinya memakai alat kontrasepsi yang diambil dari hasil Survey Demografi dan Kesehatan Indonesia (SDKI) tahun 2002. Variabel responnya adalah lamanya waktu bertahan memakai alat kontrasepsi tertentu. Dimana seorang akseptor KB mungkin gagal bertahan disebabkan karena rusak, dilepas atau diganti dengan alat kontrasepsi lain. Definisi gagal karena rusak adalah ketika si pemakai dilaporkan menjadi hamil sewaktu dia masih menggunakan alat kontrasepsi tertentu. Jadi data ini merupakan data ketahanan hidup dengan risiko-risiko bersaingan yang mempunyai tiga penyebab kegagalan (j = 1, 2, 3). Di samping itu juga dicatat enam variabel penjelas yang dianggap dapat menjelaskan peluang terjadinya berhenti memakai alat kontrasepsi (Tabel 1).

Tabel 1.
Deskripsi variabel penjelas untuk data berhenti memakai alat kontrasepsi

Nama variabel	Deskripsi
1. soseko	status sosial ekonomi rumah tangga (skor 1-7)
2. umur	umur pada saat mulai memakai alat kontrasepsi (tahun)
3. tinggal	tempat tinggal (0=perdesaan,1=perkotaan)
4. agama	agama (0=Islam, 1=non-islam)
5. didik	pendidikan si pemakai (0= ≤ SD, 1=SMP-SMA, 2=Perguruan tinggi)
6. metode	metode alat kontrasepsi (1=pil/suntik, 2=IUD dan implan, 0= metode modern lainnya)

Bagi variabel kategorik yang mempunyai jumlah kategori lebih dari dua, terlebih dahulu dilakukan konversi ke dalam variabel dummy. Variabel pendidikan (didik) dikonversi menjadi dua variabel dummy, yaitu didik1 dan didik2. Begitu juga untuk variabel metode alat kontrasepsi (metode) dikonversi menjadi variabel dummy metode1 dan metode2 seperti ditampilkan dalam Tabel 2.

Tabel 2.
Pembentukan variabel dummy

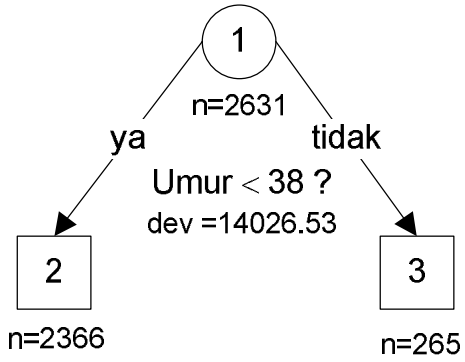
didik	didik1	didik2
0	0	0
1	1	0
2	0	1

metode	metode1	metode2
0	0	0
1	1	0
2	0	1

Untuk variabel respon lamanya bertahan memakai alat kontrasepsi sampai diganti dengan alat kontrasepsi lain diperoleh model subdistribusi terbaik mengandung variabel penjelas soseko, didik1, didik2, metode1, dan metode2 dengan nilai AIC sebesar 14056,20 (lihat Tabel 3). Semua variabel penjelas signifikan pada taraf 5%. Kemudian dibentuk regresi berstruktur pohon, dimana pohon awalnya mempunyai 11 simpul akhir. Setelah dilakukan pemangkasan, diperoleh pohon terbaik yang mempunyai 2 simpul akhir hasil dari pemilahan berdasarkan variabel penjelas umur pada nilai 38 tahun (Gambar 1). Dengan demikian model hasil perkawinannya mempunyai 6 variabel penjelas. Variabel penjelas ke-6 adalah simpul3 yang merupakan variabel dummy untuk pernyataan umur ≥ 38 tahun.

Tabel 3.
Regresi subdistribusi terbaik untuk penyebab kegagalan karena mengganti alat

	kontrasepsi		
Variabel	Koefisien	*p*-value	*p*-value
soseko	-0.04172	0.02022	0.03900
didik1	0.23090	0.09241	0.01200
didik2	0.53350	0.11320	0.00000
metode1	-0.63930	0.20060	0.00140
metode2	-0.75120	0.20710	0.00029
	AIC = 14056.20		



Gambar 1. *Pohon hasil pemangkasan*

Model hasil perkawinannya mempunyai nilai AIC yang lebih kecil daripada model subdistribusi pada Tabel 3. Tambahan pula, variabel simpul3 yang merupakan hasil dari regresi pohon juga signifikan pada taraf yang sangat kecil (Tabel 4). Hal ini juga diperkuat dengan hasil pengujian nisbah kemungkinan antara model dugaan pada Tabel 3 dengan Tabel 4. P-value dari pengujian ini sebesar $9{,}202903\times10^{-6}$.

Tabel 4. Model hasil perkawinan

	kontrasepsi		
Variabel	Koefisien	*p*-value	*p*-value
soseko	-0.03142	0.02023	0.12000
didik1	0.17560	0.09286	0.05900
didik2	0.45420	0.11400	0.00007
metode1	-0.67340	0.20410	0.00097
metode2	-0.76150	0.21030	0.00029
simpul3	-0.51770	0.12390	0.00003
	AIC = 14038.53		

Bagi variabel respon lamanya bertahan menggunakan alat kontrasepsi sampai rusak atau sampai dilepas, dilakukan cara yang sama untuk mendapat model hasil perkawinan antara model regresi subdistribusi dengan model berstruktur pohonnya.

4. Kesimpulan

Metode untuk mengawinkan model regresi subdistribusi dan model berstruktur pohon dibahas dalam makalah ini. Metode ini dimaksudkan untuk meningkatkan kemampuan model. Model berstruktur pohon menutup kekurangan model regresi berstruktur pohon dalam hal memodelkan variasi lokal dalam data, sehingga diperoleh model yang lebih mampu menggambarkan keadaan data.

Penerapan dari metode yang dibangun terhadap data lamanya bertahan memakai alat kontrasepsi sampai mengganti dengan alat kontrasepsi lain menunjukkan hasil yang baik, karena mampu meningkatkan kemampuan model yang ditunjukkan dengan menurunnya nilai AIC.

5. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Dikti Kemendikbud atas hibah yang diberikan kepada penulis. Turut juga disampaikan terima kasih kepada LPPM Universitas Islam Bandung atas terlaksananya acara Seminar Nasional Penelitian dan Pengabdian 2012.

6. Daftar Pustaka

Breiman, L, Friedman, J, Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Chapman and Hall, New York.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94, 496-509.

Ibrahim, N.A., A. Kudus, I. Daud and M.R. Abu-Bakar. (2008). Decision tree for competing risks survival probability in breast cancer study. *World Acad. Sci. Eng. Technol.*, 38: 15-19. Diakses dari http://www.waset.org/journals/waset/v38/v38-4.pdf pada 6 Agustus 2012.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company, Tokyo.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics* 44, 35-47.

Su, X. G. and Tsai, C. L. (2005). Tree-augmented Cox proportional hazards models. *Biostatistics* **6**,486-499.