PENERAPAN METODE REGRESI BERSTRUKTUR POHON PADA PENDUGAAN MASA RAWAT KELAHIRAN BAYI
(Studi Kasus di Rumah Sakit Hasan Sadikin Bandung)

Oleh
ABDUL KUDUS

PROGRAM PASCASARJANA
INSTITUT PERTANIAN BOGOR
1999

"Apabila kamu telah selesai (dari satu urusan), kerjakanlah dengan sungguh-sungguh (urusan) yang lain dan hanya kepada Tuhanmulah hendaknya kamu berharap" (QS 94: 7-8).

"Bila tiba saatnya kiamat, sedang pada tanganmu ada sebatang bibit kurma, lalu ia masih punya peluang untuk menanamkan bibit tersebut sebelum kiamat datang, maka hendaklah dia tanam bibit tersebut. Untuk itu dia mendapat pahala" (Pesan Rasulullah).

What makes a data set interesting is not only its size but also its complexity. (Breiman *et al.* 1993)

Kupersembahkan karya ini buat Rela, *a constant source of inspiration.*
Terima kasih atas ketulusannya.

- A. K -

RINGKASAN

ABDUL KUDUS. Penerapan Metode Regresi Berstruktur Pohon pada Pendugaan Masa Rawat Kelahiran Bayi (Studi Kasus di Rumah Sakit Hasan Sadikin Bandung). (Di bawah bimbingan **AUNUDDIN, AJI HAMIM WIGENA**, dan **BUNAWAN SUNARLIM**).

Metode regresi berstruktur pohon merupakan alat untuk mengeksplorasi data yang berukuran besar dan kompleks. Kekompleksan tersebut bisa berupa dimensinya yang tinggi, jenis datanya yang campuran (misal kontinu dan kategorik) atau struktur datanya yang tidak baku. Perluasan metode ini untuk analisis data tersensor mempertimbangkan sifat-sifat data tersensor dan prosedur metode regresi berstruktur pohon.

Ada dua pendekatan metode regresi berstruktur pohon bagi analisis data tersensor, yaitu pendekatan ukuran pemisahan (*separation measure*) dan pendekatan ukuran kehomogenan (*homogeneity measure*). Penerapan kedua pendekatan tersebut pada data masa rawat kelahiran bayi (data tersensor) memberikan hasil yang mirip. Hasil tersebut mendukung pengelompokan bayi menurut WHO yang sudah dilakukan sejak tahun 1961.

Terdapat lima kelompok bayi berdasarkan masa rawat kelahirannya, yaitu (1) kelompok bayi paling lemah dengan sifat berat lahirnya yang sangat rendah, (2) kelompok bayi lemah dengan sifat berat lahirnya yang rendah, (3) kelompok bayi dengan kekuatan sedang yang bersifat berat lahir cukup tetapi mengalami gangguan fisiologis pernapasan pada masa rawat kelahirannya, (4) kelompok bayi yang kuat yang mempunyai ciri berat lahir cukup, tidak mengalami gangguan fisiologis pernapasan pada masa rawat kelahirannya tetapi pendidikan ibunya maksimal lulus SD dan (5) kelompok bayi paling kuat yang bersifat berat lahir cukup, tidak mengalami gangguan fisiologis pernapasan selama masa rawat kelahirannya dan pendidikan ibunya minimal lulus SMP.

PENERAPAN METODE REGRESI BERSTRUKTUR POHON PADA PENDUGAAN MASA RAWAT KELAHIRAN BAYI
(Studi Kasus di Rumah Sakit Hasan Sadikin Bandung)

Oleh

ABDUL KUDUS

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister Sains
pada
Program Studi Statistika

PROGRAM PASCASARJANA
INSTITUT PERTANIAN BOGOR
1999

Judul Penelitian : Penerapan Metode Regresi Berstruktur Pohon Pada Pendugaan Masa Rawat Kelahiran Bayi (Studi Kasus di Rumah Sakit Hasan Sadikin Bandung)

Nama Mahasiswa : Abdul Kudus

Nomor Pokok : 96149/STK

Program Studi : Statistika

Menyetujui :

1. Komisi Pembimbing

Dr. Ir. Aunuddin
Ketua

Ir. Aji Hamim Wigena, MSc.
Anggota

Ir. Bunawan Sunarlim, MS
Anggota

2. Ketua Program Studi

Dr. Ir. Aunuddin

3. Direktur Program Pascasarjana

Prof. Dr. Ir. Sjafrida Manuwoto

Tanggal lulus : **17 Juni 1999**

RIWAYAT HIDUP

Penulis dilahirkan di Kampung Ciyuda, Desa Bendungan, Kec. Pagaden, Kabupaten Subang dengan nama Kudus pada tanggal yang tidak dicatat oleh orang tua penulis. Dengan alasan untuk keperluan pengisian ijazah SD, maka kepala sekolah SD Negeri Bendungan I menetapkan tanggal lahir penulis pada tanggal 21 Maret 1969 dan penulis meminta penambahan nama depan menjadi Abdul Kudus. Ayah penulis bernama Jamin, ibu bernama Maryam dan mempunyai satu adik bernama Yusanah. Di samping itu, penulis mempunyai satu kakak seibu dan satu kakak seayah. Ibu penulis sudah meninggal sejak tahun 1981 saat penulis bersekolah di kelas 5 SD.

Pendidikan SD ditempuh di SD Negeri Bendungan I dari tahun 1977 sampai tahun 1983. Kemudian, penulis melanjutkan sekolah ke SMP Negeri Binong dan lulus pada tahun 1986. Setelah lulus SMP, penulis melanjutkan sekolah ke SMA Negeri 1 Subang dan lulus pada tahun 1989. Pada jenjang pendidikan dasar dan menengah penulis selalu menempati peringkat pertama.

Pendidikan tinggi ditempuh di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor. Masa pendidikan tersebut dimulai setelah penulis diterima sebagai mahasiswa S1 melalui jalur USMI pada tahun 1989. Penulis lulus S1 pada tanggal 5 Mei tahun 1994 dengan skripsi berjudul "Analisis Usahatani Padi Lahan Irigasi di Kawasan Pantai Utara Jawa Barat dengan Pendekatan Fungsi Produksi (Suatu Perbandingan antara Model CES dan Model Translog)".

Pada tahun 1994 sampai 1995 penulis bekerja sebagai peneliti di CESS (*Center for Economics and Social Studies*) Jakarta. Kemudian, pada bulan Maret 1995 penulis berkarier sebagai dosen di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung sebagai tenaga dosen tetap Yayasan Pendidikan Islam (YPI). Penulis berkesempatan untuk melanjutkan kuliah di Program Studi Statistika, Program Magister, Institut Pertanian Bogor tahun 1996 atas biaya TMPD.

KATA PENGANTAR

Segala puji syukur ke hadirat Allah SWT atas segala kekuatan yang dikaruniakanNya sehingga penulis dapat menyelesaikan tesis yang berjudul "Penerapan Metode Regresi Berstruktur Pohon pada Pendugaan Masa Rawat Kelahiran Bayi (Studi Kasus di Rumah Sakit Hasan Sadikin Bandung)".

Tesis ini dibuat sebagai syarat untuk memperoleh gelar Magister Sains pada Program Studi Statistika, Program Pascasarjana, Institut Pertanian Bogor. Tanpa mengecilkan arti dari hasil penelitian ini penulis menyadari bahwa masih banyak yang harus disempurnakan. Kritik dan saran sangat diharapkan penulis dan semoga tulisan ini berguna bagi pihak-pihak yang memerlukan.

Pada kesempatan ini penulis mengucapkan terima kasih atas bimbingan dan fasilitas yang telah diberikan. Secara khusus penulis mengucapkan terima kasih kepada :

1. Dr. Ir. Aunuddin sebagai ketua pembimbing yang telah memberikan bimbingannya yang sangat berarti bagi penulis.
2. Ir. Aji Hamim Wigena, MSc. sebagai anggota komisi pembimbing yang telah memberi masukan dan koreksiannya selama penulisan tesis ini.
3. Ir. Bunawan Sunarlim, MS sebagai anggota komisi pembimbing atas kesediaannya untuk membimbing penulisan tesis ini.
4. dr. Abdurahman dan semua pihak di Rumah Sakit Hasan Sadikin Bandung yang telah mengijinkan penggunaan data untuk penulisan tesis ini.
5. Arief Shobirin Gusnanto, S.Si atas kesediaannya membantu penyediaan data bagi penulisan tesis ini.
6. Prof. Mark R. Segal, Prof. Terry M. Therneau, Dr. Tjen-Sien Lim *and all member of Recursive-Partitioning mail list group for kindness and quick information.*
7. StatLib *and* S-News Administrator *for providing SPlus library.*
8. Ama, Emih *sareng sadaya kulawargi hatur nuhun kana rojongannana.*
9. Rela *sareng kulawargi kanggo perhatosannana.*
10. Wadya Balad Motekar untuk pengertiannya terutama saat line telepon sedang dipakai mengakses internet, *your great support makes me strong.*

11. Teman, sahabat dan orang-orang yang memang bertujuan untuk memanfaatkan kelemahan penulis (?) di PPS STK 96 terima kasih atas kepercayaan dan kekompakannya.
12. Bu Tuti D. Herrie, Om Rasyid Rahman dan keluarga atas keramahan yang diberikan selama ini.
13. Pak Amir, pak Wawan, pak Is dan seluruh keluarga besar Universitas Islam Bandung terima atas dukungan moral dan finansialnya.
14. dr. Yamin Alsoph, dr. Idral Darwis dan pihak keluarga besar Rumah Sakit Kanker DHARMAIS Jakarta atas ijin penjajagan data bagi penulisan tesis ini.

Bogor, Juni 1999

Penulis

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

PENDAHULUAN

Latar Belakang Masalah

Analisis regresi sebagai suatu alat dapat digunakan untuk berbagai keperluan. Dalam pendugaan respon dari suatu percobaan atau fenomena lain yang dipelajari, analisis regresi digunakan sebagai alat prediksi. Analisis regresi juga digunakan untuk mencari peubah-peubah yang dapat menerangkan keragaman respon dan dapat digunakan dalam kajian lebih lanjut. Di samping itu, analisis regresi digunakan sebagai alat untuk mengevaluasi beberapa model yang telah diturunkan secara teoritis dengan menggunakan data empiris. Dalam suatu kajian ilmiah analisis regresi digunakan untuk mengetahui pengaruh peubah-peubah penjelas terhadap peubah respon. Adanya pengaruh ini dapat diidentifikasi dengan menggunakan spesifikasi model tertentu. Koefisien-koefisien regresi diharapkan dapat memperkuat teori tertentu dalam bidang kajian tersebut, dimana arah dan besaran koefisien regresi memberikan makna yang besar. Dengan demikian masalah pendugaan parameter merupakan tujuan dalam analisis regresi.

Keabsahan penggunaan analisis regresi sangat tergantung pada berbagai asumsi, sehingga sulit untuk mendapatkan dugaan persamaan regresi yang memenuhi semua asumsi. Masalah tersebut dapat diatasi dengan metode regresi yang tidak lagi terikat pada berbagai asumsi. Dua pendekatan yang pernah dilakukan adalah regresi berdasarkan penjumlahan fungsi-fungsi pemulusan dari kombinasi linier peubah-peubah penjelasnya yang dilakukan secara iteratif (Friedman & Stuetzle, 1981) dan regresi dengan metode pohon biner pada penyekatan ruang peubah penjelas untuk melihat adanya perbedaan dugaan respon (Breiman *et al.*, 1993). Walaupun banyak keuntungan dari pendekatan ini, tetapi ada kendala lainnya, antara lain diperlukan gugus data yang besar dan pengolahan data dengan komputer.

Pohon regresi (*regression tree*) adalah salah satu metode yang menggunakan kaidah pohon keputusan (*decision tree*). Pohon keputusan ini dibentuk dengan menggunakan algoritma penyekatan rekursif (*recursive partitioning*). Penggunaan algoritma ini dimulai dengan munculnya program *Automatic Interaction Detection* (AID) dari Morgan & Sonquist pada tahun 1963 (Davis & Anderson, 1989).

Kajian mengenai metode pohon regresi dan pohon klasifikasi (CART, *Classification and Regression Trees*) dipelopori oleh Breiman dan Friedman pada tahun 1973. Perkembangannya ditandai dengan diterbitkannya buku "Classification and Regression Trees" pada tahun 1984. Metode pohon regresi ini diilhami oleh program AID, sedangkan metode pohon klasifikasi diilhami oleh program THAID yang dikembangkan oleh Morgan dan Messenger pada awal tahun 1970-an (Breiman *et al.*, 1993). Dalam metode ini proses pengulangan penghitungannya sangat intensif sehingga diperlukan teknologi komputer untuk mempercepat prosesnya. Walaupun demikian, metode ini mempunyai kelebihan, antara lain dapat mengeksplorasi struktur data yang kompleks, mengidentifikasi peubah-peubah penjelas yang mempunyai hubungan struktural dengan peubah responnya dan memprediksi dugaan respon dari satu atau beberapa amatan baru. Metode ini juga memiliki kemampuan dalam mendeteksi interaksi antar peubah secara lokal atau bekerja untuk menemukan subgrup data yang bermakna. Interpretasi hasilnya lebih mudah daripada persamaan regresi biasa, karena identifikasi pengaruh dari peubah penjelas dalam pohon regresi dilakukan dalam masing-masing subgrup data bukan dalam keseluruhan data seperti halnya regresi biasa. Di samping itu, pohon regresi cenderung *resistant* terhadap pengaruh pencilan (Davis & Anderson,1989).

Kendala-kendala dalam penerapan metode pohon regresi sudah berkurang. Beberapa proses penghitungan yang berulang dan rumit bisa diatasi dengan bantuan komputer, sedangkan masalah perlunya gugus data yang besar dapat diatasi dengan metode validasi silang. Namun demikian diperlukan suatu algoritma yang cepat dan fleksibel untuk berbagai jenis gugus data.

Salah satu terapan analisis regresi adalah dalam bidang kesehatan, antara lain analisis ketahanan hidup (*survival analysis*). Analisis ketahanan hidup merupakan salah satu analisis yang sering digunakan untuk pendugaan masa rawat bagi kelahiran bayi. Pada umumnya data tentang masa rawat bagi kelahiran bayi ini perlu disensor, karena data mengenai lamanya waktu sampai terjadinya kesembuhan seseorang bayi seringkali tidak teramati secara lengkap melainkan tersensor oleh waktu yang ditentukan sebelum kejadian tersebut timbul.

Metode regresi untuk analisis ketahanan hidup melibatkan sejumlah peubah penjelas yang diharapkan dapat menerangkan lamanya waktu respon yang diamati. Pengaruh peubah-peubah penjelas ini terhadap peubah respon dapat diketahui melalui persamaan regresi. Adanya kenyataan tersebut memberikan dorongan untuk menerapkan metode pohon regresi pada analisis ketahanan hidup.

Batasan Masalah

Pemodelan data tersensor dengan menggunakan regresi parametrik ataupun regresi semiparametrik Cox *Proportional Hazards* tidak bisa menjelaskan adanya subgrup data yang bermakna dalam menerangkan perbedaan respon. Di samping itu juga sulitnya memenuhi asumsi-asumsi bagi keabsahan pengujian parameter. Hal tersebut mendorong untuk mengkaji penerapan metode pohon regresi pada pemodelan data tersensor.

Tujuan Penelitian

Berdasarkan permasalahan di atas, penelitian ini bertujuan untuk menerapkan metode pohon regresi dalam menduga masa rawat kelahiran bayi berdasarkan beberapa karakteristik ibunya dan karakteristik bayi tersebut.

TINJAUAN PUSTAKA

Metode Regresi

Analisis regresi digunakan untuk melihat bentuk hubungan antara peubah respon dengan peubah-peubah penjelasnya. Hubungan ini dinyatakan dalam bentuk model stokastik yang linier atau nonlinier. Pendugaan parameter model stokastik dapat dilakukan dengan berbagai cara. Metode pendugaan kuadrat terkecil adalah yang paling populer pada awal perkembangannya. Metode ini memberikan kemudahan dalam penghitungannya, walaupun tidak semua permasalahan regresi bisa diselesaikan dengan metode tersebut.

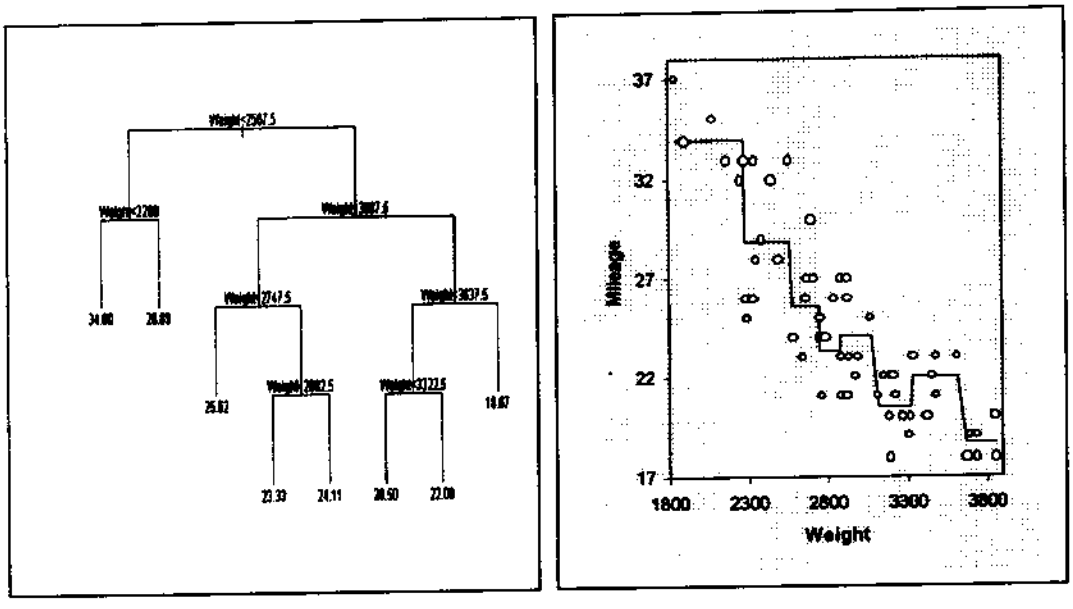
Adanya kendala dalam pemenuhan asumsi-asumsi membawa perkembangan pada penggunaan metode-metode yang lain, seperti metode kuadrat terkecil terboboti, metode kemungkinan maksimum, metode bayes dan lain-lain. Masing-masing metode merupakan jawaban terhadap permasalahan yang berbeda. Misalnya, metode kuadrat terkecil terboboti merupakan jawaban bagi masalah ketidakhomogenan ragam dan metode kemungkinan maksimum merupakan jawaban bagi pemodelan peubah respon dengan sebaran tidak mesti normal.

Semua metode di atas termasuk jenis metode parametrik yang hasilnya berupa dugaan parameter model dan beberapa kriteria lainnya. Dugaan persamaan regresi yang diperoleh bisa merupakan garis di ruang berdimensi dua, bidang di ruang berdimensi tiga atau *hyper plane* di ruang berdimensi lebih dari tiga. Hasil ini ditujukan bagi analisis mengenai bentuk model persamaan, atau mengenai pengaruh dari peubah-peubah penjelas terhadap peubah responnya. Akan tetapi, keyakinan mengenai keabsahan hasil yang diperoleh sangat tergantung pada pemenuhan asumsi.

Metode Pohon Regresi

Jika dalam regresi biasa suatu persamaan regresi menunjukkan bagaimana hubungan antara peubah-peubah penjelas dengan peubah responnya, maka hal yang sama juga berlaku pada pohon regresi. Peubah-peubah penjelas yang berpengaruh dalam regresi biasa juga akan merupakan peubah penjelas yang berpengaruh dalam

pohon regresi. Peubah yang berpengaruh dalam pohon regresi adalah peubah yang menentukan adanya pemilahan (*splitting*) ruang peubah penjelas. Pemilahan tersebut mengakibatkan perbedaan dugaan respon.



Gambar 1. Pohon regresi mengenai hubungan antara berat mobil (*weight*) dengan jarak yang ditempuh oleh suatu mobil tiap satuan volume bahan bakar (*mileage*) (Clark & Pregibon, 1992).

Pada Gambar 1 diilustrasikan penggunaan metode pohon regresi dalam mengkaji hubungan antara berat mobil (*weight*) dengan jarak tempuh per satuan volume bahan bakar (*mileage*). Gambar sebelah kiri adalah pohon regresi yang dibentuk dengan algoritma penyekatan rekursif. Dengan menggunakan beberapa kriteria dalam proses pembentukannya, diperoleh pohon regresi dengan delapan sekatan dan delapan nilai dugaan. Gambar sebelah kanan merupakan representasi lain dari pohon regresi tersebut. Sumbu datar *weight* disekat menjadi delapan sekatan dengan delapan nilai dugaan *mileage*. Garis yang menghubungkan nilai-nilai dugaan tidak mempunyai persamaan garis yang eksplisit, karena penggunaan metode pohon regresi lebih ditujukan untuk masalah-masalah prediksi.

Berikut ini akan dipaparkan mengenai proses pembentukan pohon regresi dan kriteria atau ukuran yang digunakannya. Seperti halnya metode regresi biasa dengan p peubah penjelas $X_1, X_2, ..., X_p$ dan satu peubah respon yang kontinu Y, di sini juga digunakan notasi-notasi yang sama. Pembentukan pohon regresi memerlukan empat komponen, yaitu :

1. Satu gugus pertanyaan dikotomus dengan bentuk "Apakah $\mathbf{x}_i' \in \mathbf{A}$?" dengan $\mathbf{x}_i'$ merupakan satu amatan contoh dan $\mathbf{A} \subset X$ (ruang peubah penjelas). Jawaban dari pertanyaan tersebut menentukan sekatan (*partition*), atau *split*, bagi ruang peubah penjelas. Amatan dengan jawaban *ya* masuk ke anak ruang **A** sedangkan yang *tidak* masuk ke ruang komplemen **A**. Anak ruang contoh yang terbentuk disebut simpul (*node*).
2. Kriteria *goodness-of-split* $\phi(s,t)$ yang merupakan alat evaluasi bagi pemilahan yang dilakukan oleh pemilah (*split*) *s* pada simpul *t*.
3. Ukuran yang digunakan untuk menentukan ukuran pohon yang layak (*right sized tree*).
4. Statistik yang digunakan sebagai ringkasan dari tiap simpul akhir.

Aturan Pemilahan. Pohon regresi dibentuk melalui pemilahan data pada tiap simpul ke dalam dua simpul anak. Aturannya adalah sebagai berikut :

1. Tiap pemilahan tergantung pada nilai yang hanya berasal dari satu peubah penjelas.
2. Untuk peubah kontinu X_j, pemilahan hanya berasal dari pertanyaan "Apakah $X_j \leq c$?" untuk $c \in \Re^1$. Jadi, jika ruang contohnya berukuran n dan terdapat sebanyak-banyaknya n nilai amatan berbeda pada peubah X_j, maka akan terdapat sebanyak-banyaknya n-1 *split* yang berbeda yang dibentuk oleh gugus pertanyaan {"Apakah $X_j \leq c_i$?"}, dengan i =1, 2, ...,n-1 dan c_i adalah nilai tengah-tengah antara dua nilai amatan peubah X_j berurutan yang berbeda.
3. Untuk peubah penjelas kategorik, pemilahan yang terjadi berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua anak gugus yang saling lepas (*disjoint*). Jika peubah X_j merupakan peubah kategorik nominal bertaraf L, maka akan ada 2^{L-1}-1 pemilahan, sedangkan jika berupa peubah kategorik ordinal, maka akan ada L-1 pemilahan yang mungkin.

Aturan *Growing* dan Kriteria *Goodness-of-Split* $\phi(s,t)$. Pohon regresi dibentuk dengan pemilahan yang rekursif berdasarkan kriteria tertentu. Proses pemilahan dilakukan pada tiap simpul dengan cara sebagai berikut :

1. Cari semua kemungkinan pemilahan pada tiap peubah penjelas.

2. Pilih "pemilahan terbaik" dari masing-masing peubah penjelas dan pilih "pemilahan terbaik" dari "kumpulan pemilahan terbaik" tersebut. "Pemilahan terbaik" adalah pemilahan yang memaksimumkan ukuran kehomogenan di dalam masing-masing simpul anak relatif terhadap simpul induknya dan yang memaksimumkan ukuran pemisahan (*separation*) antara dua simpul anak tersebut.

Masalah timbul pada saat terdapat data hilang (*missing data*) pada satu atau beberapa amatan pada peubah penjelas yang menjadi pemilah terbaik. Jika amatan yang datanya lengkap sudah dapat dikelompokkan berdasarkan batas pemilahan yang terjadi, maka amatan dengan data hilang tidak bisa segera dikelompokkan.

Salah satu pendekatan yang digunakan adalah menduga kelompok bagi amatan yang datanya hilang tersebut melalui peubah pengganti (*surrogate*) dimana amatan tersebut mempunyai data lengkap (Breiman *et al.*, 1993 dan Therneau & Atkinson, 1997). Misalkan pemilahan terjadi pada peubah Umur dengan titik pemilahan pada 40 tahun. Peubah pengganti dicari dengan menerapkan ulang algoritma pemilahan pada simpul yang dipilah oleh peubah Umur tadi untuk menduga kelompok 'Umur < 40' dan 'Umur > 40'. Peubah pengganti dicari dengan kriteria 'pemilahan mana yang akan mengelompokkan amatan semirip mungkin dengan pemilahan berdasarkan peubah pemilah ?'. Setiap amatan yang mempunyai data hilang pada peubah pemilah akan dikelompokkan dengan menggunakan peubah pengganti yang pertama, jika pada peubah pengganti yang pertama tersebut juga masih memiliki data yang hilang maka akan dikelompokkan berdasarkan peubah pengganti yang kedua, demikian selanjutnya.

Jumlah Kuadrat Sisaan digunakan sebagai kriteria kehomogenan di dalam masing-masing simpul. Misalkan, simpul t berisi anak contoh $\{(\underline{x}_n, y_n)\}$, $n(t)$ adalah banyaknya amatan dalam simpul t dan rataan respon dalam simpul t adalah :

$$\bar{y}(t) = \frac{1}{n(t)} \sum_{\underline{x}_n \in t} y_n \qquad (1)$$

maka Jumlah Kuadrat Sisaan di dalam simpul t adalah :

$$JKS(t) = \sum_{\underline{x}_n \in t} [y_n - \bar{y}(t)]^2 \qquad (2)$$

Misalkan ada pemilahan s yang menyekat t menjadi simpul anak kiri t_L dan simpul anak kanan t_R. Kriteria Jumlah Kuadrat Sisaan Terkecil adalah

$$\phi(s,t) = \text{JKS}(t) - \{\text{JKS}(t_L) + \text{JKS}(t_R)\} \qquad (3)$$

dan pemilahan terbaik s^* adalah pemilahan yang sedemikian sehingga

$$\phi(s^*, t) = \max_{s \in \Omega} \phi(s, t) \qquad (4)$$

dengan Ω adalah gugus yang berisi semua kemungkinan pemilahan.

Pohon regresi dibentuk melalui pemilahan simpul secara rekursif yang memaksimumkan fungsi ϕ di atas. Pemilahan tersebut dihentikan tatkala banyaknya amatan dalam simpul tersebut berjumlah "tertentu" atau pada saat nilai ϕ lebih kecil dari suatu nilai ambang (*treshold*). Simpul yang terakhir dibentuk disebut sebagai simpul akhir (*terminal node*) atau simpul daun (*leaf node*). Breiman *et al.* (1993) menetapkan banyaknya amatan pada simpul akhir kurang atau sama dengan 5, sedangkan Schmoor *et al.* (1993) mematok banyak amatan kurang dari 25. Dalam AID suatu simpul akan dijadikan sebagai simpul akhir jika $\max_s \phi(s, t) \le 0.006\text{JKS}(t_1)$, dimana t_1 adalah simpul utama (Breiman *et al.*, 1993). Penentuan nilai ambang bagi ϕ analog dengan penentuan α bagi uji-F pada penambahan peubah dalam regresi bertatar langkah maju (*forward stepwise regression*), dimana tak ada lagi penambahan peubah bebas ke dalam model jika peubah tersebut memiliki taraf nyata uji-F yang lebih besar dari α (Therneau & Atkinson, 1997).

Breiman *et al.* (1993) juga memberikan kriteria alternatif, yakni Simpangan Mutlak Terkecil (*Least Absolute Deviation*). Prosesnya sama dengan penggunaan Jumlah Kuadrat Sisaan, kecuali $\bar{y}(t)$ diganti dengan median respon dalam simpul bersangkutan dan fungsinya berupa fungsi mutlak.

Penentuan Ukuran Pohon yang Layak. Pohon yang dibentuk dengan aturan *splitting* dan aturan *growing* di atas berukuran sangat besar. Hal ini karena aturan penghentian (*stopping rule*) yang digunakan hanya berdasarkan banyaknya amatan pada simpul akhir atau besarnya peningkatan tingkat kehomogenan. Lebih banyak *split* yang dilakukan mengakibatkan makin kecilnya tingkat kesalahan prediksi. Hal tersebut terjadi karena simpul akhir bisa hanya berisi satu amatan. Masalahnya

adalah bagaimana menentukan ukuran pohon yang layak. Hal ini analog dengan regresi linier bertatar (*stepwise*), dimana R^2 terus meningkat seiring dengan banyaknya peubah penjelas yang masuk ke dalam model dan perlu ditentukan banyaknya peubah penjelas pada model terbaik. Pohon yang besar bisa menimbulkan dugaan adanya *overfitting*. Sebaliknya, kasus *underfitting* terjadi karena tidak adanya pemilahan lebih lanjut akibat adanya tetapan ambang ϕ (s^*,t), padahal sebenarnya pemilahan yang terjadi adalah layak. Cara mengatasi masalah ini adalah mencari pohon dengan ukuran yang layak.

Pencarian pohon dengan ukuran yang layak dilakukan dengan (1) penentuan pohon awal yang besar, (2) secara iteratif pohon tersebut dipangkas (*pruning*) menjadi sekuen pohon yang makin kecil dan tersarang dan (3) dipilih pohon terbaik dari sekuen ini dengan menggunakan contoh uji (*test sample*) atau contoh validasi silang (*cross validation sample*).

Pemangkasan pada langkah (2) dilakukan dengan menggunakan ukuran *cost-complexity* minimum (Breiman *et al.*, 1993). Untuk sembarang pohon T yang merupakan subpohon dari pohon terbesar T_{max} diperoleh ukuran *complexity*-nya $|\tilde{T}|$. Ukuran *complexity* tersebut adalah banyaknya simpul akhir. Dalam regresi biasa, ukuran *complexity* tersebut analog dengan derajat bebas model (Therneau & Atkinson, 1997). Untuk suatu $\alpha \geq 0$, $\alpha \in \Re^1$ ukuran *cost-complexity*-nya adalah

$$R_\alpha(t) = R(T) + \alpha|\tilde{T}| \qquad (5)$$

Dengan α adalah parameter complexity mengenai *cost* bagi penambahan satu simpul akhir pada pohon T dan R(T) adalah *resubstitution estimate* yang digunakan, seperti Jumlah Kuadrat Sisaan atau Jumlah Simpangan Mutlak.

Langkah awal pemangkasan dilakukan terhadap T_1, yakni suatu subpohon yang memenuhi kriteria $R(T_1) = R(T_{max})$. Untuk mendapatkan T_1 dari T_{max}, ambil t_L dan t_R yang merupakan simpul anak kiri dan simpul anak kanan dari T_{max} yang dihasilkan dari simpul induk t. Oleh karena $R(t) \geq R(t_L) + R(t_R)$, begitu diperoleh dua simpul anak dan simpul induknya yang memenuhi persamaan $R(t) = R(t_L) + R(t_R)$, maka pangkaslah simpul anak t_L dan t_R tersebut. Ulangi lagi proses ini

sampai tak ada lagi pemangkasan yang mungkin. Hasilnya adalah pohon T_1 yang memenuhi kriteria di atas.

Inti dari pemangkasan *cost-complexity* minimum adalah pemotongan *weakest-link*. Untuk sembarang T_t yang merupakan cabang (*branch*) dari T_1, didefinisikan

$$R(T_t) = \sum_{t' \in \tilde{T}_t} R(t') \tag{6}$$

dimana $\tilde{T}_t$ adalah gugus simpul akhir dari T_t. Untuk sembarang simpul dalam (*internal node / nonterminal node*) t dari pohon T_1 berlaku sifat $R(t) > R(T_t)$. Definisikan pula $\{t\}$ yaitu subcabang dari T_t yang hanya terdiri dari satu simpul. Ukuran *cost-complexity* dari subcabang $\{t\}$ adalah

$$R_\alpha(\{t\}) = R(t) + \alpha \tag{7}$$

dan ukuran *cost-complexity* dari cabang T_t adalah

$$R_\alpha(T_t) = R(T_t) + \alpha \left|\tilde{T}_t\right| \tag{8}$$

Oleh karena T_t merupakan cabang dan $\{t\}$ merupakan subcabang yang terdiri dari satu simpul dalam, maka

$$R_\alpha(T_t) < R_\alpha(\{t\}) \tag{9}$$

berati cabang T_t mempunyai *cost-complexity* yang lebih kecil daripada subcabang $\{t\}$. Tetapi pada nilai kritis α, kedua *cost-complexity* tersebut sama besar. Nilai kritis α tersebut adalah

$$\alpha = \frac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1} \tag{10}$$

Untuk setiap $t \in T_1$, didefinisikan fungsi $g_1(t)$ sebagai berikut :

$$g_1(t) = \begin{cases} \dfrac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1} & , t \notin \tilde{T}_1 \\ +\infty & , t \in \tilde{T}_1 \end{cases} \tag{11}$$

$\bar{t}_1$ adalah *weakest-link* dalam T_1, jika simpul tersebut memenuhi kriteria

$$g_1(\bar{t}_1) = \min_{t \in T_1} g_1(t) \tag{12}$$

dan

$$\alpha_2 = g_1(\bar{t}_1) \tag{13}$$

Dengan demikian $\bar{t}_1$ adalah simpul dalam pertama yang membuat $R_\alpha(\{t\}) = R_\alpha(T_t)$ atau persamaan (10) terpenuhi dan α_2 adalah nilai dari parameter *complexity* dimana kesamaan tersebut terjadi. Selanjutnya ganti $\{\bar{t}_1\}$ tersebut dengan cabang $T_{\bar{t}_1}$, yakni suatu cabang yang memiliki simpul utama $\bar{t}_1$, dan pangkaslah cabang $T_{\bar{t}_1}$ tersebut dari pohon T_1. Hasil pemangkasannya adalah T_2, yakni

$$T_2 = T_1 - T_{\bar{t}_1} \tag{14}$$

Dengan demikian pohon T_2 adalah subpohon yang memenuhi kriteria *cost-complexity* minimum dengan nilai parameter *complexity* sebesar α_2.

Selanjutnya lakukan pemangkasan pada subpohon T_2 dengan proses seperti di atas. Kalau ditemukan adanya *weakest-link* kembar pada langkah ke-k, k=1,2, ..., misal $g_k(\bar{t}_k) = g_k(\bar{t}'_k)$, maka pemangkasan dilakukan sebagai berikut :

$$T_{k+1} = T_k - T_{\bar{t}_k} - T_{\bar{t}'_k} \tag{15}$$

Hasil dari proses diatas adalah berupa sekuen subpohon yang tersarang dan makin kecil, yakni $\{T_1, T_2, ..., \{t_1\}\}$ dengan $T_1 > T_2 > ... > \{t_1\}$ (*baca* : pohon T_1 adalah induk bagi pohon T_2, pohon T_2 adalah induk bagi pohon T_3 , demikian selanjutnya) dan sekuen α dalam urutan meningkat, yakni $\{\alpha_1, \alpha_2, \alpha_3, ...\}$ dengan $\alpha_1 = 0$, $\alpha_2 > \alpha_1$, dan demikian selanjutnya.

Permasalahan selanjutnya adalah bagaimana memilih "pohon terbaik" dari sekuen pohon hasil pemangkasan di atas ? Di sini diperkenalkan istilah "*honest estimate*" bagi R(T). Jika kita menggunakan *resubstitution estimate* R(T) sebagai kriteria penentuan pohon terbaik, kita akan cenderung memilih pohon yang terbesar, yakni T_1. Sebab makin besar pohon, makin kecil R(T). Ada dua *honest estimate* bagi R(T), yaitu *test sample estimate* $R^{ts}(T)$ dan *cross-validation estimate* $R^{CV}(T)$.

Untuk mendapatkan *test sample estimate*, amatan dibagi dua secara acak menjadi *learning sample* L_1 dan *test sample* L_2. L_1 digunakan untuk membentuk sekuen pohon $\{T_k\}$ melalui proses pemangkasan, sedangkan L_2 digunakan untuk membentuk $R^{ts}(T_k)$. Jika L_2 berukuran n_2, maka

$$R^{ts}(T_k) = \frac{1}{n_2} \sum_{(x_n, y_n) \in L_2} [y_n - \hat{y}_k(x_n)]^2 \tag{16}$$

dimana $\hat{y}_k(x_n)$ adalah dugaan respon dari amatan ke-n pada pohon ke-k. Pohon terbaik adalah T_{k0}, yang memenuhi kriteria :

$$R^{ts}(T_{k0}) = \min_k R^{ts}(T_k)$$

Untuk membentuk *cross-validation estimate* dengan *V-fold*, amatan induk L yang berukuran n dibagi secara acak menjadi V kelompok, yakni L_1, L_2, ...,L_V yang berukuran sama. *Learning sample* ke-v adalah $L^{-v} = L - L_v$, v = 1, 2, ..., V yang digunakan untuk membentuk sekuen pohon $\{T_k\}$ dan sekuen parameter *complexity* $\{\alpha_k\}$. Jadi terdapat v sekuen $\{T_k\}$ dan v sekuen $\{\alpha_k\}$. Kemudian gunakan amatan induk L untuk membentuk sekuen $\{T_k\}$ dan $\{\alpha_k\}$. Definisikan $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$. Jika $\hat{y}_k^{-v}(x_n)$ adalah dugaan respon dari amatan ke-n pada pohon yang bersesuaian dengan α'_k yang dibentuk oleh *learning sample* ke-v, maka

$$R^{CV}(T_k) = \frac{1}{n}\sum_{v=1}^{V}\sum_{(x_n,y_n)\in L_v}\left[y_n - \hat{y}_k^{-v}(x_n)\right]^2 \qquad (17)$$

Pohon yang terbaik adalah T_{k0}, yaitu :

$$R^{CV}(T_{k0}) = \min_k R^{CV}(T_k) \qquad (18)$$

Cross validation estimate dengan 10-*fold* menghasilkan *resubstitution estimate* yang paling kecil (Breiman *et al.*, 1993). Dalam keperluan eksplorasi, penentuan pohon yang layak bisa dilakukan secara subyektif melalui pemangkasan T_{max} secara manual (Intrator & Kooperberg, 1995).

Statistik pada Simpul Akhir. Sebagaimana telah digunakan dalam persamaan (1), pada pohon regresi digunakan statistik rataan respon sebagai dugaan respon pada tiap simpul akhir.

Metode Analisis Data Ketahanan Hidup

Dalam kepustakaan, istilah waktu ketahanan hidup (*survival time*), waktu kegagalan (*failure time*), waktu sampai suatu kejadian (*time-to-event*), atau lamanya waktu (*duration*) digunakan secara bergantian. Data ketahanan hidup adalah istilah yang digunakan bagi data tentang lamanya waktu sampai terjadinya suatu kejadian. Kejadian yang paling sederhana adalah kematian, yang lainnya

antara lain terjadinya suatu penyakit, kambuhnya suatu penyakit. Dalam bidang perindustrian antara lain waktu kegagalan suatu komponen. Dalam bidang ekonomi antara lain waktu sampai mendapatkan pekerjaan. Dalam bidang demografi misalnya kejadian pernikahan (Hougaard, 1999). Dalam bidang peternakan misalnya rentang waktu masa produktif, ketahanan dalam masa pertama laktasi, *prenatal* dan *postnatal* merupakan karakteristik yang patut dipertimbangkan dalam pemuliaan (Smith, 1990 *dalam* Saefuddin, 1996). Analisis ketahanan hidup berhubungan dengan model atau metode statistika untuk menganalisis data jenis tersebut. Dengan demikian tampak jelas bahwa analisis ketahanan hidup sangat penting, karena terapannya berada pada berbagai bidang.

Yang membedakan analisis ketahanan hidup dari bidang kajian statistika lainnya adalah sebaran dari data ketahanan hidup biasanya tidak simetris. Histogram data ketahanan hidup biasanya cenderung menceng ke kanan (*positively skewed*). Sangat naif kalau data tersebut diasumsikan menyebar normal. Bisa saja dilakukan transformasi untuk mendapatkan sebaran yang lebih simetrik, tetapi pendekatan yang lebih baik adalah dengan mengadopsi model sebaran lain (Collett, 1994).

Perbedaan lainnya adalah adanya data tersensor. Keadaan ini terjadi saat peneliti tidak dapat mengamati obyek penelitian sampai timbulnya kejadian, sehingga obyek tersebut tersensor di kanan. Keadaan ini berarti informasi tentang lamanya ketahanan hidup yang diperoleh hanya sebagian, karena obyek tersebut mempunyai ketahanan hidup melebihi waktu amatan yang ditentukan atau memang tidak bisa diamati.

Dalam analisis ketahanan hidup dikenal fungsi ketahanan (*survivor function*), yakni

$$S(t) = P[T > t] , t \geq 0 \qquad (19)$$

dimana T adalah peubah acak mengenai waktu hidup dan t adalah nilai amatan waktu hidup. Fungsi ketahanan menyatakan peluang suatu amatan bertahan hidup dari waktu pangkal (*time origin*) sampai suatu waktu lebih dari t. Di sini juga dikenal fungsi *hazard* yakni peluang terjadinya suatu kejadian dalam selang waktu yang pendek dimana diketahui bahwa obyek tersebut masih hidup sampai saat

tersebut. Dengan demikian fungsi *hazard* ini menyatakan laju kematian bagi amatan yang diketahui telah bertahan hidup sampai t, yakni

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \le T \le t + \delta t \mid T \ge t}{\delta t} \right\} \qquad (20)$$

Jika F(t) adalah fungsi sebaran kumulatif bagi peubah acak T, maka

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)} \qquad (21)$$

dan jika f(t) adalah fungsi kepekatan peluang bagi peubah acak T, maka

$$h(t) = \frac{f(t)}{S(t)} \qquad (22)$$

yang berarti

$$h(t) = -\frac{d}{dt}\{\log S(t)\} \qquad (23)$$

dan juga

$$S(t) = \exp\{-H(t)\} \qquad (24)$$

dimana

$$H(t) = \int_0^t h(u)\ du \qquad (25)$$

H(t) adalah fungsi *hazard* kumulatif yang terhubung dengan fungsi ketahanan melalui kesamaan berikut :

$$H(t) = -\log S(t) \qquad (26)$$

Fungsi *hazard* dan fungsi ketahanan diduga dari waktu ketahanan hidup yang teramati. Dalam praktiknya ada dua pendekatan untuk menduga kedua fungsi tersebut, yaitu metode pendugaan yang tidak memerlukan spesifikasi fungsi kepekatan peluang bagi peubah acak T (metode nonparametrik) dan yang mengasumsikan sebaran tertentu bagi peubah acak T (metode parametrik).

Pendugaan Fungsi Ketahanan dengan Metode Kaplan-Meier

Analisis data ketahanan hidup bisa dimulai dengan pemaparan ringkasan numerik atau grafis dari data sekelompok individu yang dianalisis. Langkah ini sebagai pendahuluan bagi analisis yang lebih lanjut. Data ketahanan hidup lebih

mudah diringkas melalui dugaan dari fungsi ketahanan atau fungsi *hazard*-nya. Metode pendugaan Kaplan-Meier merupakan metode pendugaan fungsi ketahanan atau fungsi *hazard* yang biasa digunakan. Metode ini merupakan metode nonparametrik karena metode ini tidak berdasarkan pada asumsi mengenai sebaran dari data ketahanan hidup yang dianalisis.

Untuk menentukan dugaan Kaplan-Meier bagi fungsi ketahanan dilakukan pemasangan nilai amatan waktu ketahanan dengan status mengenai amatan tersebut yakni (t_i, δ_i), dimana $\delta_i = 0$ jika amatan tersebut tersensor atau $\delta_i = 1$ jika amatan tersebut mati (merupakan amatan lengkap). Jika $t_1^* < t_2^* < \ldots < t_m^*$ adalah notasi bagi waktu kematian yang berbeda dan jika $Y_i(s)$ adalah fungsi indikator yang bernilai 1 jika amatan ke-i beresiko mati pada waktu s dan bernilai 0 jika sebaliknya, yakni $Y_i(s) = 1$ jika $s \le t_i^*$ dan banyaknya yang beresiko mati pada waktu s adalah $r(s) = \sum_{i}^{n} Y_i(s)$ serta d(s) didefinisikan sebagai banyaknya yang mati sampai dengan waktu s, maka dugaan Kaplan-Meier bagi fungsi ketahanannya adalah :

$$\hat{S}_{KM}(t) = \prod_{t_i < t} \frac{r(t_i) - d(t_i)}{r(t_i)} \qquad (27)$$

Sebagai teladan misalkan diketahui data ketahanan hidup sebagai berikut :
10, 13+, 18+, 19, 23+, 30, 36, 38+, 54+, 56+, 59, 75, 93, 97, 104+, 107, 107+, 107+ [1], maka dugaan Kaplan-Meier fungsi ketahanannya tercantum pada Tabel 1.

Tabel 1. Dugaan Kaplan-Meier bagi fungsi ketahanan hidup

Selang	$r(t_i)$	$d(t_i)$	$[r(t_i) - d(t_i)]/ r(t_i)$	$\hat{S}(t)$
0-	18	0	1.0000	1.0000
10-	18	1	0.9444	0.9444
19-	15	1	0.9333	0.8815
30-	13	1	0.9231	0.8137
36-	12	1	0.9167	0.7459
59-	8	1	0.8750	0.6526
75-	7	1	0.8571	0.5594
93-	6	1	0.8333	0.4662
97-	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

[1] Tanda + di belakang nilai amatan menandakan bahwa amatan tersebut adalah tersensor

Jika dugaan fungsi ketahanan sudah diperoleh, maka median ataupun persentil lainnya dari sebaran waktu ketahanan dapat diduga secara grafis.

Uji Log-rank untuk Pembandingan Dua Kelompok Data Ketahanan Hidup

Dua kelompok data ketahanan hidup dapat dibandingkan untuk mengetahui perbedaan ketahanan hidup atau *survival experience*-nya. Seperti halnya data yang menyebar normal dimana uji kesamaan dua nilai tengah kelompok diuji dengan statistik uji t atau z, maka untuk data ketahanan hidup statistik ujinya adalah statistik log-rank.

Statistik log-rank diturunkan dari tabel 2×2 untuk setiap waktu kematian. Misalkan untuk waktu kematian t_i^* dibuat tabel 2×2 sebagai berikut :

	Mati	Hidup	
Kelompok 1	a_i	b_i	n_{i1}
Kelompok 2	c_i	d_i	n_{i2}
	m_{i1}	m_{i2}	n_i

Dimana a_i adalah banyaknya yang mati dari kelompok 1, m_{i1} adalah banyaknya yang mati dari kedua kelompok, n_{i1} adalah banyaknya amatan pada kelompok 1 dan n_i adalah banyaknya amatan dari kedua kelompok pada selang waktu t_i^* sampai t_{i+1}^*. Jika terdapat k waktu kematian pada dua kelompok amatan tersebut, maka akan ada k buah tabel seperti di atas, sehingga statistik log-rank adalah :

$$W_L = \frac{\sum_{i=1}^{k}[a_i - E_0(A_i)]}{\left[\sum_{i=1}^{k} var_0(A_i)\right]^{1/2}} \qquad (28)$$

Dengan A_i adalah peubah acak mengenai banyaknya yang mati pada kelompok 1 pada selang ke-i. Nilai harapan dan ragam A_i di bawah asumsi bahwa ketahanan hidup kedua kelompok sama adalah :

$$E_0(A_i) = \frac{m_{i1} n_{i1}}{n_i} \qquad (29)$$

dan

$$var_0(A_i) = \left[\frac{m_{i1}(n_i - m_{i1})}{n_i - 1}\right]\left[\left(\frac{n_{i1}}{n_i}\right)\left(1 - \frac{n_{i1}}{n_i}\right)\right] \qquad (30)$$

Statistik W_L menyebar Khi-Kuadrat dengan derajat bebas 1.

Model *Hazards* Proporsional

Untuk mengetahui pengaruh dari beberapa peubah penjelas terhadap ketahanan hidup dapat dilakukan melalui analisis regresi. Model regresi yang biasa digunakan dalam analisis ketahanan hidup adalah model Cox. Dalam model Cox, fungsi *hazard* bagi amatan dengan kovariat $\underline{x} = (x_1, x_2, \ldots, x_p)$ adalah

$$h(t,\underline{x}) = h_0(t)\exp(\underline{\beta}'\underline{x}) \qquad (31)$$

Dimana $\underline{\beta}$ adalah vektor koefisien berdimensi-p, dan $h_0(t)$ adalah fungsi *hazard* dasar (*baseline hazard function*). Dengan demikian nilai $\exp(\underline{\beta}'\underline{x})$ adalah *hazard* pada saat t bagi amatan dengan peubah penjelas $\underline{x}$ relatif terhadap *hazard* amatan dengan peubah penjelas bernilai nol. Persamaan (31) dapat juga dituliskan sebagai

$$\log\left\{\frac{h(t,x)}{h_0(t,x)}\right\} = \underline{\beta}'\underline{x}$$

Tampak bahwa logaritma nisbah *hazard* merupakan model linier.

Nilai-nilai dugaan $\underline{\beta}$ diperoleh dengan memaksimumkan fungsi kemungkinan parsial, yakni

$$L(\underline{\beta}) = \prod_{i=1}^{d}\left(\frac{\exp(\underline{\beta}'\underline{x}_i)}{\sum_{j\in R(t_i)}\exp(\underline{\beta}'\underline{x}_j)}\right) \qquad (32)$$

dimana $\underline{x}_i$ adalah vektor kovariat dari amatan dengan waktu ketahanan hidup t_i, $R(t_i)$ adalah gugus amatan yang beresiko mati pada saat t_i dan d adalah banyaknya selang waktu kematian (Collet, 1994).

Jika data ketahanan hidup yang dianalisis berukuran n, yakni $t_1, t_2, \ldots, t_n$ dan δ_i adalah indikator sensor dimana ia bernilai nol jika tersensor dan satu jika merupakan data lengkap, maka persamaan (32) juga bisa dinyatakan sebagai :

$$L(\underline{\beta}) = \prod_{i=1}^{n}\left[\frac{\exp(\underline{\beta}'\underline{x}_i)}{\sum_{j\in R(t_i)}\exp(\underline{\beta}'\underline{x}_j)}\right]^{\delta_i}$$

Fungsi logaritma-kemungkinan parsialnya adalah :

$$\log L(\underline{\beta}) = \sum_{i=1}^{n}\delta_i\left\{\underline{\beta}'\underline{x}_i - \log\sum_{j\in R(t_i)}\exp(\underline{\beta}'\underline{x}_j)\right\} \qquad (33)$$

Pemaksimuman fungsi (32) ini dapat dilakukan dengan prosedur Newton-Raphson.

Setelah model di-*fit* terhadap data yang ada, maka selanjutnya perlu diketahui seberapa baik model tersebut. Prosedur diagnostik bagi pemeriksaan model merupakan bagian penting dalam proses pemodelan. Kecuali pada kasus yang hanya melibatkan satu atau dua peubah penjelas, pengujian secara visual mengenai identifikasi bagi karakteristik tertentu seperti adanya amatan yang memiliki ketahanan hidup yang terlalu lama atau terlalu sebentar tidak lagi bisa dilakukan. Permasalahan menjadi makin rumit jika ada data ketahanan hidup yang tersensor.

Beberapa prosedur pemeriksaan model didasarkan pada sisaan. Model dengan sisaan yang kecil diharapkan merupakan model yang baik. Terdapat banyak jenis sisaan pada analisis data ketahanan hidup, antara lain :

1. Sisaan Cox-Snell.

Sisaan Cox-Snell bagi amatan ke-i, i = 1,2,...,n adalah

$$r_{Ci} = \exp\left(\hat{\underline{\beta}}'\underline{x}_i\right)\hat{H}_0(t_i) \qquad (34)$$

Dengan $\hat{H}_0(t_i)$ adalah dugaan bagi fungsi kumulatif *hazard* dasar pada waktu kematian t_i. Beberapa sifat dari sisaan Cox-Snell antara lain : (1) Sisaan ini tidak menyebar secara setangkup di sekitar nol, karena nilainya yang tidak negatif dan (2) Sisaan ini diasumsikan menyebar secara eksponensial dengan nilai tengah dan ragamnya bernilai satu.

2. Sisaan Cox-Snell yang dimodifikasi.

Adanya amatan yang tersensor membuat sisaan bagi amatan ini tidaklah seharusnya sama dengan amatan yang lengkap, sehingga diperkenalkan sisaan Cox-Snell yang dimodifikasi, yakni :

$$r'_{Ci} = 1 - \delta_i + r_{Ci} \qquad (35)$$

dengan δ_i adalah indikator sensor dimana ia bernilai nol jika tersensor dan satu jika merupakan data lengkap.

3. Sisaan Martingale.

Sisaan Cox-Snell yang dimodifikasi mempunyai nilai tengah satu bagi amatan yang lengkap, sehingga perlu perbaikan agar ia mempunyai nilai tengah nol bagi amatan yang lengkap. Sisaan martingale didefinisikan sebagai

$$r_{Mi} = \delta_i - r_{Ci} \qquad (36)$$

Sisaan ini bernilai antara $-\infty$ sampai 1 dan bernilai negatif bagi amatan yang tersensor, $\delta_i = 0$.

4. Sisaan Devians.

Sisaan devians diperkenalkan oleh Therneau *et. al* (1990) sebagai jawaban bagi sisaan martingale yang tidak menyebar setangkup di sekitar nol. Sisaan ini didefinisikan sebagai :

$$r_{Di} = \text{sgn}(r_{Mi})[-2\{ r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{1/2} \qquad (37)$$

Satu sifat penting dari sisaan ini adalah walaupun sisaan ini menyebar setangkup di sekitar nol, tetapi tidak mesti berjumlah nol.

Metode Pohon Regresi Ketahanan Hidup

Beberapa penyesuaian harus dilakukan bagi penerapan metode pohon pada analisis ketahanan hidup. Penyesuaian tersebut adalah penyesuaian pada aturan pemilahan, algoritma pemangkasan, pemilihan pohon yang layak dan aturan penentuan statistik pada simpul akhir. Statistik yang digunakan pada simpul akhir sebagai dugaan bagi respon didasarkan pada penduga dari fungsi sebaran. Perluasan metode pohon bagi analisis ketahanan hidup menggunakan dua pendekatan. Pendekatan pertama berdasarkan ukuran pemisahan (*separation measures*). Pada pendekatan ini, statistik uji mengenai perbedaan dua kelompok menjadi perhatian yang utama. Sedangkan pendekatan kedua menggunakan statistik yang menyatakan kehomogenan di dalam simpul. Pendekatan ini berusaha untuk menemukan kelompok amatan dalam tiap simpul dengan ketahanan hidup yang serupa.

Pohon Regresi Ketahanan Hidup dengan Ukuran Pemisahan

Segal (1988) menggunakan uji beda dua kelompok data tersensor sebagai dasar bagi aturan pemilahan. Walaupun secara aljabar tidak dapat dibuktikan bahwa pemilahan berdasarkan ukuran pemisahan menghasilkan hasil yang sama dengan pendekatan ukuran kehomogenan, tetapi Segal (1988) menemukan kasus kedua pendekatan tersebut memberikan hasil yang sama. Pemilahan yang terbaik

adalah yang mempunyai nilai statistik *log-rank* yang paling besar. Hal ini berarti dua simpul yang terbentuk memiliki ketahanan hidup yang berbeda.

Algoritma pemangkasan dari metode pohon regresi biasa tidak dapat diterapkan pada pendekatan ini. Segal (1988), memberikan alternatif pemilahan berdasarkan kriteria

$$\phi(s,t) = \sum_{\underline{x}_n \in t} |y_n - v(t)| - \left(\sum_{\underline{x}_n \in t_L} |y_n - v(t_L)| + \sum_{\underline{x}_n \in t_R} |y_n - v(t_R)| \right) \qquad (38)$$

dimana $v(.)$ adalah median Kaplan-Meier dari simpul yang bersangkutan.

Penentuan pohon terbaik, jika menggunakan kriteria pemilahan (38) adalah sama dengan metode pohon regresi biasa. Tetapi, jika kriteria pemilahannya menggunakan statistik *log-rank*, maka penentuan pohon terbaiknya adalah sebagai berikut :

(i) Bentuk pohon yang paling besar (T_{max}).

(ii) Pasangkan pada setiap simpul dalam (*internal node / nonterminal node*) dengan statistik log-rank terbesar dari semua simpul yang berada dalam cabang yang berpangkal pada simpul tersebut. (Hal ini dapat dilakukan dengan memeriksa semua statistik log-rank dari simpul-simpul yang berada di dalam cabang yang bersangkutan).

(iii) Susun sekuen pemangkasan dimulai dengan pasangan nilai statistik log-rank yang paling kecil.

(iv) Plot statistik log-rank terhadap ukuran pohon dan pohon terbaik adalah tatkala plot tersebut menampakkan garis hampir datar, dimana dua nilai statistik log-rank berurutan bernilai hampir sama.

Pohon Regresi Ketahanan Hidup dengan Ukuran Kehomogenan

Pemilahan tiap simpul pada pendekatan ukuran kehomogenan didasarkan pada ukuran yang menyatakan tingkat kehomogenan dalam kelompok amatan. Davis dan Anderson (1989) menggunakan ukuran negatif logaritma fungsi kemungkinan (*negatif log-likelihood*) dari model eksponensial. Pemilahan yang dipilih adalah yang mempunyai kerugian (*loss*) paling kecil, yakni bagi simpul t, ukuran kehomogenannya adalah

$$R(t) = -\hat{L}(t) = D_t - D_t \log\left(\frac{D_t}{Y_t}\right) \qquad (39)$$

dimana D_t adalah banyaknya yang mati pada simpul t dan Y_t adalah total waktu pengamatan pada simpul t.

LeBlanc dan Crowley (1992) mendasarkan pemodelannya pada model *semiparametric proportional hazard*. Ukuran kehomogenan yang digunakannya adalah sisaan devians. Sisaan devians-nya dihitung dengan cara sebagai berikut : Pada suatu simpul r dilakukan pemodelan Cox *proportional hazard* dengan model $h_r(t) = \theta_r h_0(t)$, $\theta_r \geq 0$. Nilai dugaan θ_r adalah :

$$\hat{\theta}_r = \frac{\sum_{i \in S_r} \delta_i}{\sum_{i \in S_r} \hat{H}_0(t)} \quad \text{dengan} \quad \hat{H}_0(t) = \sum_{i:t_i<t} \frac{\delta_i}{\sum_{r \in T} \sum_{i:t_i>t, i \in S_r} 1}$$

dan sisaan devians-nya adalah :

$$r_{Di} = 2\left[\delta_i \log\left(\frac{\delta_i}{\hat{H}_0(t)\hat{\theta}_r}\right) - \left(\delta_i - \hat{H}_0(t)\hat{\theta}_r\right)\right] \qquad (40)$$

Langkah-langkah *growing*, pemangkasan dan penentuan ukuran pohon yang layak pada pendekatan ini sepenuhnya mengadopsi kriteria yang digunakan dalam metode pohon untuk regresi biasa. Sedangkan statistik penduga respon pada simpul akhir menggunakan median waktu ketahanan.

Beberapa Teladan Terapan Pohon Regresi

Segal (1988) menerapkan metode pohon regresi pada masalah analisis survival. Peubah respon yang digunakan adalah $\log_{10}$ *survival time* yakni waktu ketahanan hidup (dalam hari) pasien pencangkokan jantung, sedangkan dua peubah bebas yang digunakan adalah umur pasien penerima pencangkokan jantung dan skor ketidakcocokan antara penerima dan donor jantung.

Davis & Anderson (1989) melakukan simulasi bagi metode pohon regresi untuk analisis survival dengan model eksponensial. Hasil simulasi menunjukkan bahwa metode pohon regresi berhasil mengidentifikasi struktur *hazard* dalam data survival.

LeBlanc & Crowley (1992) menganalisis ketahanan hidup pasien myeloma, yakni penderita kanker sel plasma sumsum tulang. Lima peubah penjelas yang sebelumnya sudah diketahui mempunyai hubungan dengan ketahanan hidup yaitu usia, serum kalsium, serum albumin, serum kreatinin dan serum β_2 mikroglobulin. Hasilnya terbentuk 8 kelompok pasien masing-masing dengan median ketahanan hidup 35.6 bulan, 22.6 bulan, 61 bulan, 26.9 bulan, 26 bulan, 9.6 bulan, 13.9 bulan dan 5 bulan.

Segal (1992) mengembangkan metode pohon regresi untuk respon hasil pengukuran berulang (*repeated measurement*) dan respon *longitudinal*. Ia mengganti fungsi yang digunakan dalam *splitting* dengan fungsi yang sudah mengakomodasi adanya respon *multiple*. Statistik yang digunakannya adalah T^2 Hotelling. Teladan data yang digunakan adalah mengenai HIV di San Francisco, yang meliputi peubah bebas umur, pendidikan, ras, *number of past episodes of syphilis*, status pengidap gonorrhoea, *genital herpes* dan hepatitis B, banyaknya *male sex partner* dalam tahun sebelumnya, sejarah transfusi darah, dan konsumsi alkohol dan rokok. Ada lima peubah respon yang digunakannya adalah yaitu *annual β_2 microglobulin determination* selama 5 tahun. Dengan contoh berukuran 95 responden hasilnya menunjukkan bahwa respon *annual β_2 microglobulin determination* selama 5 tahun ditentukan oleh peubah banyaknya *male sex partner* dalam tahun sebelumnya sebagai peubah pemilah pertama pada titik pemilahan 28 *partner* per tahun. Simpul akhir dengan banyaknya *partner* kurang dari atau sama dengan 28 (yang diistilahkan dengan kelompok *less sexually active subjects*) mempunyai tingkat *β_2 microglobulin* di bawah rataan dengan responden berjumlah 40 orang. Sedangkan kelompok komplemennya yang mempunyai *partner* lebih dari 28 orang per tahun dipilah lagi berdasarkan peubah *number of past episodes of syphilis* pada titik pemilahan "tidak pernah" atau "pernah", dimana kelompok "tidak pernah" mempunyai tingkat *β_2 microglobulin* yang hampir sama dengan kelompok *less sexually active subjects*, sedangkan kelompok "pernah" mempunyai tingkat *β_2 microglobulin* di atas rataan.

Ahn (1996) menganalisis ulang data pencangkokan jantung yang pernah dikaji oleh Segal (1988) dengan menggunakan model regresi log-normal dengan

penyesuaian pada fungsi pemilahan (dengan 2 metode, metode M dan R) dan aturan penghentian yang digunakan. Fungsi pemilahannya berdasarkan analisis sebaran sisaan regresi model log-normal, yakni menggunakan uji *t* bagi rataan simpul kiri dan kanan dan uji Levene bagi ragam pada dua calon simpul. Pemilahan yang dipilih adalah pemilahan dengan peluang nyata (*p-value*) dari statistik *t* dan statistik Levene paling kecil. Sedangkan aturan penghentiannya menggunakan penduga *bootstrap* bagi salah jenis pertama, yakni α = P(Suatu simpul dipilah|H_0 : simpul tersebut jangan dipilah). Hasilnya dengan menggunakan metode M terbentuk 5 kelompok pasien, masing-masing dengan median *survival time* 697 hari, 584.5 hari, 204 hari, 273 hari dan 136 hari. Sedangkan dengan menggunakan metode R diperoleh 4 kelompok pasien, masing-masing dengan median *survival time* 697 hari, 544 hari, 292 hari dan 129 hari.

Tibshirani & Hinton (1998) menggunakan metode penyekatan rekursif pada ruang peubah "*coaching*" dengan tujuan untuk memperbaiki dugaan respon berdasarkan peubah penjelas di masa datang. Peubah "*coaching*" adalah peubah yang sulit diukur atau mahal dan tersedia pada *training sample* yang digunakan, tetapi di masa datang peubah tersebut sulit tersedia, sehingga model yang menyatakan hubungan antara peubah respon dengan peubah penjelas didasarkan pada sekatan-sekatan contoh peubah "*coaching*".

METODE PENELITIAN

Data

Data yang digunakan dalam penelitian ini berasal dari Bagian Perinatologi Rumah Sakit Hasan Sadikin Bandung. Data tersebut merupakan karakteristik ibu dan bayi yang dilahirkan selama kurun waktu Januari sampai Juni 1998. Peubah-peubah yang diduga mempunyai hubungan dengan masa rawat kelahiran bayi antara lain :

1. **Riwayat aborsi**. Menurut Johnston (1994) *dalam* Gusnanto (1998) bahwa seorang ibu yang memiliki riwayat abortus beresiko menimbulkan kematian perinatal pada bayi yang dikandung pada kehamilan berikutnya.
2. **Asfiksia**. Peubah ini menunjukkan ada tidaknya gangguan fungsi-fungsi fisiologis pernapasan bayi setelah ia dilahirkan. Apabila terjadi gangguan atau bahkan kegagalan, maka kemampuan bayi bertahan hidup selama masa awal kehidupannya akan berkurang. Peubah ini mempunyai tiga taraf, yakni benilai 0 jika tidak mengalami asfiksia, bernilai 1 jika mengalami asfiksia ringan dan sedang serta bernilai 2 jika mengalami asfiksia berat.
3. **Usia Kehamilan**. Peubah ini berhubungan dengan tingkat kematangan sistem organ bayi yang diduga mempunyai hubungan dengan masa rawat kelahiran.
4. **Berat lahir**. Berat bayi saat dilahirkan yang diukur dalam gram.
5. **Pendidikan ibu**. Peubah ini merupakan faktor *sosio demografi* yang diduga merupakan salah satu faktor resiko tinggi bagi ketahanan hidup bayi (Nat. Acad. of Science, 1970 *dalam* Husaini, 1990).
6. **Usia ibu**. Usia ibu pada saat melahirkan dalam tahun.
7. **Masa rawat**. Peubah ini merupakan peubah respon yang diukur dalam hari sejak bayi dilahirkan sampai pulang atau pulang paksa atau meninggal. Bagi bayi yang pulang sesuai rekomendasi paramedis masa rawatnya dikatakan lengkap, sedangkan bagi bayi yang pulang paksa atau meninggal maka masa rawatnya tersensor karena tidak teramati secara lengkap.

Pembandingan Pohon Regresi

Berdasarkan dua pendekatan pohon regresi bagi data ketahanan hidup yang ada, dalam penelitian ini dibandingkan kinerja dari dua pendekatan tersebut. Pohon regresi yang dibandingkan adalah yang dikemukakan oleh Segal (1988) yang mewakili pendekatan ukuran pemisahan dan yang dikemukakan oleh LeBlanc & Crowley (1992) yang mewakili pendekatan ukuran kehomogenan. Pembandingan meliputi : (1) pohon awal yang terbentuk, (2) pemangkasan yang dilakukan dalam menentukan pohon terbaik, (3) dugaan fungsi ketahanan masing-masing kelompok amatan yang terbentuk dan dugaan masa rawatnya serta (4) resiko relatif antar kelompok.

Pohon awal dibentuk dengan menetapkan banyaknya amatan minimum pada daun sebanyak 50 amatan. Hal ini berlandaskan apa yang dilakukan oleh Schmoor *et. al* (1993) dimana untuk ukuran contoh 447 ia menetapkan banyaknya amatan pada daun sebanyak 25.

Dugaan masa rawat adalah nilai median Kaplan-Meier. Jika median Kaplan-Meier dari kelompok amatan tertentu adalah t hari, maka peluang individu pada kelompok tersebut harus dirawat lebih dari t hari adalah setengah.

Resiko relatif masing-masing kelompok ditentukan melalui analisis regresi Cox dengan kovariatnya adalah kelompok yang terbentuk. Jika pohon terbaik mempunyai a daun berarti terbentuk a kelompok amatan, maka kovariatnya memiliki a taraf. Kovariat tersebut diganti dengan $a - 1$ buah peubah boneka yakni $X_2, X_3, \ldots, X_a$ dengan mengambil nilai seperti pada Tabel 2.

Tabel 2. Nilai-nilai peubah boneka bagi kovariat dengan a taraf.

Taraf	X_2	X_3	...	X_a
1	0	0	...	0
2	1	0	...	0
3	0	1	...	0
...	...	...	...	...
a	0	0	...	1

Fungsi *hazard* bagi amatan dalam kelompok ke-j, j=1,2,...,a adalah

$$h_j(t) = \exp(\alpha_2 x_{2j} + \alpha_3 x_{3j} + \ldots + \alpha_a x_{aj})\, h_0(t)$$

atau karena $X_2, X_3, \ldots, X_n$ merupakan peubah boneka, maka berarti

$$h_j(t) = \exp(\alpha_j)\, h_0(t)$$

dengan α_j adalah pengaruh dari kelompok ke-j dan $h_0(t)$ adalah fungsi *hazard* dasar, yakni untuk kelompok pertama dimana $\alpha_1 = 0$. Resiko relatif dari kelompok ke-j terhadap kelompok pertama adalah nisbah *hazard* dari amatan dalam kelompok ke-j terhadap amatan dalam kelompok pertama yakni $\exp(\alpha_j)$. Hal ini berarti parameter α_j adalah logaritma dari resiko relatif, yakni

$$\alpha_j = \log \{ h_j(t)/ h_0(t)\}$$

dan selang kepercayaan $100(1-\alpha)\%$ bagi logaritma nisbah *hazard* adalah

$$\hat{\alpha}_j \pm z_{\alpha/2}\text{s.e.}(\hat{\alpha}_j)$$

dengan $z_{\alpha/2}$ adalah nilai kritis atas $\alpha/2$ dari sebaran normal baku (Collet, 1994).

Pengolahan data dilakukan dengan *software* S-Plus 3.2. Untuk pembentukan pohon menurut Segal (1988) menggunakan *library* tssa (*tree-structured survival analysis*) dari Wager & Segal (1996) yang diperoleh dari alamat *Universe Resource Locator* http://www.stats.ox.ac.uk/pub/SWin/tssa.zip, sedangkan bagi pembentukan pohon menurut LeBlanc & Crowley (1992) menggunakan *library* RPart (*Recursive Partitioning*) dan *library* Surv4 yang dibuat oleh Therneau & Atkinson (1997) dan diperoleh dari StatLib, http://lib.stat.cmu.edu/S/.

HASIL DAN PEMBAHASAN

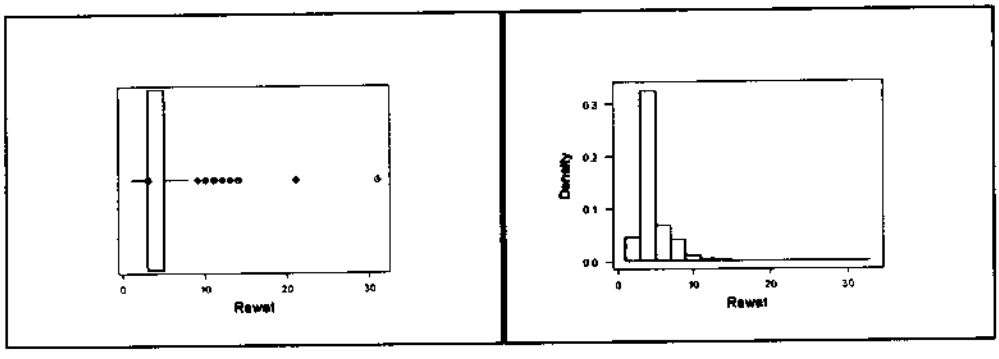
Gambaran Umum Data

Data yang lengkap tentang kelahiran bayi ada sebanyak 1143 yang terdiri dari 248 kasus (21.7%) dengan amatan masa rawat yang tersensor, yakni bayi yang pulang paksa atau meninggal dan sisanya mempunyai masa rawat yang lengkap, 148 kasus (12.9%) ibu yang pernah aborsi dalam riwayat kelahiran anak-anaknya, dan ada 140 kasus (12.2%) bayi yang mengalami asfiksia, baik asfiksia ringan, sedang maupun berat. Gambaran dari peubah-peubah lainnya ditampilkan pada Tabel 3.

Tabel 3. Gambaran umum data yang dianalisis.

Peubah	Min	Max	Jangkauan	Rataan	Simp. Baku
Usia Kehamilan (minggu)	24	45	21	38.7	2.0
Berat Lahir(gram)	630	4400	3770	2935.9	496.8
Pendidikan Ibu (tahun)	2	27	25	11.6	3.6
Usia Ibu (tahun)	15	46	31	28.0	5.9
Masa Rawat (hari)	1	31	30	4.4	2.9

Peubah masa rawat diduga mempunyai sebaran yang menceng ke kanan, dimana terdapat amatan-amatan yang mempunyai masa rawat yang lama. Hal ini diperiksa melalui diagram kotak garis dan histogram untuk data masa rawat yang lengkap (tidak tersensor). Kedua ringkasan statistik tersebut ditampilkan pada Gambar 2.



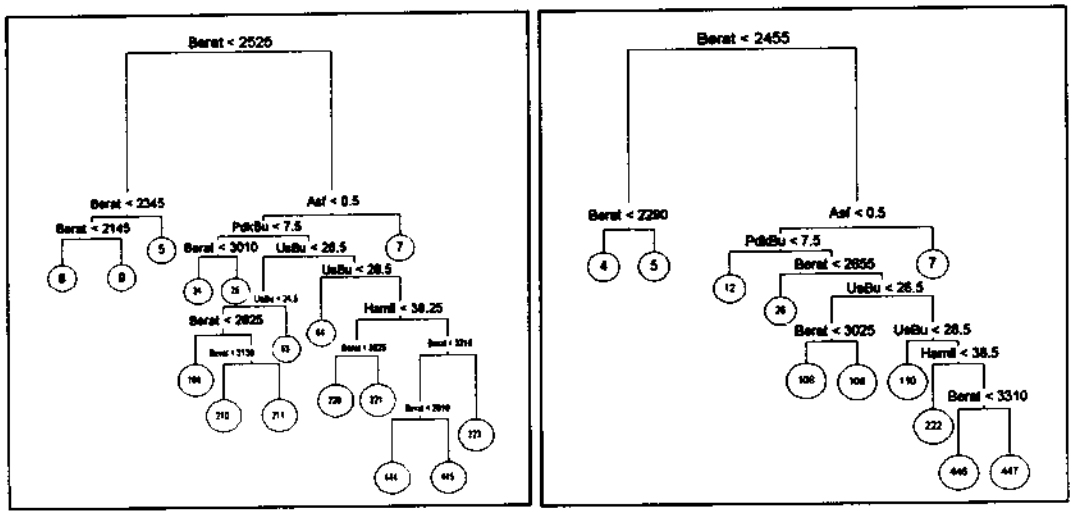
Gambar 2. Diagram kotak garis dan histogram bagi data masa rawat yang tidak tersensor.

Pembandingan Pohon Regresi

Dua pendekatan pembentukan pohon regresi, yaitu metode pendekatan ukuran kehomogenan (Segal, 1988) dan metode pendekatan ukuran pemisahan (LeBlanc & Crowley, 1992) digunakan untuk melihat perbandingan kinerja masing-masing pendekatan. Pembandingan tersebut meliputi : (1) pohon awal yang terbentuk dengan menetapkan banyaknya amatan minimum dalam daun sebanyak 50, (2) pemangkasan yang dilakukan terhadap pohon awal sebagai langkah untuk mencari pohon dengan ukuran terbaik (*right sized tree*), (3) dugaan fungsi ketahanan masing-masing kelompok (daun) yang terbentuk pada pohon terbaik dan dugaan masa rawatnya serta (4) penentuan resiko relatif antar kelompok.

Pohon Regresi Awal

Pohon regresi yang dibentuk dengan kedua pendekatan menampakkan titik pemilahan yang hampir sama, seperti tercantum pada Gambar 3.



Gambar 3. Pohon awal yang dibentuk dengan pendekatan ukuran pemisahan (kiri) dan ukuran kehomogenan (kanan).

Pada pohon yang dibentuk dengan pendekatan ukuran pemisahan (kiri) pemilahan pertama (simpul akar atau *root node*) terjadi pada berat lahir 2525 gr. Pada pohon yang dibentuk dengan pendekatan ukuran kehomogenan terjadi pada berat lahir 2455 gr. Hal ini mendukung kriteria berat bayi lahir rendah (BBLR) yang dikemukakan oleh WHO sejak tahun 1961 (Nat. Acad. Press, 1985 *dalam*

Husaini, 1990) bahwa BBLR adalah bayi yang dilahirkan dengan berat lahir kurang dari 2500 gram. Dari hasil tersebut tampaknya BBLR dan bayi dengan berat lahir cukup (BBLC, berat lahir lebih dari 2500 gr) memiliki *survival experience* atau masa rawat yang berbeda.

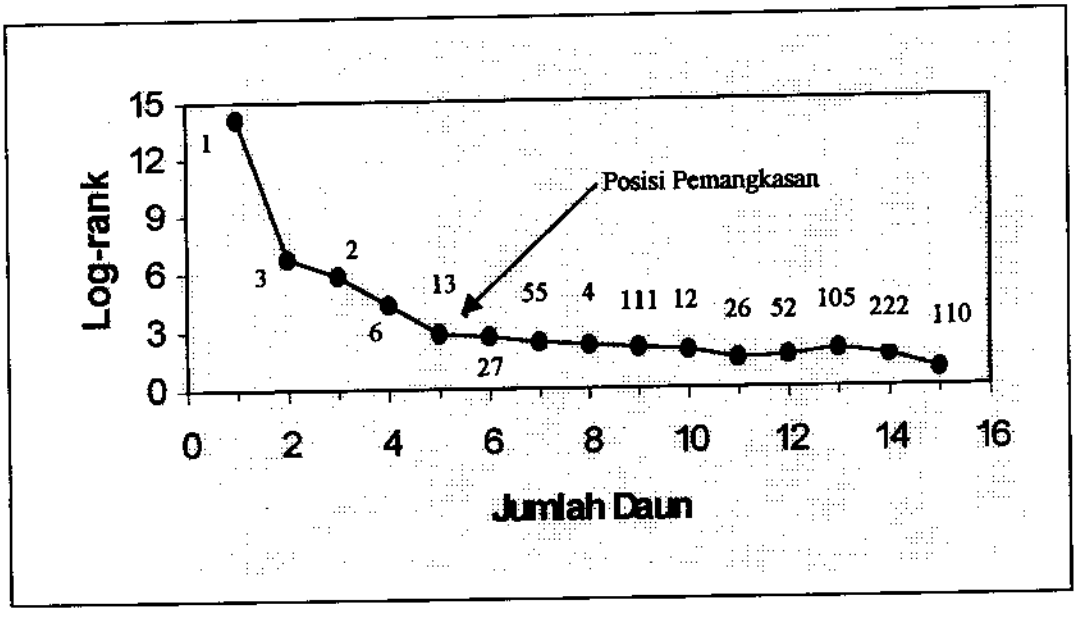
Kemiripan dua pohon regresi tampak lebih banyak lagi pada cabang dengan berat lahir yang cukup. Simpul nomor tiga [2)] dipilah pada asfiksia 0.5, baik untuk pohon regresi sebelah kiri (pendekatan ukuran pemisahan) maupun pohon regresi sebelah kanan (pendekatan ukuran kehomogenan). Selanjutnya kelompok bayi yang mempunyai berat lahir cukup dan tidak mengalami asfiksia (asfiksia = 0) dipilah berdasarkan peubah pendidikan ibu pada titik 7.5 tahun baik pada pohon regresi sebelah kiri maupun kanan. Pemilahan ini berarti memilah bayi yang ibunya berpendidikan SD (< 7.5 tahun) dan berpendidikan SMP atau lebih (≥ 7.5 tahun). Pada simpul-simpul berikutnya muncul peubah-peubah berat lahir, usia ibu dan usia kehamilan sebagai pemilah. Secara umum peubah riwayat aborsi tidak muncul sebagai pemilah.

Pohon awal sebelah kiri berukuran 16 daun, sedangkan yang sebelah kanan mempunyai 11 daun. Adanya perbedaan tersebut sebagai akibat dari penetapan banyaknya amatan dalam daun minimum sebanyak 50 amatan. Pohon sebelah kanan yang mendasarkan pembentukan pohonnya pada pemilahan dengan statistik log-rank memiliki jumlah daun lebih banyak, karena walaupun nilai statistik log-rank-nya kecil tetapi simpul-simpulnya masih bisa dipilah. Sedangkan jika pemilahannya berdasarkan pendekatan ukuran kehomogenan, perbedaan tingkat kehomogenan simpul induk dengan dua simpul anaknya sudah tidak terlalu berbeda, sehingga pemilahannya dihentikan. Pohon sebelah kiri mempunyai devians sebesar 779.46. Devians ini dapat dipandang sebagai perampatan dari jumlah kuadrat sisaan yang biasa digunakan pada pemodelan data sebaran normal. Lebih jelasnya devians ini merupakan penjumlahan dari kuadrat sisaan devians, yakni Devians = Σr_{Di}^2 (Collet, 1994).

[2)] Nomor simpul dimulai dengan nomor 1 untuk simpul akar, dan jika suatu simpul bernomor r maka simpul anak kiri bernomor 2r sedangkan simpul kanan bernomor 2r+1

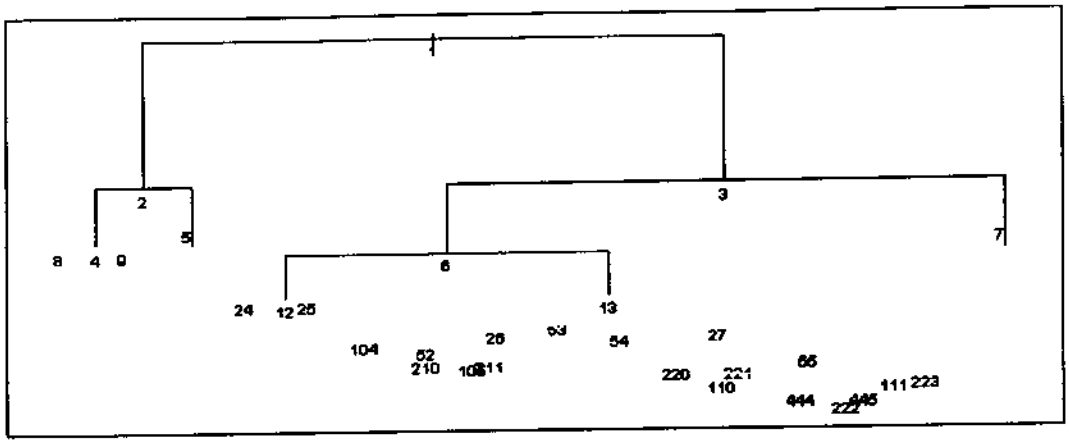
Pemangkasan

Strategi pemangkasan dari kedua pendekatan berbeda. Segal (1988) melakukan pemangkasan berdasarkan plot antara statistik log-rank dengan banyaknya daun, seperti pada Gambar 4.



Gambar 4. Plot statistik log-rank dengan jumlah daun yang digunakan dalam strategi pemangkasan

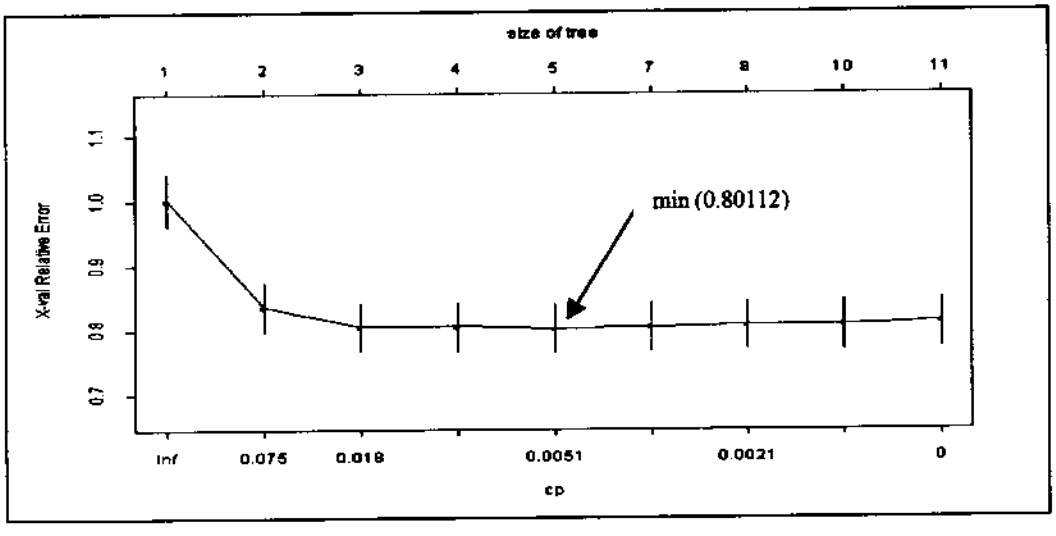
Berdasarkan Gambar 4 pemangkasan dilakukan pada simpul nomor 13 beserta simpul-simpul lainnya yang memiliki statistik log-rank yang lebih kecil. Posisi pemangkasan tersebut adalah pada saat perbedaan dua nilai statistik log-rank hampir sama (statistik log-rank simpul nomor 13 bernilai 2.79 sedangkan simpul nomor 27 bernilai 2.7). Pemangkasan tersebut menghasilkan pohon regresi dengan jumlah daun 5. Pohon regresi yang dihasilkan dengan strategi pemangkasan ini tercantum pada Gambar 5. Pada Gambar 5 tampak bahwa walaupun simpul nomor 13 memiliki statistik log-rank yang hampir sama dengan simpul nomor 27, tetapi simpul nomor 27 berada di bawahnya sehingga prioritas pemangkasan berada pada simpul nomor 13, karena secara otomatis simpul nomor 27 juga akan terpangkas.



Gambar 5. Pohon regresi dengan pendekatan ukuran pemisahan yang sudah dipangkas.

Pemangkasan sebagai upaya menentukan pohon terbaik menurut LeBlanc & Crowley (1992) menggunakan prosedur *cost-complexity* minimum yang dikemukakan oleh Breiman *et al.* (1993). *Honest estimate*-nya menggunakan *cross-validation estimate* dengan 10 lipatan (10-*fold*).

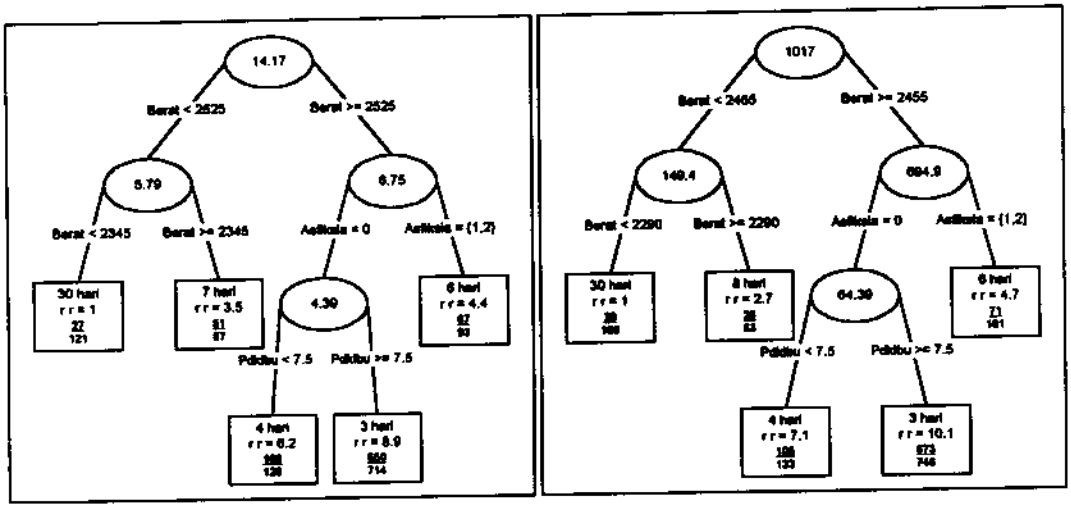
Pohon regresi berukuran 5 mempunyai R^{CV} yang paling kecil seperti terlihat pada Gambar 6. Parameter *Complexity* yang bersesuaian dengan pohon berukuran 5 ini adalah sebesar 0.0051. Nilai ini sebenarnya adalah α dibagi dengan $R(T_0)$ sehingga nilai-nilai parameter *complexity* berada pada selang [0,1] (Venables & Ripley, 1999).



Gambar 6. Plot R^{CV} (x-val Relative Error) terhadap ukuran pohon dan parameter *complexity*.

Dugaan Fungsi Ketahanan dan Masa Rawat

Pohon terbaik yang dihasilkan dari pemangkasan, masing-masing untuk pohon regresi dengan pendekatan ukuran pemisahan dan pendekatan ukuran kehomogenan sama-sama berukuran 5. Kedua pohon tersebut ditampilkan pada Gambar 7.



Keterangan : Angka di dalam ellips adalah statistik log-rank (pohon sebelah kiri) dan devians (pohon sebelah kanan), sedangkan angka di dalam kotak masing-masing adalah dugaan masa rawat, resiko relatif (rr), banyaknya bayi yang sembuh (amatan lengkap) dan banyaknya bayi dalam kelompok (daun) tersebut.

Gambar 7. Pohon terbaik yang dibentuk dengan pendekatan ukuran pemisahan (kiri) dan ukuran kehomogenan (kanan).

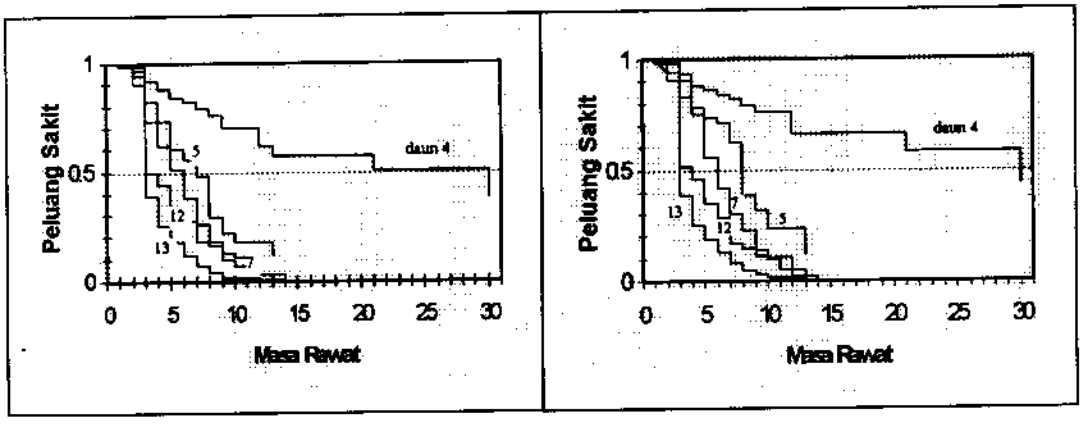
Lima kelompok bayi yang terbentuk berdasarkan pohon regresi dengan pendekatan ukuran pemisahan adalah (1) bayi dengan berat kurang dari 2345 gr (daun 4), (2) bayi dengan berat badan antara 2345 sampai 2525 gr (daun 5), (3) bayi dengan berat badan lebih dari 2525 gr dan mengalami asfiksia, baik ringan sedang ataupun berat (daun 7), (4) bayi dengan berat badan lebih dari 2525 gr, tidak mengalami asfiksia, dan beribukan dengan pendidikan SD ke bawah (daun 12) dan (5) bayi dengan berat badan lebih dari 2525 gr, tidak mengalami asfiksia dan beribukan dengan pendidikan SMP atau lebih tinggi (daun 13).

Adapun kelompok bayi yang dihasilkan dari pohon regresi berdasarkan pendekatan ukuran kehomogenan juga sebanyak lima kelompok, yakni (1) bayi dengan berat badan kurang dari 2290 gr (daun 4), (2) bayi dengan berat badan antara 2290 sampai 2455 gr (daun 5), (3) bayi dengan berat badan lebih dari 2455

gr dan mengalami asfiksia dalam masa perawatannya (daun 7), (4) bayi dengan berat badan lebih dari 2455 gr, tidak mengalami asfiksia dan beribukan dengan pendidikan SD ke bawah (daun 12) dan (5) bayi dengan berat badan lebih 2455 gr, tidak mengalami asfiksia dan beribukan yang berpendidikan SMP ke atas (daun 13).

Secara umum kedua pendekatan metode pohon regresi memberikan hasil yang mirip. Lima kelompok bayi mempunyai karakteristik yang hampir sama beserta nilai dugaan median masa rawat yang sama, kecuali dugaan median masa rawat daun 5 (7 hari dan 8 hari).

Karakteritik masa rawat masing-masing kelompok bayi dapat dilihat dari kurva fungsi ketahanannya, dalam hal ini adalah kurva peluang bahwa seseorang bayi harus dirawat (masih sakit) melebihi suatu waktu t tertentu, sehingga jika peluang seorang bayi harus melewati suatu masa rawat tertentu t bernilai besar berarti bayi tersebut adalah lemah. Kurva fungsi ketahanan Kaplan-Meier masing-masing kelompok bayi dari kedua pohon regresi yang terbentuk ditampilkan pada Gambar 8.



Gambar 8. Kurva fungsi ketahanan masing-masing kelompok bayi yang dihasilkan dari pohon regresi dengan pendekatan ukuran pemisahan (kiri) dan ukuran kehomogenan (kanan).

Secara umum berdasarkan kurva fungsi ketahanan, baik kurva di sebelah kiri maupun kanan, susunan kelompok bayi dari yang paling lemah (dalam hal peluang masih sakit setelah dirawat selama t hari) ke yang paling kuat adalah :

(1) Kelompok bayi paling lemah, yakni bayi yang berada dalam daun 4 yang mempunyai karakteristik berat lahir sangat rendah.

(2) Kelompok bayi lemah, yakni bayi yang berada dalam daun 5 yang mempunyai ciri berat lahir rendah.

(3) Kelompok bayi dengan kekuatan sedang, yakni bayi di dalam daun 7 yang bersifat : berat lahir cukup dan mengalami asfiksia.

(4) Kelompok bayi yang kuat, yakni bayi yang berada dalam daun 12 yang bersifat mempunyai berat lahir cukup, tidak mengalami asfiksia dan beribukan dengan pendidikan SD ke bawah.

(5) Kelompok bayi paling kuat, yakni bayi yang berada dalam daun 13 yang mempunyai ciri : berat lahir cukup, tidak mengalami asfiksia dan beribukan dengan pendidikan SMP ke atas.

Dugaan masa rawat masing-masing kelompok bayi adalah median masa rawat yang dihitung dengan menarik garis mendatar pada peluang sakit bernilai 0.5 pada kurva fungsi ketahanannya, $\hat{S}^{-1}(0.5)$. Nilai-nilai dugaan median masa rawat tersebut ditampilkan pada Gambar 7. Pada pohon regresi yang dihasilkan dengan pendekatan ukuran pemisahan, dugaan median masa rawat masing-masing kelompok bayi tersebut adalah :

(1) Kelompok bayi paling lemah mempunyai dugaan masa rawat 30 hari.

(2) Kelompok bayi lemah mempunyai dugaan masa rawat 7 hari.

(3) Kelompok bayi dengan kekuatan sedang mempunyai dugaan masa rawat 6 hari.

(4) Kelompok bayi yang kuat mempunyai dugaan masa rawat 4 hari.

(5) Kelompok bayi paling kuat mempunyai dugaan masa rawat 3 hari.

Sedangkan dugaan masa rawat bagi masing-masing kelompok bayi yang dihasilkan dari pohon yang dibentuk dengan pendekatan ukuran kehomogenan adalah sebagai berikut :

(1) Kelompok bayi paling lemah mempunyai dugaan masa rawat 30 hari.

(2) Kelompok bayi lemah mempunyai dugaan masa rawat 8 hari.

(3) Kelompok bayi dengan kekuatan sedang mempunyai dugaan masa rawat 6 hari.

(4) Kelompok bayi yang kuat mempunyai dugaan masa rawat 4 hari.

(5) Kelompok bayi paling kuat mempunyai dugaan masa rawat 3 hari.

Dugaan Resiko Relatif

Lima kelompok bayi yang terbentuk dari masing-masing pendekatan pembentukan pohon regresinya ingin dilihat resiko relatifnya satu sama lain. Dengan menganggap lima kelompok tersebut adalah taraf dari suatu kovariat, maka berarti ada empat peubah boneka X_2, X_3, X_4 dan X_5 yang mempunyai nilai sebagai berikut :

Taraf	X_2	X_3	X_4	X_5
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

dimana peubah boneka X_2 adalah peubah boneka untuk kelompok bayi lemah (daun 5), peubah boneka X_3 untuk kelompok bayi sedang (daun 7), peubah boneka X_4 untuk kelompok bayi kuat (daun 12) dan peubah boneka X_5 untuk kelompok bayi paling kuat (daun 13). Hal ini berarti bayi di kelompok paling lemah (daun 4) menjadi kontrol. Dengan demikian model regresinya adalah :

$$h_j(t) = \exp(\alpha_2 x_{2j} + \alpha_3 x_{3j} + \ldots + \alpha_5 x_{5j})\, h_0(t)$$

Bagi pohon regresi dengan pendekatan ukuran pemisahan, dugaan-dugaan parameternya ditampilkan pada Tabel 4.

Tabel 4. Dugaan parameter model regresi Cox untuk kelompok-kelompok bayi pada pohon regresi dengan pendekatan ukuran pemisahan.

Peubah Boneka	α	Simp. Baku	Wald	db	Nilai-P
X_2	1.2550	0.2460	26.0275	1	0.0000
X_3	1.4808	0.2378	38.7789	1	0.0000
X_4	1.8181	0.2265	64.4457	1	0.0000
X_5	2.1836	0.2087	109.4904	1	0.0000

Peubah Boneka	r r = Exp(α)	SK 95% bagi Exp(α)	
		Batas Bawah	Batas Atas
X_2	3.5	2.2	5.7
X_3	4.4	2.8	7.0
X_4	6.2	4.0	9.6
X_5	8.9	5.9	13.4

Oleh karena $\alpha_1 = 0$, maka resiko relatif (rr) dari bayi yang termasuk kelompok lemah (daun 5) terhadap bayi di kelompok paling lemah (daun 4) adalah sebesar

$e^{1.2550} = 3.5$. Nilai resiko relatif ini berarti laju kesembuhan pada waktu tertentu dari bayi di kelompok lemah adalah 3.5 kali laju kesembuhan bayi di kelompok paling lemah. Hal ini didukung dengan selang kepercayaan 95% bagi resiko relatif kedua kelompok yaitu selang (2.2, 5.7) yang tidak memuat nilai 1. Begitu pula bagi kelompok-kelompok lainnya yang masing-masing mempunyai nilai resiko relatif 4.4, 6.2 dan 8.9. Tampak bahwa laju kesembuhan pada waktu tertentu dari bayi pada kelompok paling kuat terhadap bayi pada kelompok paling lemah hampir 9 kalinya. Nilai-nilai resiko relatif tersebut juga tercantum pada pohon regresi menurut pendekatan ukuran pemisahan pada Gambar 7 sebelah kiri.

Resiko relatif antarkelompok bayi yang dihasilkan dari pohon regresi dengan pendekatan ukuran kehomogenan dapat dilihat dari Tabel 5 di bawah ini. Nilai-nilai resiko relatif ini juga dicantumkan pada pohon regresi Gambar 7 sebelah kanan.

Tabel 5. Dugaan parameter model regresi Cox untuk kelompok-kelompok bayi pada pohon regresi dengan pendekatan ukuran kehomogenan.

Peubah Boneka	α	Simp. Baku	Wald	db	Nilai-P
X_2	0.9873	0.3089	10.2145	1	0.0014
X_3	1.5562	0.2676	33.8311	1	0.0000
X_4	1.9642	0.2586	57.7079	1	0.0000
X_5	2.3158	0.2439	90.1158	1	0.0000

Peubah Boneka	r r = Exp(α)	SK 95% bagi Exp(α)	
		Batas Bawah	Batas Atas
X_2	2.7	1.5	4.9
X_3	4.7	2.8	8.0
X_4	7.1	4.3	11.8
X_5	10.1	6.3	16.3

Dari Tabel 5 tampak bahwa resiko relatif yang mengukur laju kesembuhan pada waktu tertentu antarkelompok bayi mempunyai nilai dengan kisaran yang lebih lebar. Resiko relatif kelompok bayi yang lemah terhadap bayi di kelompok yang paling lemah adalah 2.7 kali. Kemudian resiko relatifnya terhadap kelompok bayi sedang, kuat dan paling kuat masing-masing bernilai 4.7 kali, 7.1 kali dan 10.1 kali. Hal ini menguatkan hasil pengelompokan bayi yang dihasilkan oleh kedua pendekatan pembentukan pohon regresi.

KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini dapat disimpulkan sebagai berikut :

1. Pohon regresi masa rawat kelahiran bayi yang dibentuk berdasarkan ukuran pemisahan yang dikemukakan oleh Segal (1988) adalah serupa dengan yang dihasilkan melalui pendekatan ukuran kehomogenan yang dikemukakan oleh LeBlanc & Crowley (1992).
2. Pohon regresi masa rawat kelahiran bayi menghasilkan lima kelompok bayi, yakni :
   1) Kelompok bayi paling lemah yang mempunyai sifat berat lahir sangat rendah.
   2) Kelompok bayi yang lemah dengan sifat berat lahirnya yang rendah.
   3) Kelompok bayi dengan kekuatan sedang yang bersifat berat lahir cukup tetapi mengalami gangguan fungsi-fungsi fisiologis setelah kelahirannya (asfiksia).
   4) Kelompok bayi yang kuat yang mempunyai ciri berat lahir cukup, tidak mengalami asfiksia tetapi pendidikan ibunya hanya lulus SD atau kurang.
   5) Kelompok bayi paling kuat yang bersifat berat lahirnya cukup, tidak mengalami asfiksia dan beribukan yang berpendidikan SMP atau lebih.
3. Dugaan median masa rawat kelahiran bayi menurut dua pohon regresi yang dikaji adalah sebagai berikut :
   1) Kelompok bayi paling lemah mempunyai dugaan median masa rawat 30 hari.
   2) Kelompok bayi lemah mempunyai dugaan median masa rawat 7 hari (menurut ukuran pemisahan) atau 8 hari (menurut ukuran kehomogenan).
   3) Kelompok bayi sedang mempunyai dugaan median masa rawat 6 hari.
   4) Kelompok bayi kuat mempunyai dugaan median masa rawat 4 hari.
   5) Kelompok bayi paling kuat mempunyai dugaan median masa rawat 3 hari.
4. Laju kesembuhan masing-masing kelompok bayi relatif terhadap bayi dalam kelompok paling lemah yang dinyatakan dalam resiko relatif untuk kedua pohon regresi adalah sebagai berikut :

1) Kelompok bayi lemah mempunyai laju kesembuhan 3.5 kali (menurut ukuran pemisahan) atau 2.7 kali (menurut ukuran kehomogenan).
2) Kelompok bayi sedang mempunyai laju kesembuhan 4.4 kali (menurut ukuran pemisahan) atau 4.7 kali (menurut ukuran kehomogenan).
3) Kelompok bayi kuat mempunyai laju kesembuhan 6.2 kali (menurut ukuran pemisahan) atau 7.1 kali (menurut ukuran kehomogenan).
4) Kelompok bayi paling kuat mempunyai laju kesembuhan 8.9 kali (menurut ukuran pemisahan) atau 10.1 kali (menurut ukuran kehomogenan).

Saran

Hasil pengelompokan bayi berdasarkan masa rawat perlu dikonfirmasikan dengan pengelompokan yang berdasarkan data ketahanan hidup. Data ketahanan hidup mempunyai interpretasi yang lebih mengena berkaitan dengan kemampuan hidup bayi setelah dilahirkan.

DAFTAR PUSTAKA

Ahn, H. 1996. Log-normal regression modeling through recursive partitioning. *Comput. Statist. Data Anal.* **21**:381-196.

Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone. 1993. *Classification and Regression Trees*. Chapman and Hall, New York.

Clark, L.A. & D. Pregibon. 1992. Tree-based Models. dalam **Chambers, J.M. & T. J. Hastie**, editor, *Statistical Model in S, chapter 9*. Wadsworth and Brooks/Cole, Pacific Grove, California.

Collett, D. 1994. *Modelling Survival Data in Medical Research.* Chapman and Hall, London.

Davis, R. B. & J. R. Anderson. 1989. Exponential survival trees. *Statist. Med.* **8**:947-961.

Friedman, J. H. & W. Stuetzle. 1981. Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**:817-823.

Gusnanto, A. S. 1998. *Model Regresi Cox untuk Daya Tahan Hidup Bayi di Masa Perinatal*. Makalah Seminar Skripsi. Jurusan Statistika. Institut Pertanian Bogor (tidak dipublikasikan).

Hougaard, P. 1999. Fundamental of survival data. *Biometrics*. **55**:13-22.

Husaini, J. K. 1990. *Karakterisasi Wanita Hamil dalam Hubungannya dengan Berat Lahir dan Pertumbuhan Bayi Selanjutnya*. Tesis. Fakultas Pascasarjana. Institut Pertanian Bogor (tidak dipublikasikan).

Intrator, O. & Charles Kooperberg. 1995. Trees and Splines in Survival Analysis. *Technical Report, University of Washington.*

LeBlanc, M. & J. Crowley. 1992. Relative risk trees for censored survival data. *Biometrics*. **48**:411-425.

Saefuddin, A. 1996. *Statistical Analysis of Regression Models with Covariates Measured with Error*. Disertation. The Faculty of Graduate Studies of The University of Guelph (tidak dipublikasikan).

Schmoor, C., K. Ulm & M. Schumacher. 1993. Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Statist. Med.* **12**:2351-2366.

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

Segal, M. R. 1992. Tree-structured methods for longitudinal data. *J. Amer. Statist. Assoc.* 87:407-418.

Therneau, M, T., & Elizabeth Atkinson. 1997. An Introduction to Recursive Partitioning Using RPART Routines. *Technical Report, Mayo Foundation.*

Therneau, T. M., P. M. Grambsch & T. R. Flemming. 1990. Martingale based residuals for survival models. *Biometrics.* 77:147-160.

Tibshirani, R. & G. Hinton. 1998. Coaching variables for regression and classification. *Statistics and Computing.* 8:25-33.

Venables, W. N. & B. D. Ripley. 1996. *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.

Venables, W. N. & B. D. Ripley. 1999. *Statistics Complement to Modern Applied Statistics with S-Plus, 2nd Edition.* Springer-Verlag, New York.

Wager, C. G. & M. R. Segal. 1996. Tree Structured Survival Analysis Methods and Software. *Technical Report, Dept. of Biostatistics, Harvard School of Public Health.*