

Diagnostik Pengaruh bagi Model Risiko Bersaingan dengan Tapisan Sebagai Satu Risiko

NOOR AKMA IBRAHIM

dan ISA BIN DAUD

Jabatan Matematik

Fakulti Sains & Pengajian Alam Sekitar

Universiti Pertanian Malaysia

43400 UPM Serdang, Selangor Darul Ehsan/Malaysia

Kata kunci: Risiko bersaing, cerapan tertapis, kaedah satu-langkah, jarak Cook, jarak kebolehjadian.

ABSTRAK

Kaedah mengesan cerapan berpengaruh bagi model risiko bersaing yang data tertapis dianggap sebagai satu daripada risiko dalam model dua risiko yang dikemukakan. Kaedah yang disarankan termasuk kaedah penghapusan cerapan, penghapusan satu langkah, jarak Cook dan jarak kebolehjadian. Penekanan diberi ke atas kesan suatu cerapan terhadap penganggapan parameter terhasil. Dua set data digunakan untuk mengilustrasikan kaedah-kaedah ini.

ABSTRACT

Methods of detecting influential observations for competing risks model with censoring as one of the risk in a two-risks model are proposed. These methods include deletion of observations, one-step deletion, Cook distance and likelihood distance. Emphasis is on assessing the impact of an observation on the parameter estimation. Two sets of data are used to illustrate these techniques.

1. PENGENALAN

Kertas ini megemukakan penyelidikan ke atas cerapan berpengaruh dalam model risiko bersaing dengan data tertapis sebagai salah satu risiko. Model ini digunakan bagi menganalisis data mandirian yang mempunyai dua atau lebih punca atau risiko yang menyebabkan kegagalan. Teori risiko bersaing bermula pada kurun kelapan belas dan pada tahun 1760 mendapat perhatian sains yang lain apabila Daniel Bernoulli, seorang ahli matematik juga ahli perubatan mengemukakan soalan berikutan dengan isu semasa ketika itu dan soalnya ialah jika bagi suatu populasi, cacar boleh dihapuskan, apakah kesannya terhadap struktur mortaliti pada peringkat umur yang berbeza? Beliau mengenap pasti bahawa individu yang terselamat daripada cacar akan menghadapi risiko seterusnya sama seperti individu yang lain di dalam populasi tersebut.

Bagi penyesuaian masa kini cacar boleh digantikan dengan barah atau penyakit jantung, umpamanya dan kebarangkalian kematian disebabkan penyakit tersebut dengan kehadiran risiko yang lain boleh dikirakkan. Selain data klinikal, teori ini juga boleh dilaksanakan ke atas data kebolehpercayaan komponen-komponen dalam bidang industri

termasuk juga bidang ekonomi, demografi dan juga aktuari.

Apabila menganalisis data mandirian, perhatian perlu diberi kepada cerapan tertapis kerana masa tertapis adalah juga masa kegagalan. Tapisan ini berlaku jika pemerhatian terpaksa diberhentikan sebelum terjadi kegagalan akibat kesuntukan masa atau kos mengendalikannya tinggi. Selain itu data perlu ditapis apabila kegagalan yang berlaku bukan disebabkan oleh punca yang diselidiki atau apabila individu hilang daripada kajian. Dalam model yang diselidik di sini, tapisan merupakan satu risiko kerana tapisan pada sesuatu masa x mengelakkan kematian daripada sesuatu risiko semasa ujian sama seperti kematian pada masa x dari risiko yang lain. Dengan ini, tapisan dalam model risiko bersaing akan dilayan sebagai satu risiko dan ia mengikut taburan yang tertentu. Di antara artikel dan buku yang telah membincangkan analisis data mandirian dengan cerapan tertapis ialah Kaplan dan Meier (1958), Cox (1972), Kalbfleisch dan Prentice (1980) dan Cox dan Oakes (1984).

Model risiko bersaing seperti model statistik yang lain hanyalah merupakan anggaran memperihal situasi sebenar tertakluk kepada beberapa kekangan dan andaian model. Disebabkan ketidaktepatan ini, kestabilan dan variasi keputusan analisis di bawah

modifikasi atau gangguan yang kecil ke atas fomulasi model perlu dikaji. Jika gangguan kacil ini mempengaruhi keputusan analisis, model atau data sebaiknya sikaji selidik semula. Sebaliknya jika kesannya tidak ketara, sampel adalah teguh terhadap gangguan yang disengajakan dan andaian model dianggap kukuh untuk memperihalkan ciri-ciri penting data tersebut.

Kebelakangan ini beberapa artikel dan buku telah diterbitkan yang membincangkan kaedah-kaedah penilaian pengaruh ke atas model yang diganggu termasuk mengenal pasti cerapan berpengaruh dan data terencil di dalam konteks analisis regresi. Cook (1977) telah mengemukakan satu statistik bagi menentukan pengaruh sesuatu cerapan di dalam model regresi biasa. Berikutan dengan ini, ramai penulis telah mencadangkan beberapa statistik untuk mengenal pasti satu atau lebih cerapan sebagai cerapan terencil, berpengaruh atau kedua-duasekali. Mengikut takrif, cerapan dianggap berpengaruh jika penghapusannya daripada data menghasilkan perubahan yang besar kepada ciri-ciri penting di dalam analisis. Ukuran-ukuran yang akan dipertimbangkan di dalam kertas ini adalah berdasarkan fungsi pengaruh empirikal yang telah dikemukakan oleh Cook (1977), Cook dan Weisberg (1980, 1982) dan Belsley, Kuh dan Welsch (1980) bagi model regresi. Akan tetapi penulis-penulis ini tidak mempertimbangkan data tertapis di dalam model-model mereka. Walaubagaimana-pun terdapat beberapa penyelidikan diagnostik ke atas data tertapis seperti di dalam penulisan Schoenfeld (1982), Bin Daud (1987) yang telah melaksanakannya dalam model bahaya kadar Cox (1972) dan Weissfeld dan Schneider (1990) dalam model normal linear.

Tumpuan kertas ini ialah megenal pasti cerapan-cerapan berpengaruh bagi model risiko bersaing dengan tapisan yang dianggap sebagai salah satu risiko. Kaedah-kaedah pendiagnosisan termasuklah penghapusan cerapan, penghapusan satu langkah dan ukuran pengaruh.

2. TATA - TANDA DAN FORMULASI

Katakan suatu populasi tertakluk kepada k punca kegagalan atau risiko, R_1, R_2, \dots, R_k dan katakan bagi setiap individu di dalam populasi tersebut dapat dicirikan oleh masa kegagalannya Y_1, Y_2, \dots, Y_k bersama R_1, R_2, \dots, R_k masing-masingnya. Pembolehkan rawak $Y_i (i=1, \dots, k)$ mewakili hayat seseorang individu R_i penyebab yang tunggal dan mempunyai fungsi taburan $P_i(x) = Kb(Y_i \leq x)$ dengan fungsi ketumpatan

$p_i(x)$. Dan fungsi bahayanya ialah $h_i(x) = p_i(x)/1 - P_i(x)$. Mengikut David dan Moeschberger (1978) dengan kehadiran semua k risiko $R_i (i=1, \dots, k)$ secara serentak, hanya $V = \text{minimum}(Y_1, \dots, Y_k)$ dan punca kegagalan sebenar, katakan R_i yang dapat dicerap. Berikutan dengan ini kebarangkalian V melebihi x iaitu fungsi mandiriannya ialah

$$Kb(V \geq x) = P(Y_1 > x, \dots, Y_k > x) \tag{1}$$

Gail (1975) telah menunjukkan hubungan (1) adalah benar bagi Y_1, Y_2, \dots, Y_k merdeka atau sebaliknya

Katakan X_i masa hayat tercerap bersyarat dari punca kegagalan R_i , maka

$$X_i = [Y_i | R_i] = \text{minimum}(Y_1, Y_2, \dots, Y_k) \tag{2}$$

Kebarangkalian kegagalan yang disebabkan oleh R_i ialah

$$\pi_i = Kb[Y_i = \text{minimum}(Y_1, \dots, Y_k)] \tag{3}$$

dengan $\pi_i > 0$ dan $\sum_{i=1}^k \pi_i = 1$.

Dari (2) dan kemerdekaan Y_i , fungsi ketumpatan $f_i(x)$ bagi X_i , tanpa mengambil kira sebutan peringkat bawah, diberi oleh

$$f_i(x) dt = \frac{1}{\pi_i} Kb [x - dx < Y_i \leq x] \prod_{j=1, j \neq i}^k Kb [Y_j > x]$$

yang demikian

$$f_i(x) = \frac{1}{\pi_i} p_i(x) \prod_{j=1, j \neq i}^k \bar{P}_j(x) \tag{4}$$

dengan $\bar{P}_e(x) = 1 - P_e(x)$.

Katakan N_i pembolehkan rawak mewakili bilangan individu yang gagal dari punca $R_i (i=1, 2, \dots, k)$ dan X_{ij} hayat individu ke $j (j=1, \dots, n_i)$ yang gagal dari punca R_i . Fungsi ketumpatan tercantum bagi X_{ij} boleh ditulis sebagai

$$f(x_{11}, x_{12}, x_{1n_1}, \dots, x_{kn_k}) = \prod_{i=1}^k \frac{1}{n_i} \prod_{j=1}^{n_i} p_i(x_{ij}) \prod_{i=1}^k \bar{P}_i(x_{ij}) \tag{5}$$

Fungsi ketumpatan kebarangkali ini adalah bersyaratkan ke atas $N_i = n_i (i = 1, \dots, k)$ dan N_i suatu

pembolehubah rawak multinational dengan fungsi ketumpatan kebarangkalinya

$$f(n_1, \dots, n_k) = \frac{n!}{\prod_{i=1}^k (n_i!)} \prod_{i=1}^k \pi_i^{n_i} \quad (6)$$

dan $\sum_{i=1}^k n_i = n$
 Jadi, fungsi kebolehhjadian yang menjadi tumpuan ialah

$$L = \frac{n!}{\prod_{i=1}^k (n_i!)} \prod_{i=1}^k \prod_{j=1}^{n_j} p_i(x_{ij}) \prod_{i=1}^k \bar{P}_i(x_{ij})$$

$$= \frac{n!}{\prod_{i=1}^k (n_i!)} \prod_{i=1}^k L_i$$

dengan

$$L_i = \left[\prod_{j=1}^{n_j} p_i(x_{ij}) \right] \left[\prod_{l=1, l \neq i}^k \prod_{j=1}^{n_j} \bar{P}_l(x_{ij}) \right] \quad (7)$$

2.1. Model risiko bersaing dengan kovariat

Dengan menggunakan model yang disarankan oleh Cox (1972), fungsi ketumpatan kebarangkalian bagi masa kehidupan teori bersekutu dengan punca R_i ialah

$$p_i(y) = h_{oi}(y) \exp(\beta^T Z) \exp[-H_{oi}(y) \exp(\beta^T Z)] \quad (8)$$

dengan β^T vektor baris bagi parameter s regresor dan Z vektor lajur bagi s kovariat berkaitan dengan sesuatu individu,

$$H_{oi}(y) = \int_0^y h_{oi}(t) dt$$

$$h_i(y) = p_i(y) / \bar{P}(y)$$

$$= h_{oi}(y) \exp(\beta^T Z)$$

Andaikan taburan pendasar bagi masa kehidupan ialah Weibull dengan $h_o(t) = \lambda c \gamma t^{-c-1} [h_o(t)]$ adalah

fungsi bahaya dengan kovariat sama sifar] maka (8) boleh ditulis sebagai berikut

$$p_i(y) = \lambda_i c_i y^{c_i-1} \exp(\beta^T Z) \left[-\lambda_i y^{c_i} \exp(\beta^T Z) \right] \quad (9)$$

dengan

$$y > 0, \lambda_i > 0, c_i > 0$$

3. PENTAFSIRAN DAN ANGGARAN

Di sini parameter regresor β dan pemalar bentuk c akan dianggarkan menggunakan kaedah kebolehhjadian maksimum. Pendekatan yang boleh digunakan di antaranya ialah seperti pendekatan Newton-Rhapson, algoritma EM (Dempster *et al.* 1977) dan kompromi Maquardt terubahsuai (David & Moeschberger 1978, Appendix B).

Katakan X_{ij} masa hingga kegagalan bagi suatu individu ke j gagal dari punca R_i ($i=1, \dots, k; j=1, \dots, n_j$). Berdasarkan (7) dan (9) dengan menganggap

$\beta_i^T = (\beta_{i1}, \beta_{i2}, \dots, \beta_{is})$ berbeza bagi setiap R_i , maka fungsi kebolehhjadian dari sesuatu punca ialah

$$L_i = \prod_{j=1}^{n_j} \lambda_i c_i x_{ij}^{c_i-1} \exp\left(\beta_i^T z_{ij}\right) \prod_{l=1, l \neq i}^k \prod_{j=1}^{n_j} \exp\left[-\lambda_l (x_{ij})^{c_l} \exp(\beta_l^T z_{ij})\right] \quad (10)$$

Dengan mengambil log asli bagi (10),

$$\ln L_i = \sum_{j=1}^{n_j} \ln c_i + (c_i - 1) \sum_{j=1}^{n_j} \ln x_{ij} + \sum_{i=1}^{n_i} \ln \lambda_i$$

$$+ \sum_{j=1}^{n_j} \beta_i z_{ij} - \sum_{l=1, l \neq i}^k \sum_{j=1}^{n_j} \lambda_l x_{ij} \exp\left(\beta_l^T z_{ij}\right)$$

$$\sum_{j=1}^{n_j} \beta_i z_{ij} - \sum_{l=1, l \neq i}^k \sum_{j=1}^{n_j} \lambda_l x_{ij}^{c_l} \exp\left(\beta_l^T z_{ij}\right) \quad (11)$$

Untuk keadaan maksimum terhadap λ

$$\frac{\partial \ln L_i}{\partial \lambda_i} = \frac{n_i}{\lambda_i} - \sum_{l=1, l \neq i}^k \sum_{j=1}^{n_j} \lambda_l x_{ij}^{c_l} \exp\left(\beta_l^T z_{ij}\right) = 0$$

maka

$$\hat{\lambda}_i = \sum_{l=1}^k \sum_{j=1}^{n_j} \lambda_l x_{ij}^{c_l} \exp\left(\beta_l^T z_{ij}\right) \quad (12)$$

Dengan

$$g^T = \left(-\frac{\partial \ln L_i}{\partial \ln c_i}, -\frac{\partial \ln L_i}{\partial \ln \beta_i} \right)$$

$$A = - \begin{pmatrix} \frac{1 \ln L_i}{c_i^2} & \frac{1 \ln L_i}{c_i \beta_i} \\ \frac{1 \ln L_i}{\beta_i c_i} & \frac{1 \ln L_i}{\beta_i^2} \end{pmatrix},$$

anggaran parameter peringkat pertama

$\hat{\theta} = (\hat{\lambda}_i, c_i, \hat{\beta}_i)^T$ (ditandakan sebagai $\hat{\theta}_i^{(1)}$) apabila menggunakan pendekatan kompromi Marquardt terubahsuai (David & Moeschberger 1978) dan nilai $\hat{\theta}_i^{(0)}$ awal ialah

$$\hat{\theta}_i^{(1)} - \hat{\theta}_i^{(0)} = -W(WAW + \delta I)^{-1} Wg \quad (13)$$

dengan W matriks m x m penskalaan pepenjur dengan unsur pepenjurunya $|\partial^2 \ln L_i / \partial \hat{\theta}_i^{(2)}|^{-1/2}$ dan δ pemalar bukan negatif. Di sini δI adalah matriks segiempat sama yang dipilih supaya penumpuan dapat dipercepatkan atau untuk mengelakkan keanehan.

4. DIAGNOSTIK BAGI MODEL RISIKO BERSAINGAN DENGAN TAPISAN SEBAGAI SATU RISIKO

Cook (1977) telah mengemukakan satu prosedur bagi mengesan cerapan-cerapan berpengaruh bagi data penuh dalam model linear. Di sini kami akan memanfaatkan saranan Cook tetapi kaedahnya berbeza sedikit kerana tapisan dan penglibatan prosedur berlelaran bagi anggaran parameter.

Kaedah kedua yang akan diperlihatkan ialah dengan menggunakan log kepada fungsi kebolehjadian data lengkap dan data dengan satu cerapan dihapuskan. Ini dikenali sebagai jarak kebolehjadian. (Cook & Weisberg 1982)

4.1. Ukuran berpengaruh

Kaedah jarak Cook melibatkan anggaran bagi parameter $\hat{\theta}_{i(j)}$ dengan penghapusan cerapan ke j daripada set dengan setiap penghapusan perlu diikuti dengan lelaran sehingga penumpuan dan ini melibatkan masa yang sangat lama seperti yang telah diperlihatkan oleh Ibrahim dan Daud (1990). Apa yang boleh dilakukan ialah dengan

menggunakan kaedah satu-langkah. Mungkin dengan menggunakan kaedah ini, anggaran dan ukuran berpengaruh tidak begitu jitu tetapi kejutuan tidak menjadi masalah kerana yang pentingnya ialah kita dapat mengenal pasti sesuatu cerapan yang berlainan dari yang lain untuk pertimbangan seterusnya. Ini juga telah diperakui oleh Cook dan Weisberg (1982). Pregibon (1981) pula menyifatkan penghapusan ini sebagai gangguan terhadap model dan beliau juga telah menggunakan kaedah satu-langkah dalam diagnostik pengaruh bagi model regresi logistik.

Di sini anggaran kebolehjadian maksimum, $\hat{\theta}_i$, digunakan sebagai nilai awal. Berdasarkan (13), kaedah satu langkah kepada kompromi Marquardt terubahsuai adalah seperti berikut:-

$$\hat{\theta}_{i(j)} - \hat{\theta}_i = - [W(WAW) + \delta I]_{(j)}^{-1} Wg_{(j)} \quad (14)$$

yang mana W, A dan g dikira pada $\hat{\theta}_i$.

Jarak Cook bagi satu-langkah ialah

$$D_{i(j)} = (\hat{\theta}_{i(j)} - \hat{\theta}_i)^T M (\hat{\theta}_{i(j)} - \hat{\theta}_i) \quad (15)$$

Di (j) ukuran jarak setelah cerapan ke j dihapuskan dengan memilih matriks $M = (\partial^2 \ln L_i / \partial \hat{\theta}_i^{(2)})$ dan

$$\hat{\theta}_i = (\hat{\lambda}_i, \hat{c}_i, \hat{\beta}_i)^T$$

Satu lagi ukuran yang dipertimbangkan ialah jarak kebolehjadian yang ditakrifkan seperti berikut:-

$$LD_{i(j)} = 2 \left[L(\hat{\theta}_i) - L(\hat{\theta}_{i(j)}) \right] \quad (16)$$

dengan $L(\hat{\theta}_i)$ log asli kepada fungsi kebolehjadian data lengkap bagi punca R_i dan $L(\hat{\theta}_{i(j)})$ fungsi kebolehjadian setelah cerapan ke j dihapuskan.

Seperti perolahan dalam perkembangan di atas, kaedah satu=langkah boleh disarankan dengan menggantikan $L(\hat{\theta}_{i(j)})$ dengan iaitu $L(\hat{\theta}_{i(j)}^1)$ log kebolehjadian pada nilai $\hat{\theta}_{i(j)}^1$ dianggarkan dengan satu-langkah (Bin Daud 1987). Jika kontur log kebolehjadian menghampiri elips,

$$LD_{i(j)} \approx (\hat{\theta}_{i(j)} - \hat{\theta}_i)^T M (\hat{\theta}_{i(j)} - \hat{\theta}_i) \quad (17)$$

Jika kontur bagi fungsi kebolehjadian tidak elips, Cook dan Weisberg (1982), berpendapat

ukuran jarak kebolehhjadian mungkin memberi keputusan yang agak menyeleweng daripada yang dijangkakan.

5. CONTOH

Dua contoh dipertimbangkan isitu ujian klinikal untuk rawatan barah orofarinks dan data pemindahan jantung Stanford. Kedua-dua data terdapat di Appendix 1 daripada Kalbfleisch dan Prentice (1980). Model yang digunakan dalam kedua-dua data ini menganggap kegagalan yang sebenar sebagai risiko pertama dan tapisan sebagai risiko kedua. Anggaran kebolehhjadian maksimum yang berasingan dilakukan ke atas setiap risiko bagi setiap data dan diagnostik berpengaruh juga dikira secara berasingan bagi setiap risiko.

Hanya tujuh kovariat diambil kira bagi data barah orofarinks dan tiga kovariat bagi data pemindahan jantung.

5.1. Ujian klinikal bagi barah orofarinks

Analisis bagi data ini melibatkan 195 kes dengan bilangan kematian, $n_1=142$ dan bilangan tapisan $n_3=53$. Tujuh kovariat dipertimbangkan iaitu umur, jantina, rawatan, kesihatan pesakit, gred, peringkat 1 dan peringkat 2. Umur ialah umur pesakit semasa diagnos dan diukur dalam ukuran tahun. Jantina ialah lelaki (=1) dan wanita (=2). Terdapat dua jenis rawatan iaitu, rawatan terapi beradiasi (=1) dan rawatan terapi beradiasi dengan agen kemoterapeutik (=2). Keadaan yang mengambarkan kesihatan pesakit terbahagi kepada empat kategori semasa menjalani rawatan iaitu, tiada kekurangan kebolehan (=1), yang hanya boleh dilakukan kerja yang tertentu sahaja (=2), yang memerlukan bantuan untuk kebersihan diri (=3) dan yang tidak berdaya (terlantar di atas katil), (=4). Gred ialah perbandingan darjah sel ketumbuhan menyerupai sel hos iaitu, jelas (=1), kurang jelas (=2) dan tidak menentu (=3). Peringkat 1 mengukur saiz ketumbuhan iaitu, ukuran 2 cm atau kurang (=1), ukuran 2 cm hingga 4 cm (=2), ukuran lebih dari 4 cm (=3) dan sangat besar serta agresif (=4). Peringkat 2 pula mengukur wujudnya penglibatan nod limpatek iaitu, tiada bukti klinikal bagi kerebakan nod (=0), satu nod positif yang berlegar dan diameternya berukuran tidak lebih dari 3 cm (=1), satu nod positif yang berlegar dan diameternya lebih dari 3 cm (=2) dan beberapa nod positif yang tidak berlegar (=3).

Model yang telah digunakan ialah model risiko berasingan dengan taburan pendasar bagi masa

kegagalan diandaikan sebagai Weibull, $k=2$ dan pemalar bentuknya berlainan bagi setiap risiko.

Dengan nilai awal $c_1 = 1$ dan $\beta_1 = 0$ kaedah kompromi Marquardt terubahsuai berleleran memberikan anggaran kebolehhjadian maksimum kepada lima titik perpuluhan seperti berikut:-

$$\hat{\lambda}_1 = 0.08508, \hat{c}_1 = 1.16552, \hat{\beta}_{11} = 0.00406, \hat{\beta}_{12} = -0.19071, \hat{\beta}_{13} = 0.17200, \hat{\beta}_{14} = -0.10148, \hat{\beta}_{15} = 0.25932, \hat{\beta}_{16} = 0.31154 \text{ dan } \hat{\beta}_{17} = 0.13882$$

Bagi cerapan tertapis dengan nilai awal $c_2 = 1$ dan $\beta_2 = 0$ anggaran kebolehhjadian maksimum adalah seperti berikut:-

$$\hat{\lambda}_2 = 0.01875, \hat{c}_2 = 3.191124, \hat{\beta}_{21} = -0.01417, \hat{\beta}_{22} = -0.020541, \hat{\beta}_{23} = 0.14541, \hat{\beta}_{24} = 0.004646, \hat{\beta}_{25} = 0.38582, \hat{\beta}_{26} = 0.06303$$

dan $\hat{\beta}_{27} = 0.12080$ serta log kebolehhjadian maksimum $= -93.37044$. Anggaran sisihan piawai asimptot bagi punca pertama dan kedua adalah

$$s(\hat{\lambda}_1) = 0.06814 \quad s(\hat{\lambda}_2) = 0.02278$$

$$s(\hat{c}_1) = 0.07143 \quad s(\hat{c}_2) = 0.13007$$

$$s(\hat{\beta}_{11}) = 0.00832 \quad s(\hat{\beta}_{21}) = 0.01478$$

$$s(\hat{\beta}_{12}) = 0.20268 \quad s(\hat{\beta}_{22}) = 0.32575$$

$$s(\hat{\beta}_{13}) = 0.17111 \quad s(\hat{\beta}_{23}) = 0.29921$$

$$s(\hat{\beta}_{14}) = 0.12027 \quad s(\hat{\beta}_{24}) = 0.20804$$

$$s(\hat{\beta}_{15}) = 0.07207 \quad s(\hat{\beta}_{25}) = 0.28590$$

$$s(\hat{\beta}_{16}) = 0.11647 \quad s(\hat{\beta}_{26}) = 0.19756$$

$$s(\hat{\beta}_{17}) = 0.07654 \quad s(\hat{\beta}_{27}) = 0.14253$$

Bagi $i = 1$, jarak kebolehhjadian bagi setiap cerapan adalah lebih besar daripada jarak Cook (*Rajah 1* dan *Rajah 2*). Didapati cerapan ke-159 memberikan nilai jarak Cook dan jarak kebolehhjadian yang paling besar berbanding dengan cerapan-cerapan yang lain dan cerapan ini boleh dianggap sebagai titik terpencil.

Bagi $i = 2$, daripada *Rajah 3* iaitu graf bagi jarak Cook, jelas kelihatan cerapan-cerapan 6, 43, 46, 53 dan 176 cerapan yang berpengaruh yang mungkin. Tetapi graf jarak kebolehhjadian (*Rajah 4*) menunjukkan cerapan-cerapan yang lain telah disebut tadi, cerapan ke-186 dan 190 adalah juga cerapan berpengaruh. Jarak kebolehhjadian memberikan gambaran yang tidak begitu jelas bagi $i = 2$,

dan ini mungkin disebabkan bilangan cerapan tertapis adalah kecil pada keseluruhannya. Oleh itu bagi keadaan yang begini lebih baik berpandukan ukuran jarak Cook untuk mengesan cerapan berpengaruh.

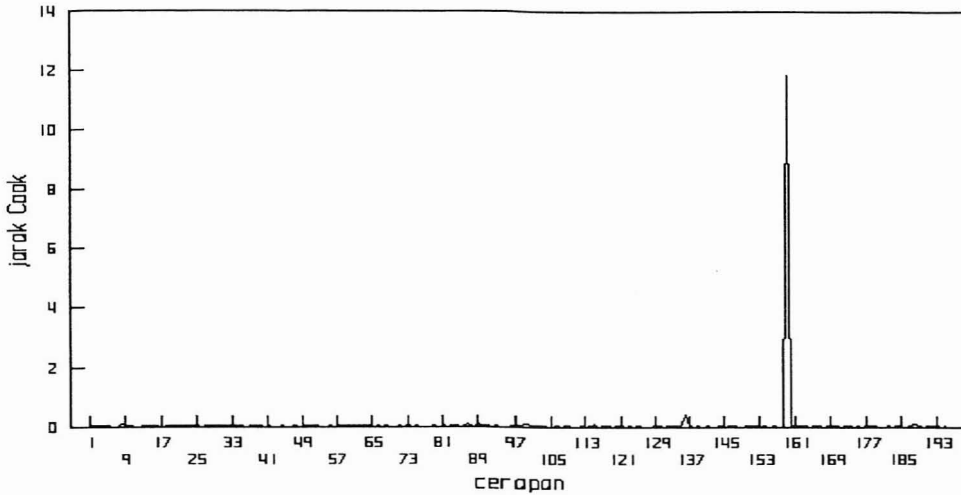
5.2. Data pemindahan jantung Stanford

Data ini melibatkan 90 kes, $n_1 = 62, n_2 = 28$ dan 3 kovariat iaitu umur, pembedahan dan pemindahan. Umur ialah umur seseorang itu diterima ketika menjalani program Stanford ini. Kovariat

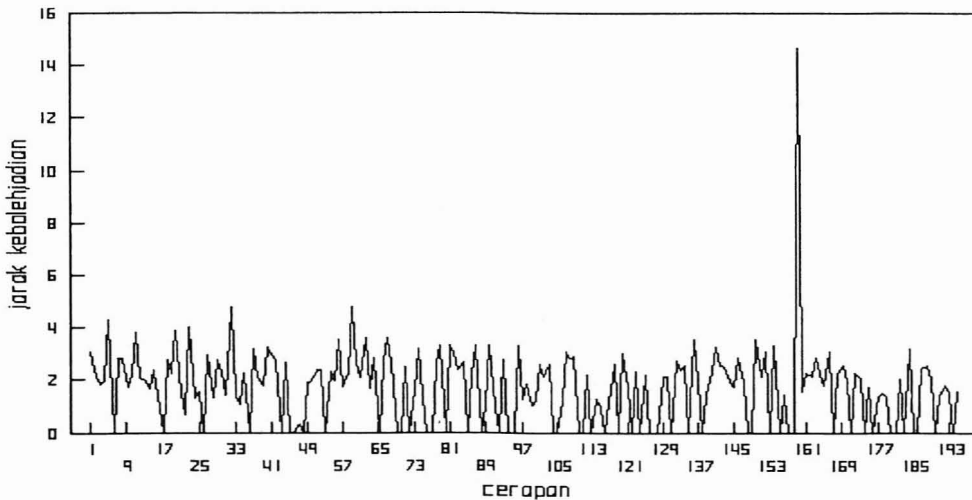
pembedahan pula bermaksud pembedahan yang pernah dilakukan (=1) dan tidak pernah dilakukan (=0) sebelum mengikuti program ini. Pemindahan bermaksud pemindahan dilakukan (=1) dan tiada berlaku pemindahan (=0) semasa mengikuti program.

Dengan nilai awal $c_1 = 1$ dan $\beta_1 = 0$, pendekatan komprom Marquardt terubahsuai bagi anggaran maksimum adalah seperti berikut:-

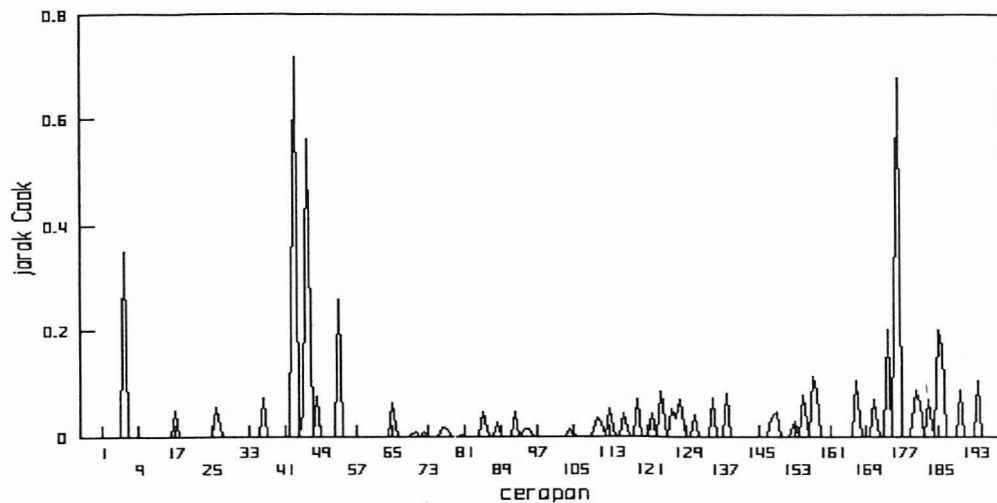
$$\hat{\lambda}_1 = 0.23933, \hat{c}_1 = 0.76289, \hat{\beta}_{11} = 0.05587, \hat{\beta}_{11} = -0.8569$$



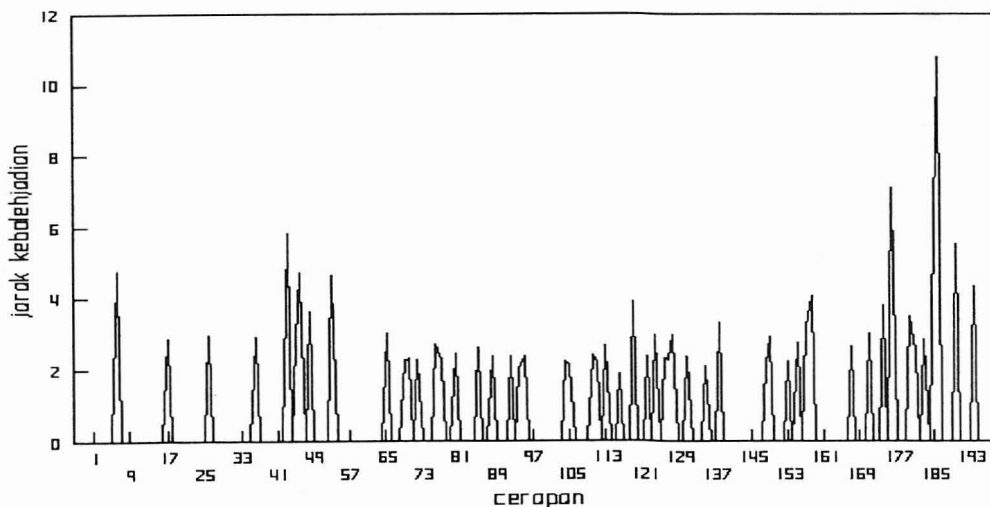
Rajah 1: Rajah Cook bagi data barah orofarinks risiko pertama



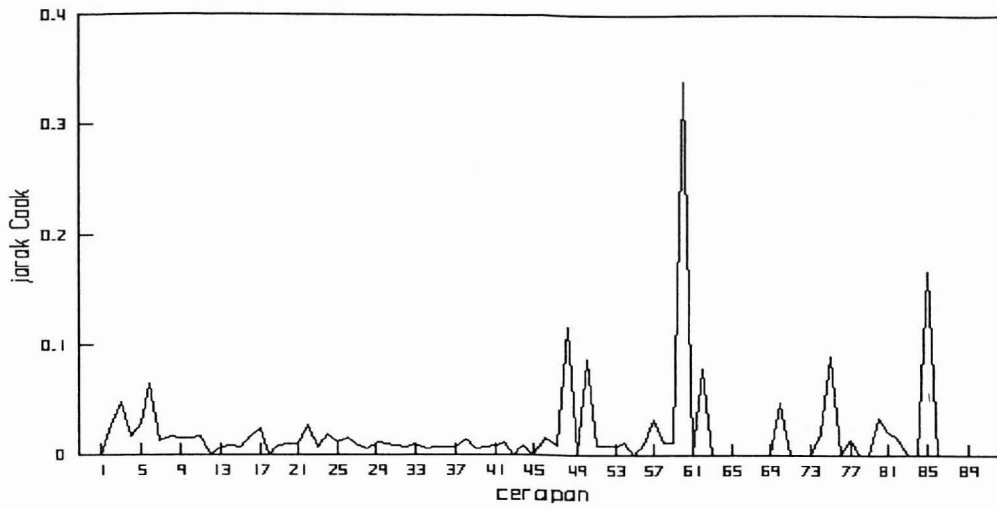
Rajah 2: Jarak kebolehdajian bagi data orofarinks risiko pertama



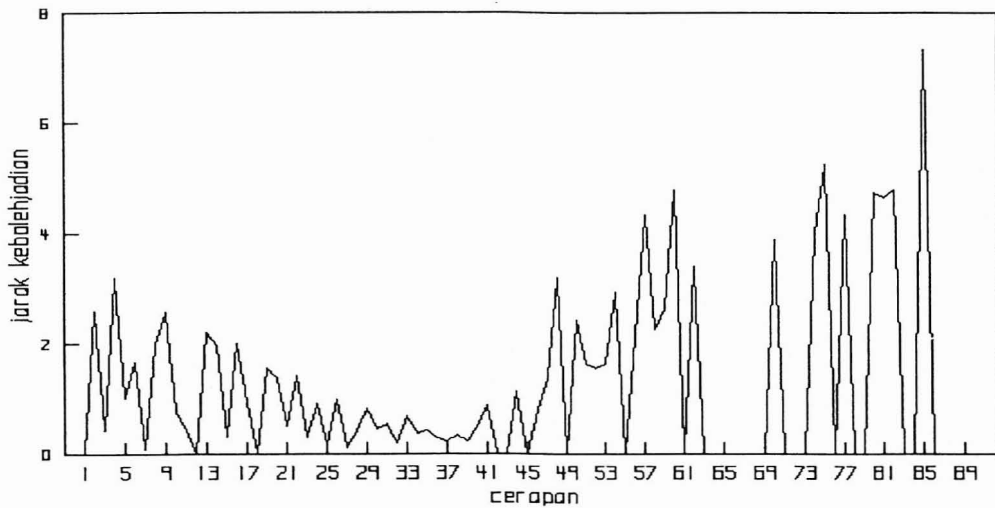
Rajah 3: Jarak Cook bagi data barah orofarinks risiko kedua



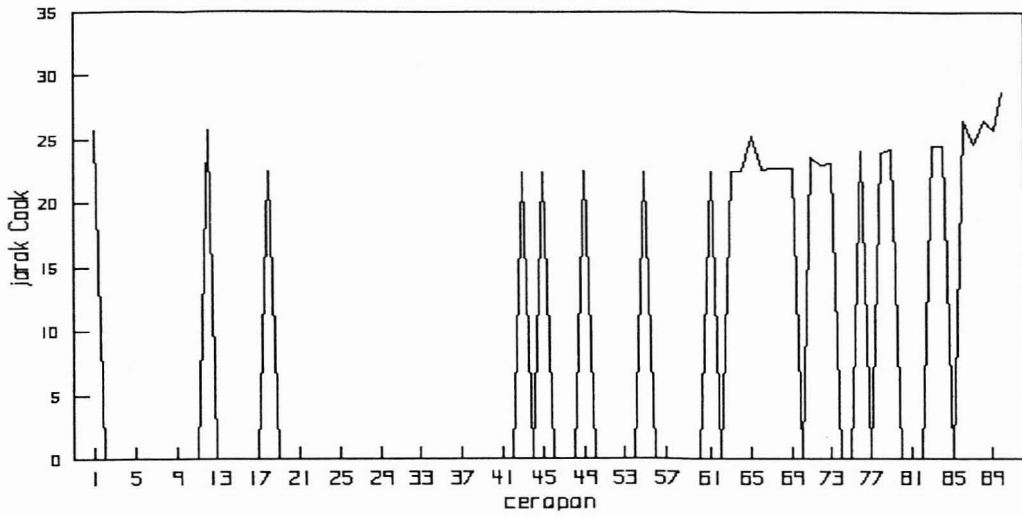
Rajah 4: Jarak akebolehdajian bagi data orofarinks risiko kedua



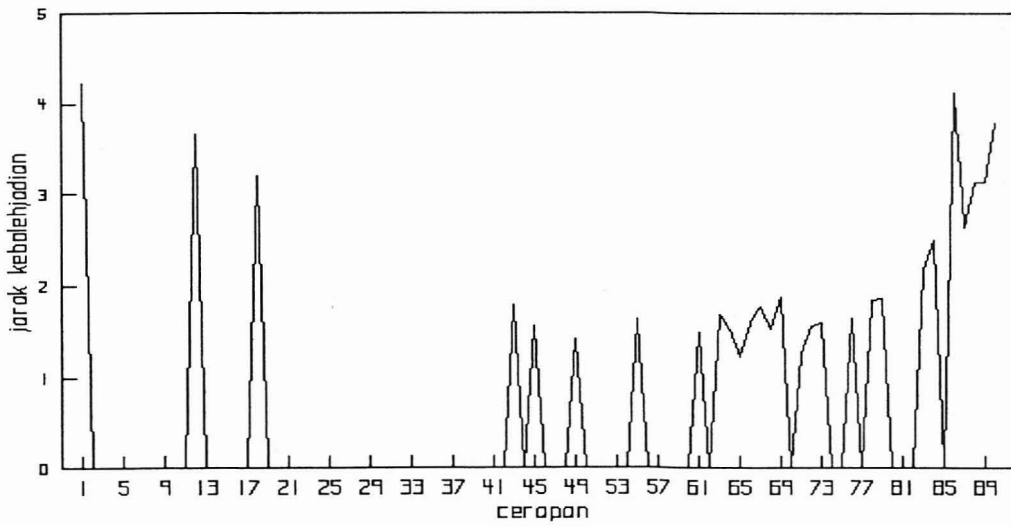
Rajah 5: Jarak Cook bagi data barah Stanford risiko pertama



Rajah 6: Jarak kebolehdajian bagi data Stanford risiko pertama



Rajah 7: Jarak Cook bagi data barah Stanford risiko kedua



Rajah 8: Jarak kebolehhadion bagi data Stanford risiko kedua

dan $\hat{\beta}_{13} = -1.42677$ serta log kebolehdajian = -56.5087. Bagi cerapan tertapis dengan nilai awal $c_2 = 1$ dan $\beta_2 = 0$, anggaran kebolehdajian maksimum adalah seperti berikut:-

$\hat{\lambda}_2 = 0.74576 = \hat{c}_2 = 1.38135$, $\hat{\beta}_{21} = -0.02263$, $\hat{\beta}_{22} = -0.19567$, $\hat{\beta}_{23} = -0.15521$ serta log kebolehdajian = -56.19152. Anggaran sisihan piawai asimptot bagi punca pertama dan kedua adalah:-

$s(\hat{\lambda}_1) = 0.16941$ $s(\hat{\lambda}_2) = 0.65654$
 $s(\hat{c}_1) = 0.08743$ $s(\hat{c}_2) = 0.15330$
 $s(\hat{\beta}_{11}) = 0.01666$ $s(\hat{\beta}_{21}) = 0.02402$
 $s(\hat{\beta}_{12}) = 0.44220$ $s(\hat{\beta}_{22}) = 0.47688$
 $s(\hat{\beta}_{13}) = 0.33011$ $s(\hat{\beta}_{23}) = 0.63287$

Bagi $i = 1$, daripada *Rajah 5*, jarak Cook memberikan gambaran yang jelas bahawa cerapan ke-60 dan 85 adalah titik-titik berpengaruh. Tetapi daripada *Rajah 6*, yang jelas kelihatan ialah cerapan ke-85 sahaja dan boleh dianggap sebagai berpengaruh. Kelainan ini mungkin disebabkan apabila pemindahan beralaku pesakit akan mengikuti rawatan yang berlainan dan kovariat jenis ini mestilah diteliti dahulu sebelum model ditentukan.

Bagi $i = 2$, jarak Cook (*Rajah 7*) dan jarak kebolehdajian (*Rajah 8*) tidak dapat menunjukkan dengan jelas cerapan-cerapan yang mungkin berpengaruh. Terdapat sedikit gangguan pada kedua-dua graf dan ini berlaku mungkin disebabkan kovariat yang bersandaran kepada masa dan pengaruh cerapan kegagalan terhadap keseluruhan data.

6. KESIMPULAN

Analisis diagnostik untuk masa kegagalan yang sebenar dalam model risiko bersaing seperti contoh barah orfarinks menunjukkan kedua-dua ukuran iaitu $D_{i(q)}$ dan $LD_{i(q)}$ berjaya mengesan cerapan yang sama. Tetapi ini tidak berlaku ke atas data Stanford bagi $i = 1$, walau-pun terdapat satu cerapan yang dikesan oleh kedua-dua kaedah. Kelainan ini mungkin disebabkan oleh cirian yang ada pada data itu sendiri.

Bagi data tertapis pula beberapa kesimpulan boleh diperolehi daripada penyelidikan ini. Masa tapisan itu sendiri boleh menyebabkan kelainan cerapan yang dikesan oleh $D_{i(q)}$ dan $Ld_{i(q)}$. Ini boleh disebabkan oleh pengaruh masa kegagalan sebenar ke atas keseluruhan data atau disebabkan taburan

yang bukan kuadratik yang memberikan kontur kebolehdajian yang tidak elips dan ukuran yang diberi agak menyeleweng. Dalam hal ini lebih baik berpandukan jarak Cook bagi mengesan cerapan berpengaruh. Bilangan cerapan juga memainkan peranan untuk menentukan keberkesanan kaedah kebolehdajian dalam mengesan cerapan berpengaruh. Didapati jika n_2 adalah besar atau sama dengan n_1 , kaedah kebolehdajian lebih berkesimbangan dengan jarak Cook. Keputusan memperlihatkan cerapan tertapis juga memberikan kesan ke atas anggaran parameter dan perlu diteliti apabila penyesuaian model dilakukan ke atasnya.

RUJUKAN

- BELSLEY, D.A., E. KUH and R.E. WELSCH. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- BIN DAUD, I. 1987. Influence Diagnostics in Regression with Censored Data. Ph. D. Thesis, Loughborough University.
- COOK, R. D. 1977. Detection of Influential Observations in Linear Regression. *Technometrics* **19**: 15-18.
- COOK, R. D. and S. WEISBERG. 1980. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics* **22**: 495-508.
- COOK, R. D. and S. WEISBERG. 1982. Residuals and Influence in Regression. London: Chapman and Hall.
- COX, D. R. 1972. Regression Models and Life Tables (with discussion). *J. R. Statist. Soc.* **B34**: 187-220.
- COX, D. R. and D. OAKES. 1984. *Analysis of Survival Data*. London: Chapman and Hall.
- DAVID, H. A. and M. L. MOESCHBERGER. 1978. *The Theory of Competing Risks*. London: Griffin.
- DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN. 1977. Maximum Likelihood from Incomplete Data via Em Algorithm. *J. R. Statist. Soc.* **B39**: 1-38.
- GAIL, M. 1975. A Review and Critique of some Models Used in Competing Risks Analysis. *Biometrics* **31**: 209-222.
- IBRAHIM, N. A. and I. DAUD 1990. Analisis Regresi Mudah dan Mengesan Cerapan Berpengaruh bagi Data Tertapis. Laporan Teknikal, 45/1990, Jabatan Matematik, Universiti Pertanian Malaysia.
- KALBFLEISCH, J. D. and R. L. PRENTICE. 1980. *The Statistical Analysis of failure Time Data*. New York: Wiley.
- KAPLAN, E. L. and P. MEIER. 1958. Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc.* **53**: 457-481.

- PREGIBON, D. 1981. Logistics Regression Diagnostics. *Ann. Statist.* **9**: 705-724.
- SCHOENFELD, D. 1982. Partial Residuals for the Proportional Hazards Regression Model. *Biometrika* **69**(1): 239-241.
- WEISSFELD, L. A. and H. SCHNEIDER. 1990. Influence Diagnostics for the Normal Linear Model with Censored Data. *Austral. J. Statistics* **32**(1): 11-20.

(Di terima 24Ogos 1991)