Statistika, Vol. 9 No. 2, 123 – 135
Nopember 2009

Pemeriksaan Data Berpengaruh dalam Model Gamma

NUSAR HAJARISMAN

Mahasiswa Sekolah Pascasarjana Institut Pertanian Bogor
nrisman@yahoo.co.uk

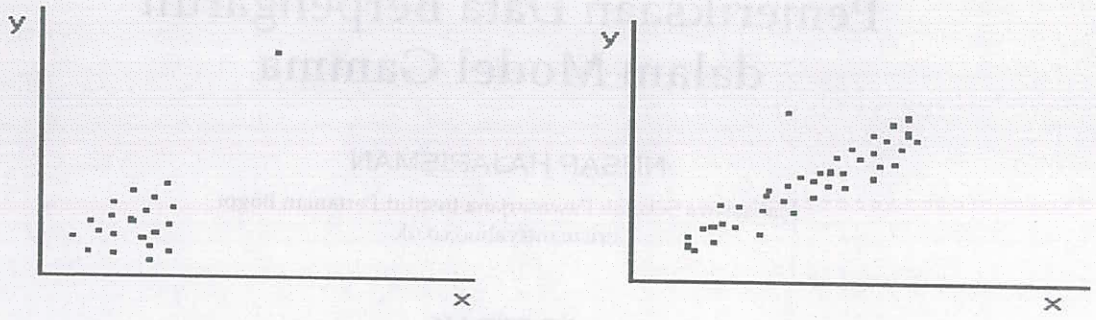
ABSTRAK

Dalam pemodelan statistika, khususnya dalam pemodelan data kategorik, ada sejumlah cara dimana model dugaan tidak layak. Salah satu diantaranya adalah data mungkin berisi suatu data pencilan yang berpotensi menjadi data berpengaruh sehingga mengakibatkan data tidak cocok terhadap model dugaan. Teknik yang digunakan untuk pemeriksaan data berpengaruh ini disebut juga sebagai proses diagnosa. Pada makalah ini pembahasan akan lebih difokuskan pada pemeriksaan data berpengaruh dalam pemodelan yang responsnya mengikuti distribusi gamma. Beberapa ukuran statistik yang digunakan untuk memeriksa data pencilan adalah nilai leverage, residu devians dibakukan, residu Pearson dibakukan, dan residu likelihood. Kemudian data pencilan yang berpotensi sebagai data berpengaruh akan diperiksa dengan menggunakan statistik Cook's distance.
Kata Kunci: distribusi gamma, nilai leverge, residu devians dibakukan, residu Pearson dibakukan, residu likelihood, statistik Cook's distance.

1. PENDAHULUAN

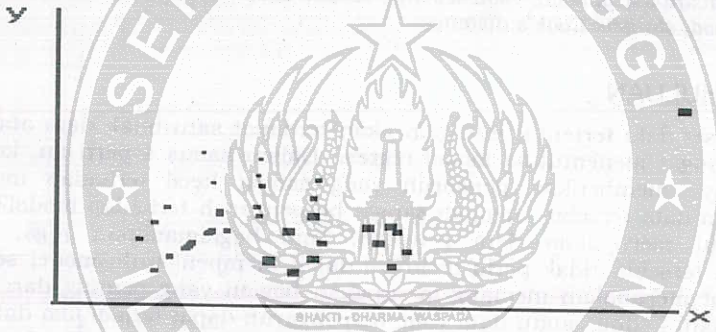
Dalam satu set data tertentu, mungkin akan terdapat satu buah data atau sekelompok kecil data yang sangat menentukan model regresi. Dalam kasus seperti ini, kelompok data besar lainnya hanya memberikan pengaruh yang sangat kecil terhadap model. Apakah yang menyebabkan data tersebut menjadi sangat berpengaruh terhadap model? Pertama, mungkin data tersebut merupakan data pencilan. Tapi bagaimanapun juga, semua data yang berpengaruh tersebut tidak perlu dicurigai dapat mempengaruhi model sepenuhnya. Padahal data tersebut merupakan memang merupakan bagian yang penting dari satu set data yang sedang diamati. Kedua, suatu data yang berpengaruh dapat terjadi jika data tersebut jaraknya jauh dari kumpulan data lainnya. Walaupun data itu benar, maka data itu bukan berarti merupakan gambaran dari model yang keliru. Sebagai contoh, perhatikan Gambar 1(a) untuk kasus pada satu peubah. Nilai leverage yang besar akan menentukan slope regresi sepenuhnya oleh titik data tersebut. Tapi titik data tersebut bukan merupakan data pencilan yang menyebabkan model menjadi keliru. Di lain pihak, Gambar 1(b) menunjukkan bahwa titik data tersebut berada di luar trend.

Pada Gambar 1 menunjukkan apa yang mungkin terjadi di lapangan. Dalam kasus pada Gambar 1(a), maka dapat diatasi dapat dilakukan dengan cara menambah data sehingga dapat mengisi celah yang kosong tersebut. Sedangkan apabila kita mempunyai informasi yang tidak lengkap mengenai data tersebut, maka suatu data yang berpengaruh harus diperiksa secara hati-hati. Selanjutnya, untuk Gambar 1(b) yang merupakan data pencilan, maka pemeriksaannya dapat dilakukan melalui analisis residu dan nilai leverage yang nanti akan dibahas pada bagian berikutnya.



Gambar 1. Plot antara y dan x

Dalam pemeriksaan data berpengaruh ini, akan sangat berhubungan dengan pemeriksaan data pencilan. Kedua konsep tersebut, baik itu data pencilan maupun nilai laverage, menggambarkan suatu kondisi yang tidak biasa dalam suatu pengamatan. Pengamatan xi yang mempunyai nilai leverage yang besar (mendekati satu) akan berada jauh dari kumpulan data yang lainnya. Tapi tidak semua data yang mempunyai nilai leverage yang besar itu merupakan data yang berpengaruh, serta tidak semua data pencilan itu juga merupakan data yang berpengaruh, sehingga dalam hal ini perlu dilakukan pemeriksaan secara lebih teliti.



Gambar 2 Diagonal HAT yang besar tapi bukan data berpengaruh

Lalu, apakah penting kita melakukan pemeriksaan terhadap data yang berpengaruh tersebut? Jelas bahwa nilai laverage dari suatu titik pengamatan akan mengakibatkan model menjadi kurang baik. Perhatikan Gambar 2. Dalam hal ini, jelas bahwa titik B merupakan data yang berpengaruh karena jika kita pindahkan titik data tersebut akan dapat menghasilkan perubahan yang besar pada slope regresi. Sedangkan pada titik A perubahan yang dihasilkannya tidak terlalu besar. Jadi, suatu data yang berpengaruh akan menghasilkan perubahan pada slope maupun intersep dari model regresi sehingga model regresi itu menjadi kurang baik. Menurut Myers (1990), dalam pemeriksaan data berpengaruh ini ada beberapa hal yang perlu diperhatikan, yaitu:

- Tidak semua data pencilan merupakan data yang berpengaruh (tergantung pada nilai leverage).
- Tidak semua yang mempunyai nilai laverage yang besar merupakan data yang berpengaruh [lihat Gambar 1(a)].
- Tidak semua data yang berpengaruh merupakan data pencilan.

Pada makalah ini pembahasan akan lebih difokuskan pada pemeriksaan data berpengaruh dalam pemodelan yang responsnya mengikuti distribusi gamma. Pada bagian dua akan dibahas terlebih dahulu mengenai model regresi gamma. Kemudian pada bagian tiga dibahas mengenai beberapa ukuran statistik yang digunakan untuk memeriksa data pencilan adalah nilai leverage, residu devians dibakukan, residu Pearson dibakukan, dan residu likelihood.

Kemudian data pencilan yang berpotensi sebagai data berpengaruh akan diperiksa dengan menggunakan statistik Cook's *distance*.

2. MODEL REGRESI GAMMA

Misalkan diamati suatu variabel respons y_i untuk n buah pengamatan. Asumsi dasar yang diperlukan dalam model gamma ini adalah

$$\text{var}(y_i) = \sigma^2 \left[E(y_i) \right]^2, \qquad \text{untuk } i = 1,...,n \tag{1}$$

yaitu, koefisien variasi pengamatannya merupakan suatu konstanta dan koefisien variasi umum dinyatakan dengan σ^2. Apabila nilai yang mungkin dari variabel respons berupa bilangan nyata positif dan apabila respons tersebut berasal dari distribusi gamma, maka akan diperoleh bentuk khusus dimana $\sigma^2 = 1/\nu$ dan ν merupakan parameter bentuk (shape parameter).

Untuk unit pengamatan ke-i, dimisalkan bahwa

$$E(y_i) = \mu_i, \qquad \text{untuk } i = 1,...,n$$

Pada umum rata-rata dari unit pengamatan ke-i dimisalkan bergantung pada nilai-nilai $(x_{i1},...,x_{id})$ dari variabel penjelas yang dihubungkan dengan unit pengamatan ke-i, yaitu:

$$E(y_i) = \mu_i = \mu(x_{i1},...,x_{ip}), \qquad \text{untuk } i = 1,...,n \tag{2}$$

dimana $\mu(x_{i1},...,x_{ip})$ merupakan fungsi dari segugus variabel penjelas.

2.1 Distribusi Gamma

Fungsi pembangkit moment dari model Gamma(ν, μ) mempunyai bentuk sebagai berikut:

$$M(\xi;\nu,\mu) = \left(1 - \frac{\xi\mu}{\nu}\right)^{-\nu} \tag{3}$$

dan fungsi pembangkit kumulant diberikan oleh

$$K(\xi) = \ln\left[M(\xi)\right] = -\nu \ln\left(1 - \frac{\xi\mu}{\nu}\right) \tag{4}$$

Kemudian, moment ke-k diberikan oleh

$$m_k = \frac{\mu^k (1+\nu)(2+\nu)\cdots(k-1+\nu)}{\nu^{k-1}}, \qquad \text{untuk } k = 1,2,... \tag{5}$$

dan kumulant ke-k diberikan oleh

$$m_k = \frac{(k-1)!\mu^k}{\nu^{k-1}}, \qquad \text{untuk } k = 1,2,... \tag{6}$$

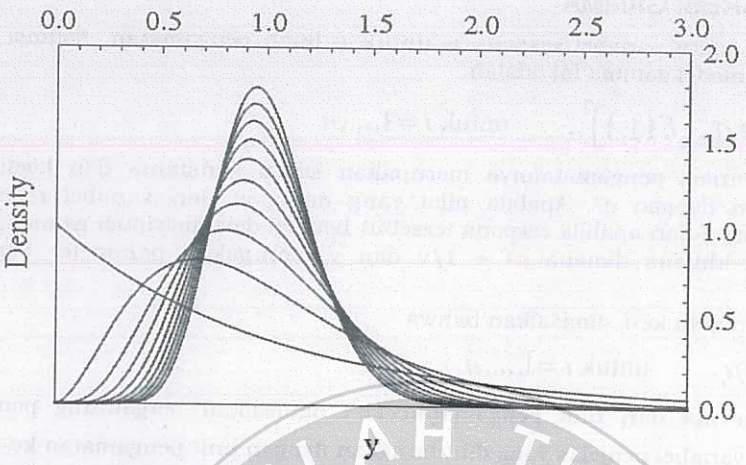
Jadi, empat kumulant pertama dari model Gamma(ν, μ) adalah

$$\kappa_1 = \mu \tag{7}$$

$$\kappa_2 = \frac{\mu^2}{\nu} \tag{8}$$

$$\kappa_3 = \frac{2\mu^3}{\nu^2} \tag{9}$$

$$\kappa_4 = \frac{6\mu^4}{\nu^3} \tag{10}$$



Gambar 3. Fungsi Densitas dari distribusi gamma dengan rata-rata 1 dan parameter bentuk 1, 3, ..., 19

Kumulant dari variabel yang dibakukan

$$z = \frac{\sqrt{\nu}(y-\mu)}{\mu}$$

diperoleh perluasan deret Taylor

$$\frac{\xi}{2} + \frac{\xi^3}{3\nu^{1/2}} + \frac{\xi^4}{4\nu} + \frac{\xi^5}{3\nu^{3/2}} + \frac{\xi^6}{6\nu^2} + O[\xi]^7$$

Yang diurutkan sebagai 0, 1, $O\left(\nu^{-1/2}\right)$, $O\left(\nu^{-1}\right)$, dan seterusnya. Pada saat $r \geq 2$, maka kumulant ke-r dari variabel dibakukan z adalah urutan $O\left(\nu^{(1-r)/2}\right)$. Kumulant dari variabel Z mendekati 0, 1, 0, 0, ... dari distribusi normal dibakukan untuk $\nu \to \infty$. Oleh karena konvergen dari kumulant juga berarti konvergen dalam distribusi, maka peluang pendekatannya dapat diperoleh melalui rumusan

$$P(Y \leq y) \approx \Phi\left(\frac{y-\mu}{\mu/\sqrt{\nu}}\right)$$

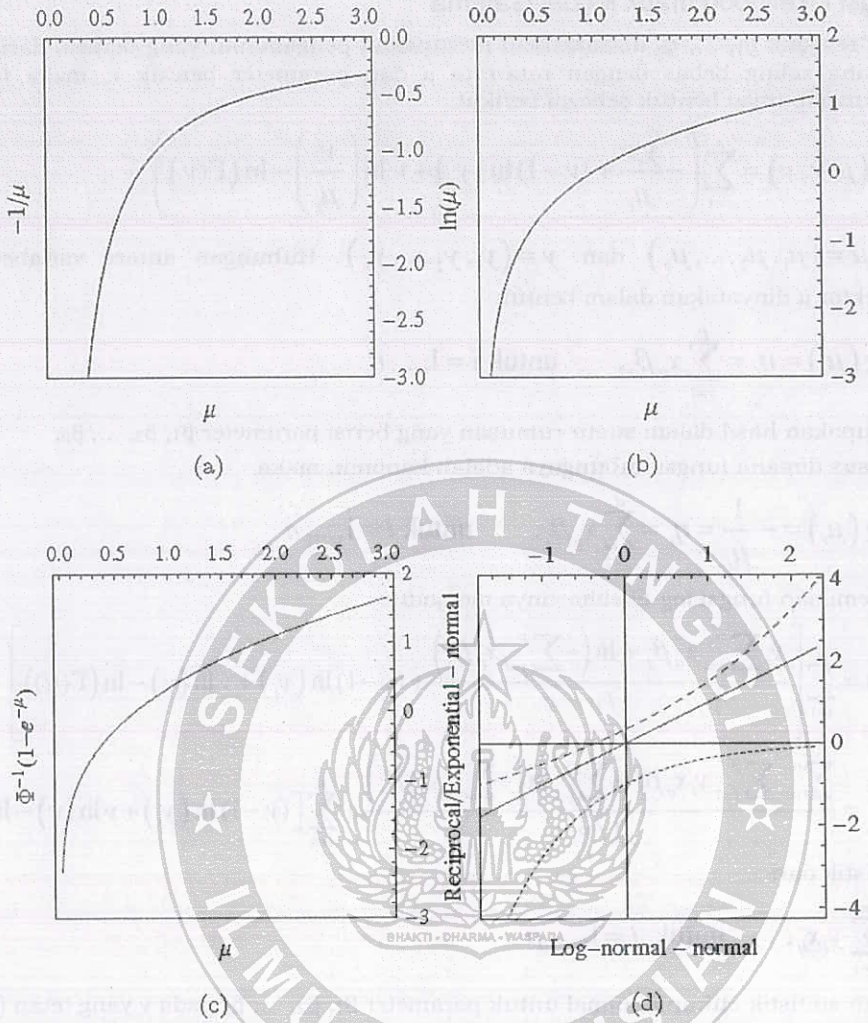
dimana Φ merupakan fungsi distribusi kumulatif dari distribusi normal baku. Gambar 3 menunjukkan grafik dari densitas gamma dengan rata-rata satu dan berbagai nilai dari parameter bentuk ν.

2.2 Fungsi Hubung

Fungsi hubungan yang biasa digunakan dalam model gamma adalah fungsi resiprokal, yaitu:

$$g(\mu) = -\frac{1}{\mu} \tag{11}$$

Fungsi hubung ini merupakan fungsi hubung kanonik. Fungsi hubung resiprokal digunakan pada saat prediktor linear dibatasi hanya pada suatu nilai negatif.



Gambar 4. Plot fungsi hubung untuk distribusi gamma: (a) fungsi hubung resiprokal, (b) fungsi hubung log, (c) fungsi hubung eksponensial-normal, dan (d) plot parametrik dari fungsi hubung

Misalkan diberikan dua buah distribusi dengan fungsi distribusi kumulatif F_1 dan F_2 sedemikian rupa sehingga distribusi yang pertama hanya mempunyai nilai positif dan distribusi yang kedua mempunyai sembarang bilangan nyata, maka fungsi

$$g(\mu) = -F_2\left[F_1(\mu)\right] \tag{12}$$

merupakan fungsi hubung lainnya yang mungkin dapat dibentuk.

Dalam hal fungsi hubung log, maka distribusinya adalah log-normal dibakukan dan distribusi normal sebab

$$F_2^{-1}\left[F_1(\mu)\right] = \Phi^{-1}\left[\Phi\left(\ln(\mu)\right)\right] = \ln(\mu)$$

Sebagai contoh, misalkan diambil F_1 sebagai fungsi distribusi kumulatif dari model eksponensial dengan paraneter μ dan F_2 merupakan distribusi normal baku, maka diperoleh

$$F_2^{-1}\left[F_1(\mu)\right] = \Phi^{-1}\left(1-e^{-\mu}\right)$$

Gambar 4 menampilkan grafik dari berbagai fungsi hubung di atas secara terpisah dan digabungkan bersama.

2.3 Fungsi Likelihood untuk Model Gamma

Pada saat respons $y_1, \ldots, y_n$ diasumsikan merupakan pengamatan yang berasal dari distribusi gamma yang saling bebas dengan rata-rata μ dan parameter bentuk ν, maka fungsi log-likelihood mempunyai bentuk sebagai berikut:

$$l(\mu,\nu;y)=\sum_{i=1}^{n}\left(-\frac{y_i\nu}{\mu_i}+(\nu-1)\ln(y_i)+\nu\ln\left(\frac{\nu}{\mu_i}\right)-\ln(\Gamma(\nu))\right) \tag{13}$$

dimana $\mu=(\mu_1,\mu_2,\ldots,\mu_n)$ dan $y=(y_1,y_2,\ldots,y_n)$. Hubungan antara variabel penjelas dengan vektor μ dinyatakan dalam bentuk

$$g(\mu_i)=\eta_i=\sum_{j=1}^{p}x_{ij}\beta_j, \qquad \text{untuk } i=1,\ldots,n$$

Yang merupakan hasil dalam suatu rumusan yang berisi parameter $\beta_1, \beta_2, \ldots, \beta_p$.
Dalam kasus dimana fungsi hubungnya adalah kanonik, maka

$$g(\mu_i)=-\frac{1}{\mu_i}=\eta_i=\sum_{j=1}^{d}x_{ij}\beta_j, \qquad \text{untuk } i=1,\ldots,n$$

Dengan demikian fungsi log-likelihoodnya menjadi

$$l(\beta,\nu;y)=\sum_{i=1}^{n}\left[\frac{y_i\sum_{j=1}^{p}x_{ij}\beta_j+\ln\left(-\sum_{j=1}^{p}x_{ij}\beta_j\right)}{1/\nu}+(\nu-1)\ln(y_i)+\nu\ln(\nu)-\ln(\Gamma(\nu))\right]$$

$$=\frac{\sum_{i=1}^{n}\sum_{j=1}^{p}y_ix_{ij}\beta_j+\sum_{i=1}^{n}\ln\left(-\sum_{j=1}^{p}x_{ij}\beta_j\right)}{1/\nu}+\sum_{i=1}^{n}\left[(\nu-1)\ln(y_i)+\nu\ln(\nu)-\ln(\Gamma(\nu))\right]$$

Jadi, statistik dari

$$\sum_{i=1}^{n}y_ix_{ij}, \qquad \text{untuk } j=1,\ldots,p$$

merupakan statistik cukup minimal untuk parameter $\beta_1, \beta_2, \ldots, \beta_p$ pada ν yang tetap (*fixed*).
Fungsi likelihood untuk model dugaan untuk model gamma dengan parameter bentuk ν tetap dapat dinyatakan dalam bentuk

$$l(\hat{\mu},\nu;y)=\sum_{i=1}^{n}\left(-\frac{y_i\nu}{\hat{\mu}_i}+(\nu-1)\ln(y_i)+\nu\ln\left(\frac{\nu}{\hat{\mu}_i}\right)-\ln(\Gamma(\nu))\right)$$

dimana nilai $\tilde{\mu}_i=y_i$ akan memberikan nilai likelihood yang paling besar. Dengan demikian, fungsi deviansnya akan menjadi

$$D(y;\nu,\hat{\mu})=2\{l(\nu,\tilde{\mu};y)-l(\nu,\hat{\mu};y)\}$$

$$=2\nu\sum_{i=1}^{n}\left(\ln\left(\frac{\hat{\mu}_i}{y_i}\right)-\frac{(y_i-\hat{\mu}_i)}{\hat{\mu}_i}\right) \tag{14}$$

Distribusi asimtotik dari devians $D(y;\nu,\hat{\mu})$ adalah distribusi χ^2 dengan derajat bebas sama dengan $(n-p)$.

2.4 Pendugaan Parameter

Dikatehaui bahwa

$$\frac{\partial l}{\partial \mu_i} = \frac{\nu(y_i - \mu_i)}{\mu_i^2}$$

maka dengan menggunakan aturan rantai akan menghasilkan

$$\frac{\partial l}{\partial \beta_j} = \nu \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mu_i^2} \frac{\partial \mu_i}{\partial \beta_j}$$

dimana

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

Sehingga diperoleh

$$\frac{\partial l}{\partial \beta_j} = \nu \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mu_i^2} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

serta matriks informasi Fisher dapat ditulis dalam bentuk

$$-E\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n} \frac{1}{\mu_i^2} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\mu_i^2} x_{ij} x_{ik} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

dimana $\mathbf{W}$ merupakan matriks diagonal pembobot yang unsur-unsurnya adalah

$$\mathbf{W} = \text{diag}\left\{\frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\mu_i^2}\right\}$$

Dalam kasus dimana fungsi hubungnya adalah kanonik, maka diperoleh

$$\frac{\partial l}{\partial \beta} = X^T (y - \mu)$$

Dengan matriks diagonal pembobotnya mempunyai unsur-unsur $\mathbf{W} = \text{diag}\{\mu_1^2, ..., \mu_n^2\}$.

3. PEMERIKSAAN MODEL GAMMA

3.1 Residu dan Nilai Leverage

Diasumsikan bahwa rata-rata komponen ke-*i* dari vektor respons merupakan beberapa fungsi nonlinear dari parameter regresi $\mu_i = \eta_i = \eta_i(\beta)$. Kemudian dapat dinyatakan devians residu komponen ke-*i* dari vektor respons sebagai berikut:

$$d_i = sign(y_i - \mu_i)\left[2\nu\left\{\ln\left(\frac{\hat{\mu}_i}{y_i}\right) - \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}\right\}\right]^{1/2} \quad (15)$$

dimana $\hat{\mu}_i = \eta_i(\hat{\beta})$.

Matriks hat adalah sama dengan

$$H(\hat{\beta}) = W(\hat{\beta})^{1/2} X(\hat{\beta}) \left[X(\hat{\beta})^T W(\hat{\beta}) X(\hat{\beta}) \right]^{-1} X(\hat{\beta})^T W(\hat{\beta})^{1/2} \quad (16)$$

dimana

$$X(\beta) = \frac{\eta(\beta)}{\partial \beta^T} = \left(\frac{\eta_i(\beta)}{\partial \beta_j} \right)$$

dan

$$W(\beta) = \text{diag}\left(\frac{1}{\eta_1(\beta)^2}, \ldots, \frac{1}{\eta_n(\beta)^2} \right) \quad (17)$$

Unsur-unsur diagonal utama dari matriks hat disebut juga sebagai nilai leverage, h_{ii}. Nilai leverage ini banyak digunakan dalam perhitungan nilai beberapa residu dalam model linear terampat seperti nilai residu devians dibakukan, nilai residu Pearson dibakukan, serta residu likelihood.

Residu devians dibakukan mempunyai bentuk

$$r_{D_i} = \frac{d_i}{\sqrt{(1-h_{ii})}} \quad (18)$$

dimana d_i adalah nilai devians komponen ke-i. Kemudian, residu Pearson dibakukan mempunyai bentuk

$$r_{P_i} = \frac{\mu_i(\hat{\beta})}{\sqrt{w_i(\hat{\beta})(1-h_{ii})}} = \frac{v(y_i - \hat{\mu}_i)}{\hat{\mu}_i \sqrt{1-h_{ii}}} \quad (19)$$

Sedangkan bentuk dari residu likelihoodnya diberikan oleh

$$r_L = \text{sgn}(y_i - \hat{y}_i) \sqrt{h_{ii} r_{P_i}^2 + (1 - h_{ii} r_{D_i}^2)} \quad (20)$$

Suatu titik data yang mempunyai nilai leverage yang besar, tapi juga mengikuti garis trend dalam model regresi tidak akan berpengaruh pada koefisien regresi. Besarnya pengaruh yang disebabkan oleh nilai leverage yang besar dapat merupakan suatu fungsi dari seberapa baik pengamatan tersebut mengikuti model yang dibentuk oleh kelompok data lainnya. Jelasnya, kombinasi yang dapat menyebabkan adanya pengaruh yang besar terhadap model adalah nilai leverage yang besar yang diikuti oleh residu yang relatif besar pula.

Lalu seberapa besar nilai leverage sehingga bisa dikatakan bahwa titik data tersebut merupakan data yang berpengaruh? Myers (1990) dan Collet (2002) menunjukkan fakta bahwa $\sum_{i=1}^{n} h_{ii} = p$. Rata-rata dari nilai leverage ini adalah p/n. Tentunya untuk setiap h_{ii} yang lebih besar daripada $2p/n$, maka dapat dikatakan bahwa data tersebut mempunyai potensi sebagai data yang berpengaruh.

3.2 Statistik Cook's *Distance*

Untuk masing-masing koefisien dalam model, pemeriksaan data berpengaruh akan memberikan suatu statistik dimana akan memberikan besarnya galat baku taksiran yang dapat merubah nilai koefisien model jika pengamatan ke-i dihapus dari analisis. Untuk melihat pengaruh data ke-i terhadap koefisien regresi (model), digunakan statistik:

$$D_{1i} = \frac{1}{p} \left(\hat{\beta}_i - \hat{\beta}_{(i)} \right)^T \mathbf{X}^T \mathbf{W} \mathbf{X} \left(\hat{\beta}_i - \hat{\beta}_{(i)} \right) \quad (21)$$

Cara lain untuk melihat pengaruh data ke-i terhadap model, digunakan statistik:

$$D_{2i} = \frac{2}{p}\left\{\log L\left(\hat{\beta}_i\right) - \log L\left(\hat{\beta}_{(i)}\right)\right\} \quad (22)$$

dimana $L\left(\hat{\beta}_i\right)$ merupakan fungsi likelihood untuk n pengamatan yang menyebar gamma dan $L\left(\hat{\beta}_{(i)}\right)$ merupakan fungsi likelihood $(n - 1)$ tanpa pengamatan ke-i yang juga menyebar gamma.

Dalam perhitungan D_{1i} dan D_{2i} (dalam Pers. 21) dan Pers. (22)), maka kita perlu mengamati $n \times p$ statistik untuk memperkirakan pengaruh data ke-i terhadap koefisien-koefisien regresi tersebut sehingga hal ini akan membuat perhitungan menjadi rumit. Untuk mengatasi hal tersebut ada statistik lain yang berhubungan dengan satu titik data tapi juga dapat mengukur pengaruh terhadap sekumpulan koefisien-koefisien regresi. Statistik itu disebut dengan Cook's *distance* atau Cook's D yang dapat dirumuskan dalam bentuk skalar sebagai berikut:

$$D_i = \frac{h_{ii} r_{P_i}^2}{p\left(1 - h_{ii}\right)} \quad (23)$$

Dalam hal ini statistik Cook's *distance* dihitung berdasarkan nilai residu Pearson dibakukan dan nilai leveragenya. Nilai D_i akan menjadi besar baik pada saat nilai residu Pearson yang besar pada titik data ke-i maupun pada saat nilai leverage yang besar.

4. CONTOH APLIKASI

Berikut ini akan dibahas mengenai contoh aplikasi dari pemeriksaan data berpengaruh dalam model regresi gamma. Data yang disajikan pada Tabel 1 merupakan data mengenai banyaknya klaim asuransi mobil yang diklasifikasikan ke dalam dua variabel, yaitu x_1 = lamanya (dalam tahun) dimana sejak klaim terakhir diajukan oleh pemegang polis, dan x_2 = gabungan dari umur, jenis kelamin, dan status marital. Sedangkan variabel n dan y masing-masing menunjukkan banyaknya klaim dan biaya total klaim. Variabel x_1 dan x_2 merupakan variabel kategorik yang masing-masing diklasifikasikan dengan empat dan lima kategori.

Tabel 1
Data tentang Asuransi Mobil

No.	x_1	x_2	n	Y
1	3	1	217151	63191
2	3	2	14506	4598
3	3	3	31964	9589
4	3	4	22884	7964
5	3	5	6560	1752
6	2	1	13792	4055
7	2	2	1001	380
8	2	3	2695	701
9	2	4	3054	983
10	2	5	487	114
11	1	1	19346	5552
12	1	2	1430	439
13	1	3	3546	1011
14	1	4	3618	1281
15	1	5	613	178
16	0	1	37730	11809
17	0	2	3421	1088
18	0	3	7565	2383
19	0	4	11345	3971
20	0	5	1291	382

Variabel x_1 diklasifikasikan menjadi empat kategori, yaitu: 3 = jenis mobil berlisensi dan bebas dari kecelakaan selama 3 tahun; 2 = jenis mobil berlisensi dan bebas dari kecelakaan selama 2 tahun, 1 = jenis mobil berlisensi dan bebas dari kecelakaan selama 1 tahun; serta 0 = untuk lainnya. Sedangkan variabel x_2 diklasifikasikan menjadi lima kategori, yaitu: 1 = wanita berumur < 25 tahun dan belum menikah, 2 = laki-laki berumur < 25 tahun dan belum menikah; 3 = laki-laki/wanita yang telah bercerai berumur < 25 tahun, 4 = wanita menikah yang berumur < 25 tahun; serta 5 = laki-laki menikah yang berumur < 25 tahun.

Data tersebut kemudian akan dianalisis melalui model regresi gamma dengan menggunakan fungsi hubung log. Tabel 2 menyajikan hasil-hasil ringkasan statistik mengenai model gamma. Berdasarkan tabel tersebut terlihat bahwa model sudah cukup baik dalam menggambarkan hubungan antara lamanya (dalam tahun) dimana sejak klaim terakhir diajukan oleh pemegang polis dan gabungan dari umur, jenis kelamin, dan status marital dengan biaya total klaim yang diasumsikan menyebar gamma. Hal ini terlihat dari rasio antara nilai devians dan derajat bebasnya (maupun rasio nilai chi-kuadrat Pearson dengan derajat bebasnya) yang cukup kecil, yaitu 24.269/17 = 1.439. Kemudian apabila kita lihat nilai penduga parameter β_1 dalam model regresi gamma ini menunjukkan hasil yang secara statistik tidak signifikan di bawah 5%, sedangkan untuk penduga parameter β_2 adalah signifikan.

Tabel 2
Ringkasan Statistik untuk Data Asuransi Mobil

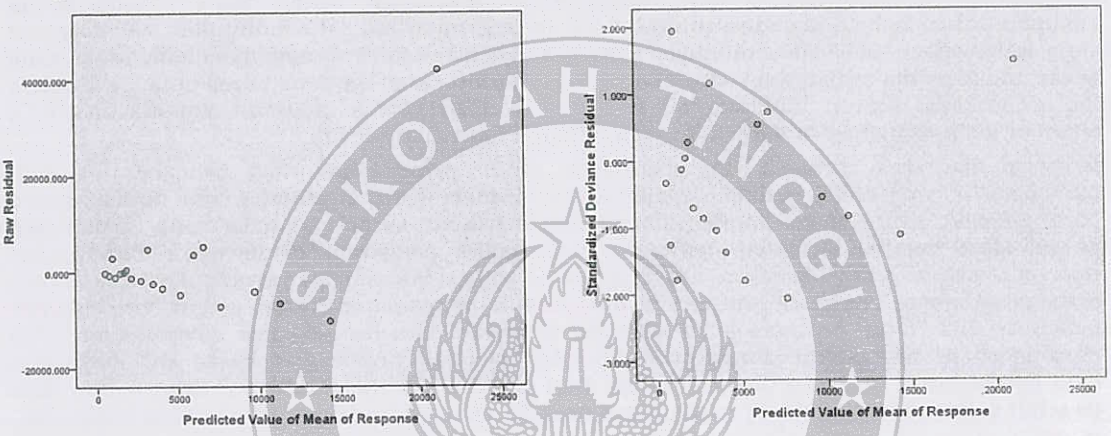
Parameter	Nilai Penduga	Galat Baku	Statistik chi-kuadrat	p-value
Intersep	9.420	0.6200	230.815	< 0.0001
X1	0.391	0.2053	3.632	0.0572
X2	-0.634	0.1623	15.253	< 0.0001
Skala	1.053	0.2919		
Devians = 24.469 (db = 17) Chi-kuadrat Pearson = 22.852 (db = 17) Log-likelihood = -182.090				

Tabel 3
Analisis residu, nilai leverage, dan statistik Cook's *distance*

No.	Residu Devians Baku	Residu Pearson Baku	Residu Likelihood	Nilai Leverage	Cook's Distance
1	2.221	1.494	1.697	0.240	0.519
2	-0.630	-0.829	-0.799	0.165	0.026
3	0.641	0.543	0.558	0.140	0.022
4	1.621	1.164	1.251	0.165	0.173
5	0.050	0.049	0.050	0.240	0.000
6	-0.762	-1.109	-1.061	0.160	0.037
7	-0.968	-2.060	-1.990	0.085	0.029
8	-0.830	-1.365	-1.339	0.060	0.015
9	-0.550	-0.701	-0.689	0.085	0.009
10	-0.956	-1.778	-1.674	0.160	0.058
11	-0.453	-0.540	-0.527	0.160	0.013
12	-0.932	-1.790	-1.734	0.085	0.027
13	-0.632	-0.855	-0.844	0.060	0.008
14	-0.116	-0.120	-0.120	0.085	0.000
15	-0.816	-1.251	-1.192	0.160	0.042
16	0.900	0.732	0.775	0.240	0.085
17	-0.732	-1.038	-0.994	0.165	0.035
18	0.309	0.283	0.286	0.140	0.005
19	3.266	1.943	2.217	0.165	0.703
20	-0.294	-0.325	-0.318	0.240	0.009

Untuk melihat apakah data tersebut terdapat pencilan akan digunakan analsis residu dan nilai leverage, kemudian dari hasil analisis residu tersebut untuk setiap data yang teridentifikasi sebagai data pencilan akan dilihat potensinya sebagai data berpengaruh dengan menggunakan statistik Cook's *distance*. Hasil analisis residu, nilai leverage, dan statistik Cook's *distance* disajikan pada Tabel 3.

Dari hasil analisis residu, terutama nilai-nilai dari residu devians, diperoleh nilai mutlak dari residu devians baku untuk pengamatan ke-1 dan ke-19 adalah lebih besar daripada 2.0, yaitu masing-masing sebesar r_{D_1} = 2.221 dan $r_{D_{19}}$ = 3.266. Walaupun nilai mutlak residu Pearson baku dan residu likelihood untuk kedua pengamatan tersebut kurang dari 2.0, kecuali nilai mutlak residu likelihood untuk pengamatan ke-19 yang sebesar $r_{L_{19}}$ = 2.217, tetapi kedua pengamatan tersebut dapat dianggap sebagai data pencilan yang mungkin berpengaruh pada model regresi gamma. Perlu dicatat bahwa nilai leverage untuk kedua pengamatan tersebut adalah kurang dari (2)(3)/20 = 0.3, tetapi sekali lagi kedua pengamatan tersebut berpotensi sebagai data yang berpengaruh.



(a) (b)

Gambar 5. Plot antara residu dengan nilai dugaan respons: (a) plot antara residu biasa dengan nilai dugaan respons, (b) plot antara residu devians baku biasa dengan nilai dugaan respons

Pada Gambar 5 menampilkan plot antara residu dengan nilai dugaan respons: (a) plot antara residu biasa dengan nilai dugaan respons, dan (b) plot antara residu devians baku biasa dengan nilai dugaan respons. Dari kedua gambar tersebut terlihat bahwa pengamatan ke-1 merupakan data pencilan karena berada di luar kelompok besarnya. Setelah teridentifikasi bahwa pengamatan ke-1 dan ke-19 dianggap sebagai data pencilan, maka akan dilihat bagaimana pengaruh dari kedua pengamatan tersebut terhadap model dengan menggunakan statistik Cook's *distance*. Dari Tabel 1 terlihat bahwa nilai statistik Cook's *distance* untuk kedua pengamatan tersebut masing-masing adalah 0.519 dan 0.703, keduanya dianggap besar karena lebih besar daripada 0.5. Artinya memang kedua pengamatan tersebut merupakan suatu data yang berpengaruh terhadap model.

Tabel 4
Ringkasan Statistik untuk data asuransi mobil setelah menghilangkan pengamatan ke-1 dan ke-19

Parameter	Nilai Penduga	Galat Baku	Statistik chi-kuadrat	p-value
Intersep	8.919	0.5727	242.526	< 0.0001
X1	0.452	0.2070	4.763	0.0293
X2	-0.609	0.1586	14.754	< 0.0001
Skala	0.854	0.2540		
Devians = 17.430 (db = 15) Chi-kuadrat Pearson = 15.075 (db = 15) Log-likelihood = -157.112				

Selanjutnya, analisis dilakukan kembali dengan menghilangkan pengamatan ke-1 dan ke-19 dari analisis yang hasilnya disajikan pada Tabel 4. Tampak bahwa terdapat perubahan hasil yang cukup berarti, terutama pada tingkat signifikansi untuk parameter β_1. Parameter β_1 yang sebelumnya tidak signifikan di bawah 5%, setelah pengamatan ke-1 dan ke-19 dihilangkan dari analisis menjadi signifikan secara statistik di bawah 5%.

Demikian juga terjadi penurunan nilai devians dan nilai chi-kuadrat Pearson yang cukup signifikan. Selisih nilai devians antara model awal dengan model revisi adalah (24.469 – 17.430) = 7.039, begitu juga Selisih nilai chi-kuadrat Pearson antara model awal dengan model revisi adalah (22.852 – 15.075) = 7.777. Keduanya adalah signifikan di bawah 5%. Selain itu rasio antara nilai devians maupun chi-kuadrat Pearson terhadap derajat bebasnya adalah mendekati satu. Hal ini menunjukkan bahwa tingkat kecocokan model terhadap data juga semakin tinggi.

5. KESIMPULAN

Berdasarkan pembahasan di atas, maka dapat dikatakan bahwa para peneliti harus memperhatikan bahwa diagnosa di atas tidak menggambarkan satu kumpulan alat diagnosa yang independen. Sebagai contohnya, misalnya apabila Cook's D menghasilkan harga yang besar, maka paling sedikit ada satu nilai residu atau nilai leverage yang besar pula. Jadi dalam hal ini berbagai ukuran statistik, baik nilai residu, nilai leverage, maupun statistik Cook's D, tersebut akan saling melengkapi dan perlu dilihat secara menyeluruh.

Berbagai alat atau statistik yang digunakan untuk pemeriksaan data pencilan dan data berpengaruh yang dibahas dalam makalah ini dirancang untuk memberikan tanda kepada para peneliti, yaitu suatu tanda dimana jika terdapat sumber-sumber untu melakukan penyelidikan kembali terhadap beberapa data, maka pengaruh itu harus diteliti dengan seksama. Hal ini perlu dilakukan jika terjadi hasil yang tidak diinginkan yang disebabkan oleh satu pengamatan. Apakah kita perlu menghapus pengamatan yang sangat berpengaruh tersebut? Kita harus bersikap lebih seksama terhadap data berpengaruh daripada terhadap data pencilan. Jika pada evaluasi hasil diperoleh masalah yang serius, maka kehadiran dari data berpengaruh itu perlu dipertanyakan. Tapi jika hasil evaluasi menunjukkan bahwa data tersebut valid, maka tindakan penghapusan data itu menjadi tindakan yang kurang bijaksana.

Dalam beberapa hal mungkin data tersebut dapat memberikan dukungan utama pada model yang telah dirumuskan. Selanjutnya, nilai laverage yang ideal adalah yang memenuhi distribusi uniform. Hal ini terjadi jika semua nilai diagonal matriks HAT diambil pada nilai p/n, dan data yang berpotensi sebagai data berpengaruh diturunkan dari leverage yang dibagi secara merata di antara kumpulan data, tapi hal ini sulit dilakukan. Kondisi seperti ini tidak berarti bahwa model regresi tidak bisa diperbaiki. Singkatnya, informasi yang diperoleh melalui berbagai diagnosa tersebut menjadikan para peneliti perlu melakukan penyelidikan lebih jauh, sehingga tujuan dari pembentukan model yang efektif bisa dicapai.

Dalam analisis regresi klasik, prosedur yang ditempuh untuk memperoleh model yang baik, yaitu melalui pengujian hipotesis, pemilihan variabel, dan lain-lain, seringkali gagal dalam pembentukan modelnya. Hal ini juga berlaku dalam pemodelan linear terampat, khususnya untuk model regresi gamma yang telah dibahas dalam makalah ini. Prosedur tersebut tidak memberikan penjelasan yang memadai mengapa model menjadi tidak baik. Dari contoh pemakaian yang telah dibahas pada bagian sebelumnya dapat ditunjukkan bahwa betapa satu buah pengamatan dapat mengendalikan variabel. Dengan demikian, maka pemeriksaan terhadap data berpengaruh ini perlu dilakukan dalam proses pembentukan model yang baik.

DAFTAR PUSTAKA

[1] Agresti, A. (2002). *Categorical Data Analysis.* Second Edition. New York: John Wiley and Sons.
[2] Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Second Edition. New York: John Wiley and Sons.
[3] Aitkin, M., D. Anderson, B. Francis, and J. Hinde. (1989). *Statistical Modelling in GLIM.* Oxford: Clorendeon Press.
[4] Baker, R.J., and J.A. Nelder. (1978). *Generalized Linear Interactive Modeling (GLIM).* Release 3. Oxford: Numerical Algorithms Group.
[5] Collet, D. (2003). *Modeling Binary Data.* Second Edition. London: Chapman and Hall.
[6] De Jong, P., and Heller, Z. G. (2008). *Generalized Linear Models for Insurance Data.* Cambridge: Cambridge University Press

[7] Dobson, A. (2002) *An Introduction to Generalized Linear Models*. Second Edition. London: Chapman and Hall.
[8] Draper, N.R., and H. Smith. (1981). *Applied Regression Analysis*. Second Edition. New York: John Wiley and Sons.
[9] Lawal, B. (2003) *Categorical Data Analysis With SAS And SPSS Applications*. London: Lawrence Erlbaum Associates.
[10] McCullagh, P., and J.A. Nelder (1983). *Generalized Linear Models*. Second Edtion. New York: Chapman and Hall.
[11] Myers, R.H. (1990). *Classical and Modern Regression With Applications*. Boston: PWS-KENT Publishing Company.
[12] Nelder, J.A., and R.W.M. Wedderbun. (1972). Generalized Linear Models. *Journal of Royal Statistical Society*, Series A 153: 370-384.
[13] Santner, T.J., and D.E. Duffy. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag.
[14] Uusipaikka, E. (2009). *Confidence Intervals in Generalized Regression Models*. London: Chapman and Hall.

SEKOLAH TINGGI
ILMU KEPOLISIAN
BHAKTI - DHARMA - WASPADA