PEMERIKSAAN DATA BERPENGARUH DALAM MODEL REGRESI GAMMA

Nusar Hajarisman[1]
[1] Staf Pengajar Jurusan Statistika Universitas Islam Bandung
Jalan Purnawarman No. 69 Bandung 40116
nrisman@yahoo.co.uk

Abstract

In certain cases, often encountered a group of data where the observed respone variable shaped non-negative and not symmetrical (skewed to the right). The case data such as this one can be found in the field of insurance, such as variable stating the amount of claims or states of time between when a customer claims to obtain these claims. To handle such cases one of them is by using a generalized linear model, with respone variables that follow the gamma distribution. The paper discusses the inspection data outliers and influential data in the modeling of the respone following the gamma distribution. Several statistical measures used to examine the outlier data is the value of leverage, standardized deviance residual, standardized Pearson residual, and residual likelihood. Then the data outliers potentially influential data will be examined using Cook's distance statistics.

Keywords: Gamma Distributions, Leverage, Standardized Deviance Residual, Standardized Pearson Residual, And Residual Likelihood, Cook's Distance Statistics

1. Pendahuluan

Dalam beberapa tahun terakhir ini, model linear yang berbentuk: $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$, dengan asumsi bahwa unsur-unsur dari **e** adalah berdistribusi normal identik dan saling bebas, $\mathbf{e} \sim NID(0, \sigma^2)$, merupakan basis dari kebanyakan analisis untuk data-data kontinu. Dengan adanya berbagai kelebihan dalam teori statistik dan perangkat lunak komputer, dapat digunakan metode yang analog dengan pengembangan model linear dalam beberapa situasi sebagai berikut: (1) variabel respon mempunyai distribusi selain distribusi normal , mungkin dapat berbentuk kategori daripada kontinu; dan (2) hubungan antara variabel respon dengan variabel prediktornya tidak perlu berbentuk linear. Salah satu dari kelebihan ini yang telah banyak dikenal juga sebagai sifat-sifat dari distribusi normal dibagi ke dalam kelas yang lebih luas dari suatu distribusi yang disebut juga sebagai distribusi dari keluarga eksponensial.

Pada kasus tertentu, sering kali ditemui suatu gugus data dimana variabel respon yang diamati berbentuk nonnegatif dan tidak simetris (miring ke kanan). Kasus data seperti ini salah satunya dapat ditemui dalam bidang asuransi, misalnya variabel yang menyatakan besarnya klaim atau waktu antar menyatakan klaim dengan saat seorang nasabah memperoleh klaimnya tersebut. Untuk menangani kasus seperti itu, ada dua cara yang dapat dilakukan yaitu (1) menggunakan transformasi ke normalitas, kemudian menggunakan model linear normal pada variabel respon yang sudah ditransformasi; (2) menggunakan model linear terampat (*generalized linear model*, GLM), dimana salah satunya menggunakan variabel respon yang mengikuti distribusi gamma.

Dalam makalah ini akan dibahas mengenai penerapan model regresi gamma pada bidang asuransi. Model regresi gamma merupakan model regresi yang variabel responnya didasarkan pada distribusi gamma, yang juga merupakan anggota dari eksponensial, sehingga terminologi dari model linear terampat dapat diterapkan di sini. Sebagaimana

yang telah diketahui bahwa langkah-langkah pemodelan statistika adalah proses pencocokan data terhadap variabel respon yang diikuti oleh proses pengujian hipotesis yang berhubungan dengan evaluasi kelayakan suatu model. Dengan kata lain, setelah model dicocokan terhadap data pengamatan dengan variabel respon, maka pada langkah berikutnya adalah melakukan pemeriksaan apakah model dugaan tersebut layak atau tidak. Terdapat sejumlah cara yang dapat menjadikan model dugaan itu menjadi tidak layak. Yang paling penting dari semuanya adalah komponen sistematik linear dari model dinyatakan dengan tidak benar.

Sebagai contoh, misalnya model mungkin tidak menyertakan variabel prediktor yang seharusnya berada di dalam model, atau mungkin nilai-nilai yang diambil oleh satu atau lebih variabel penjelas perlu ditransformasi. Transformasi dari peluang respon mungkin kurang tepat. Data mungkin berisi observasi tertentu yang membentuk sebagai data pencilan (*outliers*), yang dapat menjadikan model tidak bagus. Selain itu terdapat observasi yang merupakan data berpengaruh (*influence data*), yang dapat mempengaruhi kesimpulan dari hasil analisis. Teknik-teknik yang digunakan untuk menentukan kelayakan model dugaan secara kolektif disebut sebagai diagnostik (*diagnostics*). Teknik semacam ini dapat dilakukan melalui uji statistik formal, tetapi lebih sering juga menyangkut evaluasi yang kurang formal mengenai tabel dari nilai statistik tertentu atau melalui gambaran secara grafik atau plot dari nilai-nilai tersebut.

Pada makalah ini pembahasan akan lebih difokuskan pada pemeriksaan data berpengaruh dalam pemodelan yang responnya 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 devian dibakukan, residu Pearson dibakukan, dan residu likelihood. Kemudian data pencilan yang berpotensi sebagai data berpengaruh akan diperiksa dengan menggunakan statistik Cook's *distance*. Contoh penerapan dari pemeriksaan data berpengaruh untuk model regresi gamma pada bidang asuransi akan dibahas pada bagian empat.

2. Model Regresi Gamma

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

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

yaitu, koefisien variasi pengamatannya merupakan suatu konstanta dan koefisien variasi umum dinyatakan dengan σ^2. Apabila nilai yang mungkin dari variabel respon berupa bilangan nyata positif dan apabila respon 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

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

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

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

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

2.1 Distribusi Gamma

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

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

dan fungsi pembangkit kumulan diberikan oleh

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

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 kumulan ke-k diberikan oleh

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

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

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

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

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

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

Kumulan dari variabel yang dibakukan adalah

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

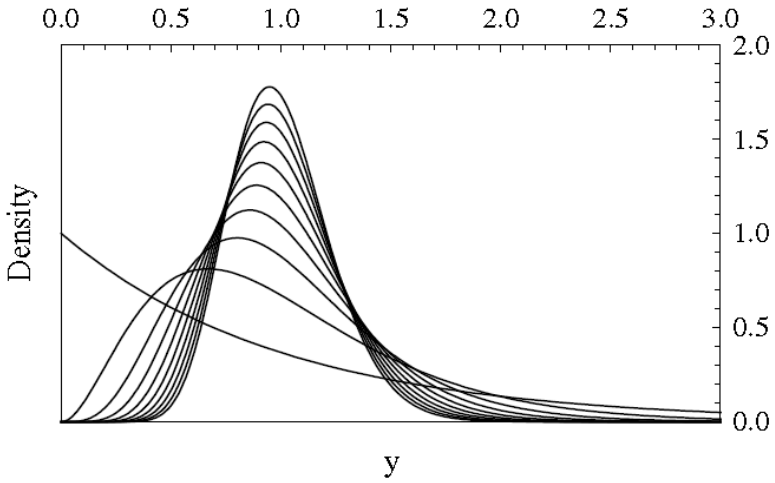
diperoleh perluasan deret Taylor

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

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

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

dengan Φ merupakan fungsi distribusi kumulatif dari distribusi normal baku. Gambar 1 menunjukkan grafik dari densitas gamma dengan rata-rata satu dan berbagai nilai dari parameter bentuk ν.



Gambar 1. Fungsi Densitas Dari Distribusi Gamma Dengan Rata-Rata 1 Dan Parameter Bentuk 1, 3, ..., 19

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.

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.

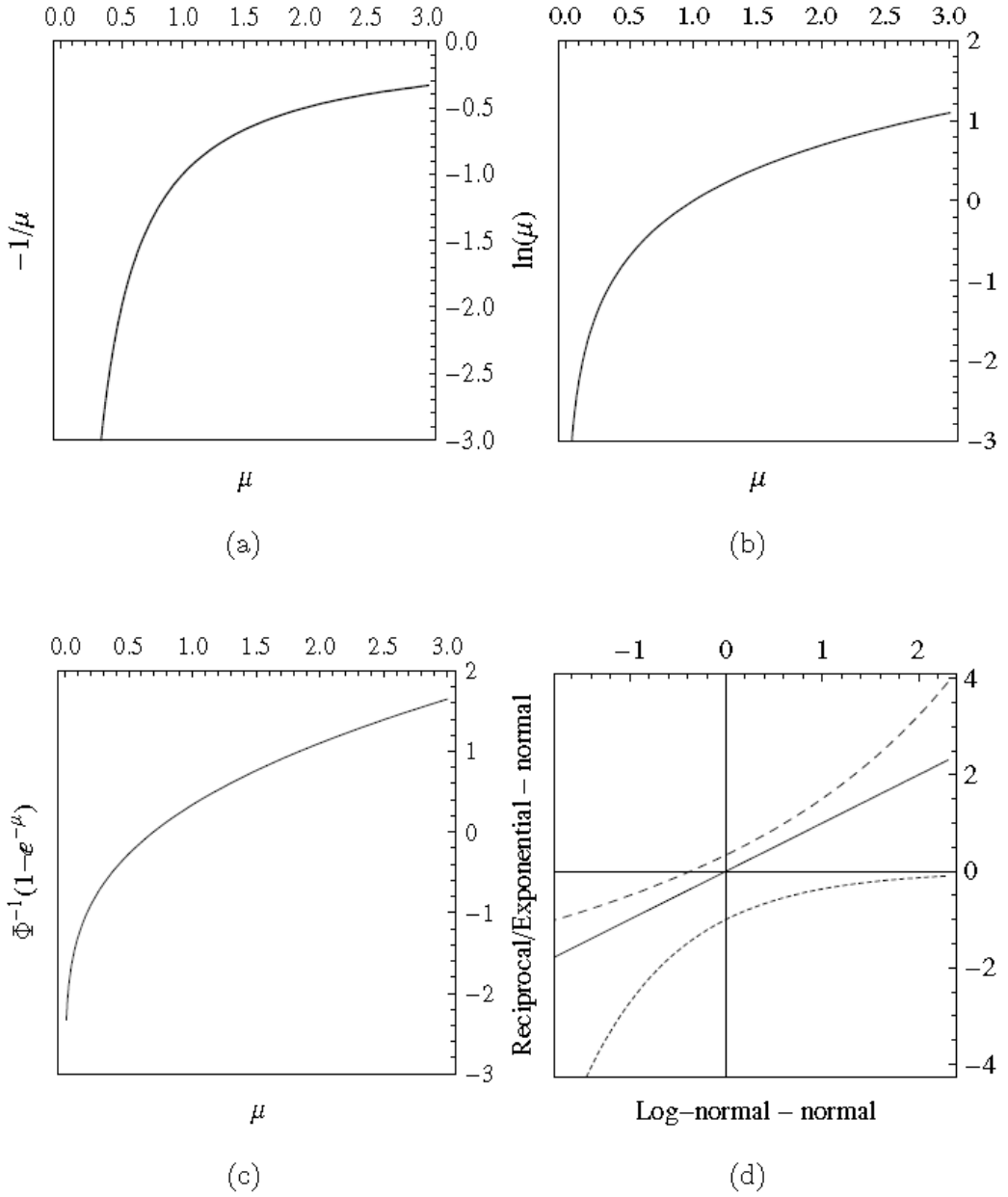
Dalam hal fungsi hubung log, maka distribusinya adalah log-normal dibakukan dan distribusi normal sebab

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

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



Gambar 2. 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

2.3 Fungsi Likelihood untuk Model Gamma

Pada saat respon y_1, ..., 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:

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

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

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

Yang merupakan hasil dalam suatu rumusan yang berisi parameter β_1, β_2, ..., β_p.
Dalam kasus dimana fungsi hubungnya adalah kanonik, maka

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

Dengan demikian fungsi log-likelihoodnya menjadi

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

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

Jadi, statistik dari

$$\sum_{i=1}^{n} y_i x_{ij}, \qquad \text{untuk } j = 1,..., p$$

merupakan statistik cukup minimal untuk parameter β_1, β_2, ..., β_p pada ν yang tetap (*fixed*).

Fungsi likelihood untuk model dugaan untuk model gamma dengan parameter bentuk ν tetap dapat dinyatakan dalam bentuk

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

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

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

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

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

2.4 Pendugaan Parameter

Diketahui bahwa

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

maka dengan menggunakan aturan rantai akan menghasilkan

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

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

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

dengan **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

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

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

3. Diagnostik pada Model Regresi Gamma

3.1 Residu dan Nilai Leverage

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

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

dengan $\hat{\mu}_i=\eta_i\left(\hat{\beta}\right)$.

Matriks hat adalah sama dengan

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

dengan

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

dan

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

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 devian dibakukan, nilai residu Pearson dibakukan, serta residu likelihood.

Residu devian dibakukan mempunyai bentuk $r_{D_i}=\frac{d_i}{\sqrt{\left(1-h_{ii}\right)}}$ (18)

dengan d_i adalah nilai devian komponen ke-*i*. Kemudian, residu Pearson dibakukan mempunyai bentuk

$$r_{P_i}=\frac{\mu_i\left(\hat{\beta}\right)}{\sqrt{w_i\left(\hat{\beta}\right)\left(1-h_{ii}\right)}}=\frac{\nu\left(y_i-\hat{\mu}_i\right)}{\hat{\mu}_i\sqrt{1-h_{ii}}} \tag{19}$$

Sedangkan bentuk dari residu likelihoodnya diberikan oleh

$$r_L = sign\,(y_i - \hat{y}_i)\sqrt{h_{ii}\,r_{Pi}^2 + (1 - h_{ii}\,r_{Di}^2)} \tag{20}$$

Suatu titik data yang mempunyai nilai leverage yang besar, tapi juga mengikuti garis tren 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. Selanjutnya, Myers (1990) dan Collet (2003) 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 2*p*/*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:

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

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

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

dengan $L\left(\hat{\beta}_i\right)$ 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 perlu diamati $n \times p$ 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:

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

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 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. Data diperoleh dari suatu perusahaan asuransi di kota Bandung, Jawa Barat.

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.

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

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 devian dan derajat bebasnya (maupun rasio nilai chi-kuadrat Pearson dengan derajat bebasnya) yang cukup kecil, yaitu 24,269/17 = 1,439. Kemudian apabila dilihat 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.

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.

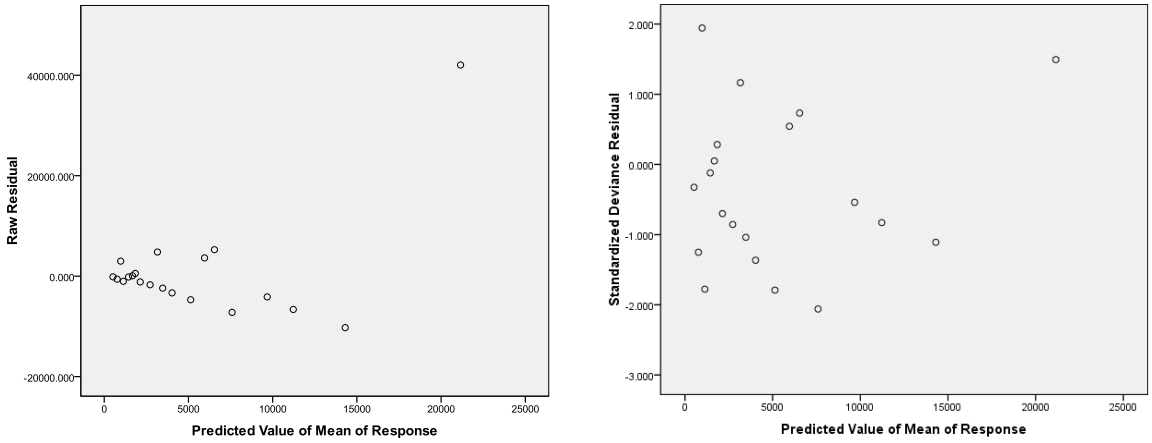
Tabel 2. Ringkasan Statistik untuk Data Asuransi Mobil

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		
Devian = 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 Devian 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

Dari hasil analisis residu, terutama nilai-nilai dari residu devian, diperoleh nilai mutlak dari residu devian 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 Respon: (a) Plot Antara Residu Biasa Dengan Nilai Dugaan Respon, (b) Plot Antara Residu Devian Baku Biasa Dengan Nilai Dugaan Respon

Pada Gambar 5 menampilkan plot antara residu dengan nilai dugaan respon: (a) plot antara residu biasa dengan nilai dugaan respon, dan (b) plot antara residu devian baku biasa dengan nilai dugaan respon. 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		
Devian = 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 devian dan nilai chi-kuadrat Pearson yang cukup signifikan. Selisih nilai devian 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 devian 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, 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 untuk 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. Sebaiknya 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 leverage 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., *Categorical Data Analysis*, Second Edition, John Wiley and Sons, New York, 2002.
2. Agresti, A., *An Introduction to Categorical Data Analysis*, Second Edition, John Wiley and Sons, New York, 2007.
3. Aitkin, M., Anderson, D., Francis, B., and Hinde, J., *Statistical Modelling in GLIM*, Clorendeon Press, Oxford, 1989.
4. Baker, R.J., and Nelder, J.A., *Generalized Linear Interactive Modeling (GLIM)*, Release 3, Numerical Algorithms Group, Oxford, 1978.

5. Collet, D., *Modeling Binary Data*, Second Edition, Chapman and Hall, London, 2003.
6. de Jong, P., and Heller, Z. G., *Generalized Linear Models for Insurance Data*, Cambridge University Press, Cambridge, 2008.
7. Dobson, A., *An Introduction to Generalized Linear Models*, Second Edition, Chapman and Hall, London, 2002.
8. Draper, N.R., and H. Smith, *Applied Regression Analysis*, Second Edition, John Wiley and Sons, New York,1981
9. Lawal, B., *Categorical Data Analysis With SAS And SPSS Applications*, Lawrence Erlbaum Associates, London, 2003.
10. McCullagh, P., and J.A. Nelder, *Generalized Linear Models*, Second Edtion, Chapman and Hall, New York,1983
11. Myers, R.H., *Classical and Modern Regression With Applications*, PWS-KENT Publishing Company, Boston, 1990.
12. Nelder, J.A., and R.W.M. Wedderbun, Generalized Linear Models. *Journal of Royal Statistical Society*, Series A, 1972, No. 153: 370-384.
13. Santner, T.J., and D.E. Duffy, *The Statistical Analysis of Discrete Data*, Springer-Verlag, New York, 1989.
14. Uusipaikka, E., *Confidence Intervals in Generalized Regression Models*, Chapman and Hall, London, 2009.