Statistika, Vol. 10 No. 1, 51 – 62
Mei 2010

Pendekatan Fungsi Quasi-Likelihood dan Implementasinya dalam Sistem SAS

NUSAR HAJARISMAN

Jurusan Statitika, Universitas Islam Bandung
Jalan Purnawarman 69 Bandung 40116.
E-mail: nrisman@yahoo.co.uk

ABSTRAK

Seringkali perhatian utama peneliti ditekankan pada bagaimana rata-rata respons atau bentuk fungsional lainnya dipengaruhi oleh satu atau lebih kovariat. Biasanya terdapat informasi prior dalam mengamati kealamiahan bentuk hubungan tersebut, akan tetapi seringkali diperlukan informasi dari kumulant atau moment yang berordo lebih tinggi (McCullagh dan Nelder, 1983). Dalam makalah ini akan dibahas mengenai bagaimana statistik inferens dapat dibuat berdasarkan suatu percobaan dimana tidak tersedia cukup informasi dalam membentuk fungsi likelihood, yaitu melalui pembentukan fungsi quasi-likelihood.
Kata Kunci: generalized linear model; fungsi likelihood; quasi-likelihood; fungsi varians; prosedur GLIMMIX.

1. PENDAHULUAN

Dalam model linear terampat (generalized linear models, GLM), fungsi kemungkinan mempunyai peranan penting khususnya untuk keperluan statistik inferens. Berbagai metode pendugaan parameter dan pengujian hipotesis untuk mengevaluasi kecocokan model banyak didasarkan pada fungsi kemungkinan atau juga dikenal dengan fungsi likelihood. Untuk membentuk fungsi likelihood biasanya dilakukan melalui mekanisme probabilistik yang menyatakan, untuk nilai-nilai parameter dengan range tertentu, peluang dari seluruh sampel yang relevan yang telahdiamati. Spesifikasi semacam itu menunjukkan bahwa peneliti harus mempunyai pengetahuan yang memadai tentang bagaimana data tersebut dibangkitkan atau peneliti mempunyai pengalaman berdasarkan hasil percobaan atau penelitian sebelumnya.

Seringkali tidak tersedia teori yang memadai tentang mekanisme acak dimana data tersebut dibangkitkan. Namun demikian, peneliti dapat menyatakan range dari nilai respons yang mungkin (misalnya diskret, kontinu, positif, negatif, atau bentuk lainnya), serta pengalaman masa lalu berdasarkan data yang serupa, biasanya cukup digunakan untuk menyatakan (dalam bentuk kualitatif) karakeristik tambahan mengenai gambaran data, seperti:

- Bagaimana rata-rata atau median dari respons dipengaruhi oleh stimulus atau perlakuan eksternal;
- Bagaimana keragaman dari perubahan respons dengan rata-rata respons;
- Apakah data adalah saling bebas;
- Apakah distribusi dari variabel respons di bawah kondisi perlakuan yang tetap adalah simetris, miring positif, atau miring negatif/

Seringkali perhatian utama peneliti ditekankan pada bagaimana rata-rata respons atau bentuk fungsional lainnya dipengaruhi oleh satu atau lebih kovariat. Biasanya terdapat informasi prior dalam mengamati kealamiahan bentuk hubungan tersebut, akan tetapi seringkali diperlukan informasi dari kumulant atau moment yang berordo lebih tinggi (McCullagh dan Nelder, 1983). Dalam makalah ini akan dibahas mengenai bagaimana statistik inferens dapat dibuat berdasarkan suatu percobaan dimana tidak tersedia cukup informasi dalam membentuk fungsi likelihood, yaitu melalui pembentukan fungsi quasi-likelihood.

Fungsi quasi-likelihood ini biasanya digunakan pada suatu data non-Gaussian untuk memodelkan rata-rata dan dispersinya (Pawitan, et al, 2006). Sebagai contoh, misalnya dalam menganalisis data cacahan (*count data*) melalui regresi Poisson biasa yang mengasumsikan varians sama dengan rata-ratanya: $V(\mu) = \mu$. Akan tetapi, seringkali terdapat keragaman ekstra Poisson (atau dikenal dengan istilah overdispersion), dimana $V(\mu) = \phi\mu$, dimana $\phi > 1$, sehingga

distribusi apa yang nantinya digunakan menjadi tidak jelas. Dalam faktanya, menurut Jorgensen (1986) menunjukan bahwa tidak ada keluarga GLM yang memenuhi hubungan rata-rata dan varians ini, sehingga tidak ada penyesuaian yang sederhana terhadap densitas dari Poisson biasa dalam menangani masalah overdispersi ini. Sekali lagi, untuk menangani masalah ini dapat digunakan pendekatan quasi-likelihood (Wedderburn, 1974). Melalui pendekatan quasi-likelihood ini, peneliti hanya perlu menyatakan bentuk hubungan antara rata-rata dan varians daripada harus memerlukan informasi yang lengkap dari distribusi data.

2. FUNGSI QUASI-LIKELIHOOD

Misalkan diperoleh variabel respons $y_1, ..., y_n$ yang saling bebas dengan rata-rata $E(y_i) = \mu_i$ dan varians $\text{var}(y_i) = \phi V(\mu_i)$, dimana μi merupakan suatu fungsi dari parameter regresi $\beta = (\beta_1, ..., \beta_p)$ yang tidak diketahui dan V( ) adalah fungsi yang diketahui. Dalam makalah ini akan dibahas mengenai penggunaan fungsi quasi-likelihood untuk keperluan inferensi dari model yang bukan merupakan keluarga GLM.

McCullagh dan Nelder (1983) mendefinisikan quasi-likelihood (atau selanjutnya disingkat dengan QL) sebagai fungsi $q(\mu_i; y_i)$ yang memenuhi

$$\frac{\partial q(\mu_i; y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)} \tag{1}$$

dan untuk data yang saling bebas, total quasi-likelihood adalah $\sum_i q(\mu_i; y_i)$. Penduga koefisien regresi $\hat{\beta}$ memenuhi persamaan skor seperti dalam GLM

$$\sum_i \frac{\partial q(\mu_i; y_i)}{\partial \beta} = \sum_i \frac{\partial \mu_i}{\partial \beta} \frac{(y_i - \mu_i)}{\phi V(\mu_i)} = 0 \tag{2}$$

Adalah mungkin untuk memperlakukan pendekatan quasi-likelihood sederhana sebagai pendekatan persamaan pendugaan, yaitu dengan tidak memperhatikan fungsi pendugaan sebagai fungsi skor. Dalam hal ini kita masih menurunkan penduga dengan menggunakan persamaan pendugaan yang sama, tetapi inferensinya tidak didasarkan pada besaran likelihood seperti statistik uji rasio likelihood, dan lebih berdasarkan pada sifat-sifat distribusional dari penduganya secara langsung.

Selanjutnya kita perlu menyelidiki penggunaan pendekatan quasi-likelihood dalam beberapa konteks:

- Terdapat suatu struktur peluang tertentu, suatu quasi-distribusi dari keluarga distribusi GLM yang tidak sesuai dengan distribusi yang sesungguhnya. Sebagai contoh, misalnya distribusi yang sebenarnya adalah binomial negatif, sedangkan menurut quasi-distribusinya adalah Poisson. Demikian juga misalnya, quasi-distribusi mungkin berupa skala kontinu, padahal distribusi yang sebenarnya merupakan berskala diskret, atau sebaliknya.
- Tidak terdapat struktur peluang tertentu, akan tetapi mempunyai quai-likelihood, yaitu terdapat suatu nilai riil dari fungsi $q(\mu_i; y_i)$ yang mempunyai turunan sebagaimana dalam Pers. (2).
- Persamaan pendugaan dalam (2) lebih jauh dapat diperluas untuk variabel respons yang berkorelasi. Untuk kasus seperti ini, maka tidak ada nilai riil dari fungsi $q(\mu_i; y_i)$.

Pendekatan quasi-likelihood pertama kali dibentuk untuk mengatasi dua permasalahan pertama di atas dan mempunyai sifat-sifat yang perlu dicatat sebagai berikut:

- Berbeda dengan pendekatan fungsi likelihood biasa, dalam fungsi quasi-likelihood tidak menyatakan struktur peluang tertentu, tetapi hanya memerlukan asumsi mengenai dua buah moment pertama. Hal ini pendekatan fungsi quasi-likelihood mempunyai fleksibilitas yang tinggi.

- Pendugaan untuk parameter regresi hanya untuk rata-rata. Untuk pendekatan berdasarkan likelihood untuk pendugaan parameter dispersi ϕ diperlukan beberapa teorema tambahan.

Wedderburn (1974) menurunkan beberapa sifat dari quasi-likelihood, tapi perlu dicatat bahwa teori ini mengasumsikan bahwa ϕ diketahui. Dengan asumsi ini terlihat bahwa quasi-likelihood merupakan log-likelihood sebenarnya jika dan hanya jika respons y_i berasal dari model keluarga eksponensial dengan parameter-satu (keluarga GLM dengan $\phi = 1$) dengan densitas-log:

$$q(\mu; y) = \theta y - b(\theta) + c(y) \tag{3}$$

dimana $\mu = b'(\theta)$ dan $V(\mu) = b''(\theta)$. Pemilihan hubungan hubungan rata-rata dan varians adalah sama dengan untuk memilih suatu fungsi $b(\theta)$. Dengan kata lain, quasi-distribusi berhubungan dengan quasi-likelihood dalam keluarga eksponensial. Sepanjang inferens ordo pertama yang diperhatikan, maka quasi-likelihood diartikan oleh hubungan rata-rata dan varians sebagaimana dalam likelihood yang sebenarnya. Sebagai contoh, diperoleh

$$E\left(\frac{\partial q}{\partial \mu}\right) = 0$$

dan

$$E\left(\frac{\partial q}{\partial \mu}\right)^2 = -E\left(\frac{\partial^2 q}{\partial \mu^2}\right) = \frac{1}{V(\mu)}$$

Apabila likelihood sebenarnya adalah $l(\mu)$, maka menurut teorema batas-bawah Cramer-Rao dapat dinyatakan bahwa

$$-E\left(\frac{\partial^2 q}{\partial \mu^2}\right) = \frac{1}{V(\mu)} \le -E\left(\frac{\partial^2 l}{\partial \mu^2}\right)$$

dengan kesamaan jika likelihood sebenarnya mempunyai bentuk dari keluarga eksponensial. Jika parameter ϕ tidak diketahui, maka quasi-distribusi adalah dalam bentuk umum dan bukan dalam bentuk keluarga eksponensial.

Tabel 1. Fungsi Varians dari Berbagai Model Keluarga Eksponensial

Nama	**V(μ)**	**Q(μ; y)**	**Batas**
Normal	1	$-(y-\mu)^2/2$	
Overdispersed Poisson	μ	$y\log\mu - \mu$	$\mu \ge 0$; $y \ge 0$
Overdispersed Binomial	μ(1 − μ)	$y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$	$0 \le \mu \le 1$; $0 \le y \le 1$
Gamma	μ^2	$-(y/\mu) - \log\mu$	$\mu \ge 0$; $y \ge 0$
Power varians	μ^p	$\mu^{-p}\left(\frac{y\mu}{1-p} - \frac{\mu^2}{2-p}\right)$	$\mu \ge 0$; $y \ge 0$; $p \ne 0, 1, 2$
Binomial negatif	$\mu + \mu^2/k$	$y\log\left(\frac{\mu}{\kappa+\mu}\right) + \kappa\log\left(\frac{\kappa}{\kappa+\mu}\right)$	$\mu \ge 0$; $y \ge 0$

Walaupun dalam pendekatan QL hanya dinyatakan bentuk hubungan rata-rata dan varians, akan tetapi persamaan pendugaan menyatakan suatu quasi-distribusi untuk moment yang berordo lebih tinggi. Lee dan Nelder (1999) menyatakan bahwa oleh karena persamaan pendugaan QL merupakan persamaan skor yang diturunkan dari QL, tetapi bentuk distribusinya mengikuti pola kumulant berordo lebih tinggi yang mungkin berasal dari

keluarga GLM. Diantara distribusi yang mempunyai hubungan rata-rata dan varians tersebut, keluarga GLM mempunyai posisi khusus sebagai berikut:

- Kita dapat menyatakan $-E\left(\partial^2 q / \partial \mu^2\right)$ sebagai informasi yang tersedia dalam y di sekitar μ ketika hanya hubungan antara rata-rata dan varians diketahui. Dalam keadaan demikian, keluarga GLM merupakan asumsi paling lemah dari distribusi yang dapat dibuat, dalam hal tidak informasi yang dapat digunakan untuk menyatakan hubungan rata-rata dan varians ini.
- Persamaan QL untuk parameter rata-rata hanya menyangkut bentuk $\left(y_i - \mu_i\right)$, bukan dalam bentuk ordo yang lebih tinggi $\left(y_i - \mu_i\right)^d$, untuk d = 2, 3, ... Diantara persamaan pendugaan yang menyangkut persamaan QL $\left(y_i - \mu_i\right)$ yang optimal, yaitu dalam hal persamaan ini memberikan penduga varians minimum asimptotik (McCullag dan Nelder, 1983). Apabila terdapat keluarga GLM dengan hubungan rata-rata dan varians tertentu, maka penduga QL merupakan penduga yang efisien di bawah keluarga GLM (Pawitan, et al., 2006). Namun demikian, apabila distribusi sebenarnya bukan merupakan keluarga GLM, maka penduga QL efisiensinya menjadi berkurang.
- Terakhir, penduga likeliood maksimum dapat dikatakan untuk menggunakan seluruh informasi yang tersedia apabila model sebenarnya diketahui, sedangkan untuk hubungan rata-rata dan varians tertentu, penduga QL merupakan penduga yang paling robust dibandingkan dengan kekeliruan spesifikasi dari kemiringan (*skewness*).

Beberapa fungsi varians yang banyak digunakan dan model keluarga eksponensial yang bersesuaian disajikan pada Tabel 1.

2.1 Pendugaan Parameter

Algoritma untuk pendugaan parameter regresi QL dapat dinyatakan sebagai kuadrat terkecil terboboti iteratif (*iterative weighted least square*, IWLS). Pendugaan ini dapat diturunkan sebagai algoritma Gauss-Newton untuk menyelesaikan persamaan pendugaan (Pawitan, et al., 2006). Metode ini merupakan algoritma umum untuk menyelesaikan persamaan nonlinear. Akan diselesaikan persamaan:

$$\sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1}\left(y_i - \mu_i\right) = 0$$

Dengan cara melinearkan μ_i di sekitar nilai awal penduga $\beta^{(0)}$ dan mengevaluasi V_i pada nilai penduga awal. Misalkan $\eta_i = g\left(\mu_i\right) = x_i^T \beta$ merupakan skala prediktor linear. Kemudian

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} x_i$$

sehingga

$$\mu_i \approx \mu_i^{(0)} + \frac{\partial \mu_i}{\partial \beta}\left(\beta - \beta^{(0)}\right)$$

$$= \mu_i^{(0)} + \frac{\partial \mu_i}{\partial \eta_i} x_i^T \left(\beta - \beta^{(0)}\right)$$

dan

$$y_i - \mu_i = y_i - \mu_i^{(0)} - \frac{\partial \mu_i}{\partial \eta_i} x_i^T \left(\beta - \beta^{(0)}\right)$$

Dengan menempatkan persamaan di atas ke dalam persamaan pendugaan, akan diperoleh

$$\sum_i \frac{\partial \mu_i}{\partial \eta_i} V_i^{-1} x_i \left\{ y_i - \mu_i^{(0)} - \frac{\partial \mu_i}{\partial \eta_i} x_i^T \left(\beta - \beta^{(0)}\right)\right\} = 0$$

yang dapat digunakan untuk menyelesaikan β pada iterasi berikutnya yang diberikan oleh persamaan berikut:

$$\beta^{(1)} = \left(X^T \Sigma^{-1} X \right)^{-1} X^T \Sigma^{-1} z$$

dimana **X** merupakan matriks model dari variabel prediktor, **Σ** adalah matriks diagonal dengan unsur-unsur sebagai berikut:

$$\Sigma_{ii} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V_i$$

dimana $V_i = \phi V\left(\mu_i^{(0)}\right)$, dan z adalah variabel takbebas yang disesuaikan

$$z_i = x_i^T \beta^{(0)} + \frac{\partial \eta_i}{\partial \mu_i}\left(y_i - \mu_i^{(0)} \right)$$

Perlu dicatat bahwa parameter dispersi ϕ tidak digunakan dalam metode IWLS.

2.2 Pengujian Hipotesis

Pendekatan quasi-likelihood akan membawa pada dua buah penduga varians yang berbeda, yaitu:

- Rumus yang biasa dengan menggunakan matrisk Hessian. Rumusan ini menghasilkan penduga efisien pada saat spesifikasi hubungan rata-rata dan varians adalah benar.
- Rumus yang disebut sebagai 'rumus sandwich' yang menggunakan metode moment. Metode ini menghasilkan penduga yang robust, tanpa mengasumsikan spesifikasi hubungan rata-rata dan varians yang benar.

Dengan pendekatan QL, sepanjang fungsi rata-rata dinyatakan dengan benar, maka statistik skor quasi mempunyai rata-rata nol dan penduga β yang konsisten. Pemilihan hubungan rata-rata dan varians ini akan berpengaruh pada efisiensi dari penduga parameternya.

2.2.1 Asumsi varians yang benar

Diasumsikan variabel respons $y_1, \ldots, y_n$ yang saling bebas dengan rata-rata $\mu_1, \ldots, \mu_n$ dan varians $\text{var}(y_i) = \phi V(\mu_i) \equiv V_i(\beta, \phi)$, kemudian statistik skor-quasi $S(\beta) = \partial q / \partial \beta$ mempunyai rata-rata nol dan varians

$$\begin{aligned}
\text{var}(S) &= \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} \text{var}\left(y_i\right) V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T}(\beta) \\
&= \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} V_i V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T} \\
&= \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T} \\
&= X^T \Sigma^{-1} X
\end{aligned}$$

dimana **X** dan **Σ** didefinisikan sama dengan sebelumnya. Penduga likelihood biasa didasarkan pada nilai harapan matriks Hessian

$$-E\left(\frac{\partial^2 q}{\partial \beta \partial \beta^T} \right) \equiv D = X^T \Sigma^{-1} X$$

sehingga likelihood yang biasa akan menghasilkan varians dari skor adalah sama dengan nilai harapan turunan keduanya. Dengan menggunakan pendekatan ordo-pertama

$$S\left(\beta\right) \approx S\left(\hat{\beta}\right) - D\left(\beta - \hat{\beta}\right) = -D\left(\beta - \hat{\beta}\right)$$

Oleh karena $\hat{\beta}$ menyelesaikan persamaan $S\left(\hat{\beta}\right)=0$. Dengan demikian

$$\hat{\beta} \approx \beta + D^{-1}S\left(\beta\right)$$

dan

$$\text{var}\left(\hat{\beta}\right) \approx \left(X^T\Sigma^{-1}X\right)^{-} \quad (4)$$

Oleh karena $S(\beta)$ merupakan jumlah dari variat yang saling bebas, maka agar supaya dalil limit pusat terpenuhi sedemikian rupa sehingga pendekatan

$$\hat{\beta} \sim N\left(\beta,\left(X^T\Sigma^{-1}X\right)^{-1}\right)$$

2.2.2 Tidak Mengasumsikan Spesifikasi Varians Adalah Benar

Apabila kita spesifikasi varians tidak diasumsikan benar, maka akan diperoleh rumusan yang sedikit lebih rumit. Varians skor-quasi adalah

$$\begin{aligned}\text{var}(S) &= \sum_i \frac{\partial\mu_i}{\partial\beta} V_i^{-1}\,\text{var}\left(y_i\right)V_i^{-1}\frac{\partial\mu_i}{\partial\beta^T}\\ &= X^T\Sigma^{-1}\Sigma_z\Sigma^{-1}X\end{aligned}$$

dimana $\boldsymbol{\Sigma}_z$ adalah varians sebenarnya dari variabel z, dengan unsur-unsur sebagai berikut:

$$\Sigma_{z_{ii}} = \left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2 \text{var}\left(y_i\right)$$

Kerumitan terjadi karena $\boldsymbol{\Sigma}_z \neq \boldsymbol{\Sigma}$.

Diasumsikan dalam kondisi yang biasa, berdasarkan Pers. (4), maka penduga parameter $\hat{\beta}$ mendekati distribusi normal dengan rata-rata β dan varians

$$\text{var}\left(\hat{\beta}\right) = \left(X^T\Sigma^{-1}X\right)^{-1}\left(X^T\Sigma^{-1}\Sigma_z\Sigma^{-1}X\right)\left(X^T\Sigma^{-1}X\right)^{-1}$$

Rumusan di atas disebut juga rumus varians 'sandwich', yang secara asimptotoik adalah benar bahkan jika diasumsikan varians y_i adalah tidak benar. Dalam prakteknya kita dapat menduga $\boldsymbol{\Sigma}_z$ melalui matriks diagonal dengan unsur-unsur

$$\left(\varepsilon_i^*\right)^2 = \left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2\left(y_i-\mu_i\right)^2$$

Sehingga bentuk yang ditengah dari rumus varians dapat diduga oleh

$$X^T\Sigma^{-1}\Sigma_z\Sigma^{-1}X = \sum_i \frac{\left(\varepsilon_i^*\right)^2}{\Sigma_{ii}^2}x_i x_i^T$$

Penduga sandwich dan penduga biaa nilai akan mendekati pada saat spesifikasi varians adalah mendekati benar dan pada sampel yang berukuran besar.

3. BEBERAPA CONTOH PENGGUNAAN FUNGSI QUASI-LIKELIHOOD

3.1 Kasus Satu-Sampel

Pendugaan rata-rata populasi adalah masalah statistik yang mungkin paling sederhana. Misalkan diberikan sampel $y_1, \ldots, y_n$ yang saling bebas dan identik, dengan asumsi bahwa rata-rata dan varians masing-masing diberikan oleh $E(y_i) = \mu$ dan $\text{var}(y_i) = \sigma^2$. Dari Pers. (2) diketahui bahwa $\hat{\mu}$ adalah solusi dari

$$\sum_{i=1}^{n}\left(y_i-\mu\right)/\sigma^2=0$$

yang menghasilkan solusi $\hat{\mu}=\bar{y}$.

Distribusi-quasi dalam hal ini merupakan distribusi normal, sehingga untuk pendekatan QL bekerja dengan baik, kita perlu memerika apakah hal ini merupakan asumsi yang masuk akal atau tidak. Apabila misalnya distribusi dari data miring, maka kita lebih baik menggunakan suatu distribusi dengan fungsi varians yang berbeda. Alternatifnya, kita dapat memandang pendekatan yang sederhana sebagai pendekatan dari persamaan-pendugaan (*estimating-equation*). Contoh ini menujukkan antara kelebihan dan kekurangan dari persamaan-pendugaan dan QL dibandingkan dengan pendekatan likelihood biasa, sebagai berikut:

- Penduga adalah konsisten untuk kelas yang lebih luas untuk suatu distribusi yang mendasarinya, sebut saja sembarang distribusi dengan rata-rata μ. Dalam kenyataannya, variabel respons sampel tidaklah saling bebas.
- Kita mempunyai dasar inferens pada pertimbangan asimptotik, karena tidak ada inferens untuk sampel kecil. Dengan menggunakan pendekatan QL kita dapat menggunakan metode REML untuk meningkatkan kelayakan inferens.
- Terdapat kemungkinan kehilangan efisiensi dibandingkan dengan inferens untuk likelihood biasa pada saat distribusi sebenarnya bukan merupakan keluarga GLM.
- Bahkan jika distribusi sebenarya adalah simetris, rata-rata sampel adalah tidak robust terhadap data pencilan. Perlu dicatat bahwa banyak penduga robust yang diusulkan untuk data simetris tetapi mempunyai distribusi dengan ekor-panjang (miring) mungkin tidak robust terhadap kemiringan data.
- Tidak ada preskripsi baku untuk menduga parameter varians σ^2. Beberapa teorema lainnya diperlukan, misalnya penggunaan penduga metode moment:

  $$\sigma^2=\frac{1}{n}\sum_i\left(y_i-\bar{y}\right)^2$$

  Tetap tidak membuat asumsi mengenai distribusinya. Persamaan pendugaan dapat diperluas dengan menyertakan parameter dispersi. Metode moment dapat mendapat kesulitan dalam model-model semi-parametrik, sedangkan pendekatan likelihood tidak.

3.2 Model Linear

Diberikan sampel saling bebas $\left(y_i,x_i\right)$, untuk i = 1, 2, ..., n. Misalkan

$$E\left(y_i\right)=x_i^T\beta\equiv\mu_i\left(\beta\right)$$

$$\mathrm{var}\left(y_i\right)=\sigma_i^2\equiv V_i\left(\beta\right)=V_i$$

Persamaan pendugaan untuk β adalah

$$\sum_i x_i\sigma_i^{-2}\left(y_i-x_i^T\beta\right)=0$$

Yang akan memberikan suatu penduga kuadrat terkecil terboboti

$$\hat{\beta}=\left(\sum_i x_i x_i^T/\sigma_i^2\right)^{-1}\sum_i x_i y_i/\sigma_i^2$$

$$=\left(X^TV^{-1}X\right)^{-1}X^TV^{-1}y$$

dimana **X** adalah matriks model berukuran $n\times p$, **V** adalah matriks varians $\mathrm{diag}\left[\sigma_i^2\right]$, dan ***y*** adalah vektor respons.

3.3 Regresi Poisson

Untuk data cacahan saling bebas y_i dengan variabel prediktor x_i, misalkan diasumsikan bahwa:

$$E\left(y_i\right)=\mu_i=\exp\left(x_i^T\beta\right)$$

$$\text{var}\left(y_i\right)=\phi\mu_i=V_i\left(\beta,\phi\right)$$

Persamaan pendugaan untuk β adalah

$$\sum_{i=1}^{n}\exp\left(x_i^T\beta\right)x_i\exp\left(-x_i^T\beta\right)\left(y_i-\mu_i\right)/\phi=0$$

atau

$$\sum_{i=1}^{n}x_i\left(y_i-\exp\left(x_i^T\beta\right)\right)=0$$

Persamaan pendugaan di sini betul-betul merupakan persamaan skor di bawah model Poisson. Aspek yang menarik dari pendekatan QL ini adalah kita dapat menggunakan model ini bahkan untuk respons yang kontinu, sepanjang varians dapat dimodelkan sebagai proporsional terhadap rata-ratanya.

Pernyataan di atas mempunyai dua buh interpretasi. Pertama, diantara keluarga distribusi yang memenuhi hubungan rata-rata dan varians Poisson, penduga berdasarkan quasi-likelihood Poisson adalah robust terhadap asumsi distribusinya. Kedua, metode persamaan pendugaan adalah efisien, jika memang distribusinya adalah Poisson.

3.4 Model dengan Koefisien Variasi Konstan

Misalkan y_1, ..., y_n adalah respons yang saling bebas dengan rata-rata dan varians masing-masing adalah

$$E\left(y_i\right)=\mu_i=\exp\left(x_i^T\beta\right)$$

$$\text{var}\left(y_i\right)=\phi\mu_i^2=V_i\left(\beta,\phi\right)$$

Persamaan pendugaan untuk β adalah

$$\sum_{i=1}^{n}x_i\left(y_i-\mu_i\right)/\left(\phi\exp\left(x_i^T\beta\right)\right)=0$$

Model di atas dilatarbelakangi dengan cara mengasumsikan bahwa respons mempunyai distribusi gamma, tetapi model ini dapat diterapkan untuk setiap respons dimana koefisien variasi yang mendekati konstan. Metode ini betul-betul efisien apabila model sebenarnya adalah gamma diantara keluarga distribusi yang mempunyai koefisien variasi konstan.

4. CONTOH NUMERIK

Untuk lebih memudahkan dalam memahami hasil analisis, maka dalam makalah ini akan langsung diterapkan pada suatu data mengenai tanaman parasit tanpa klorofil yang tumbuh akar tanaman yang sedang berbungan dan dikenal sebagai orobanche (Collet, 1991). Dalam makalah ini akan diamati tentang faktor-faktor yang mempengaruhi munculnya kecambah dari benih dua jenis variaetas orobanche, yaitu variaetas orobanche 75 dan varietas orobanche 73. Masing-masing orobanche itu diamati pada dua jenis tanaman yang berbeda, yaitu tanaman buncil dan ketimun. Banyaknya benih yang berkecambah kemudian dicatat, dan hasilnya disajikan dalam Tabel 2. Dalam hal ini peristiwa 'sukses' adalah benih dari tanaman buncis dan ketimun yang dapat berkecambah, sedangkan peristiwa 'gagal' adalah benih-benih yang tidak berkecambah.

Dalam makalah ini jenis varietas orobanche dijadikan sebagai peubah penjelas x_1, jenis tanaman (buncis dan ketimun) merupakan peubah x_2. Banyaknya benih yang diamati pada gerombol ke-i dinotasikan dengan n_i, sedangkan banyaknya benih yang berkecambah pada gerombol ke-i dinotasikan sebagai y_i. Perlu dicatat bahwa banyaknya gerombol (*batch*) untuk setiap kombinasi yang digunakan dalam percobaan ini masing-masing berbeda, mulai dari 4 sampai dengan 81. Proporsi tersebut yang berdasarkan pada batch yang lebih besar akan mempunyai presisi yang lebih besar pula. Hal ini akan sangat penting apabila digunakan untuk memperhitungkan besarnya peluang suatu benih untuk berkecambah.

Tabel 2. Banyaknya Benih Orobanche yang Berkecambah, *y*, dari *n* Benih yang Diamati dari Akar Tanaman Buncis dan Ketimun

Orobanche 75				**Orobanche 73**			
Buncis		**Ketimun**		**Buncis**		**Ketimun**	
y	**n**	**Y**	**n**	**y**	**n**	**y**	**n**
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

Sumber: Collet, D. (1991). *Modelling Binary Data*. London: Chapman and Hall.

Informasi mengenai distribusi dari banyaknya peristiwa 'sukses' (benih yang berkecambah) adalah tidak diketahui dengan pasti. Responsnya bukan merupakan proporsi binoial, karena hal tersebut tidak mewakili rasio dari frekuensi benih yang berkecambah dari percobaan Bernoulli. Akan tetapi oleh karena rata-rata proporsi μ_{ij} untuk varietas ke-*i* pada jenis tanaman ke-*j* akan berada dalam interval [0, 1], maka kita dapat memperlakukan variabel proporsi, *p*, sebagai variabel 'pseudo-binomial:

$$E\left[p_{ij} \right] = \mu_{ij}$$

$$\mu_{ij} = 1/\left(1+\exp\left\{-\eta_{ij}\right\}\right)$$

$$\eta_{ij} = \beta_0 + \beta_{1i} + \beta_{2j}$$

$$\text{var}\left[p_{ij} \right] = \phi\mu_{ij}\left(1-\mu_{ij}\right)$$

Dalam hal ini ηij adalah prediktor linear untuk varietas ke-*i* pada jenis tanaman ke-*j*, *i* menyatakan efek dari varietas ke-*i* dan *j* menyatakan efek jenis tanaman ke-*j*. Logit dari nilai harapan proporsi banyaknya benih berkecambah secara linear dipengaruhi oleh efek-efek tersebut. Fungsi varians dari model adalah berdistribusi binomial(n, μ_{ij}), dan ϕ merupakan parameter overdispersinya. Model ini dapat dianalisis dengan menggunakan prosedur GLIMMIX dalam sistem SAS dengan mengikuti pernyataan berikut:

```
proc glimmix data=kecambah;
       class x1 x2;
       model p = x1 x2/ link=logit dist=binomial;
       random _residual_;
       lsmeans x1 / diff=control('1');
run;
```

Tabel 3 dan 4 masing-masing menampilkan hasil-hasil mengenai statistik kecocokan model dan efek dari masing-masing prediktor yang dianalisis. Statistik kecocokan model yang digunakan ada beberapa diantaranya adalah -2 log-likelihood, AIC, dan statistik chi-kuadrat Pearson beserta rasio dengan derajat bebasnya. Sedangkan untuk menguji efek dari masing-masing prediktor digunakan statistik uji F Tipe III.

Tabel 3. Statistik Kecocokan Model

Ukuran statistik	**Nilai dari likelihood biasa**	**Nilai dari quasi-likelihood**
-2 log-likelihood	1092.629	27.20
AIC	1098.629	33.20
Chi-kuadrat Pearson	38.3106	1.78
Chi-kuadrat Pearson / db	2.1284	0.10

Sebagai bahan perbandingan pada Tabel 3 disajikan pula hasil-hasil analisis fungsi likelihood biasa dengan menggunakan prosedur GENMOD. Dari keempat ukuran statistik kecocokan model yang diperhatikan terlihat bahwa terdapat penurunan yang cukup besar antara hasi dari fungsi likelihood biasa dengan hasil yang diberikan dari pendekatan QL. Indikasi adanya masalah overdispersi dalam data ditunjukkan oleh rasio chi-kuadrat Pearson dengan derajat bebasnya untuk model likelihood biasa, yaitu sebesar 2.1284. Kemudian setelah data dianalisis dengan menggunakan pendekatan fungsi QL, maka rasio antara chi-kuadrat Pearson dengan derajat bebasnya menjadi kurang dari satu.

Tabel 4. Tes Type III untuk Uji Efek Prediktor

Efek	**Likelihood Biasa**		**Quasi-likelihood**	
	Nilai chi-kuadrat	**p-value**	**Nilai-F**	**p-value**
X1	3.06	0.0800	5.07	0.0371
X2	56.49	0.0001	12.99	0.0020

Sedangkan hasil pengujian efek untuk setiap prediktor yang diamati, terlihat pula bahwa terdapat hasil pengujian yang berbeda antara model likelihood biasa dengan model pendekatan fungsi QL. Di bawah taraf signifikansi 5%, hasil pengujian dari model fungsi likelihood biasa menunjukkan bahwa efek dari variabel x_1 adalah tidak signifikan secara statistik, sedangkan menurut pendekatan fungsi QL memberikan hasil pengujian yang signifikan secara statistik. Sedangkan efek dari variabel x_2 adalah signifikan secara statistik, baik menurut model likelihood biasa maupun menurut pendekatan fungsi QL. Perlu dicatat bahwa efek variabel prediktor dari model likelihood biasa dihitung dengan menggunakan statistik chi-kuadrat, sedangkan model pendekatan fungsi QL dihitung dengan menggunakan statistik-F.

Tabel 5 menampilkan rata-rata kuadrat terkecil untuk analisis data tersebut. Dimisalkan kita ingin melihat efek dari masing-masing kategori dari variabel x_2, yaitu jenis tanaman. Artinya untuk melihat bagaimana efek jenis tanaman 1 (buncis) dan jenis tanaman 2 (ketimun) terhadap munculnya kecambah. Nilai-nilai dari efek ini diperoleh dengan cara merata-ratakan

$$\text{logit}\left(\hat{\mu}_{ij}\right) = \hat{\eta}_{ij}$$

menurut kategori jenis varietas. Dengan kata lain, rata-rata kuadrat terkecil dihitung pada skala dimana efek dari modelnya adalah aditif. Perlu dicatat bahwa rata-rata kuadrat terkecil diurutkan menurut jenis tanaman. Penduga dari nilai harapan proporsi banyaknya yang berkecambah adalah:

$$\hat{\mu}_{.1} = \frac{1}{1+\exp(-0.6407)} = 0.6549$$

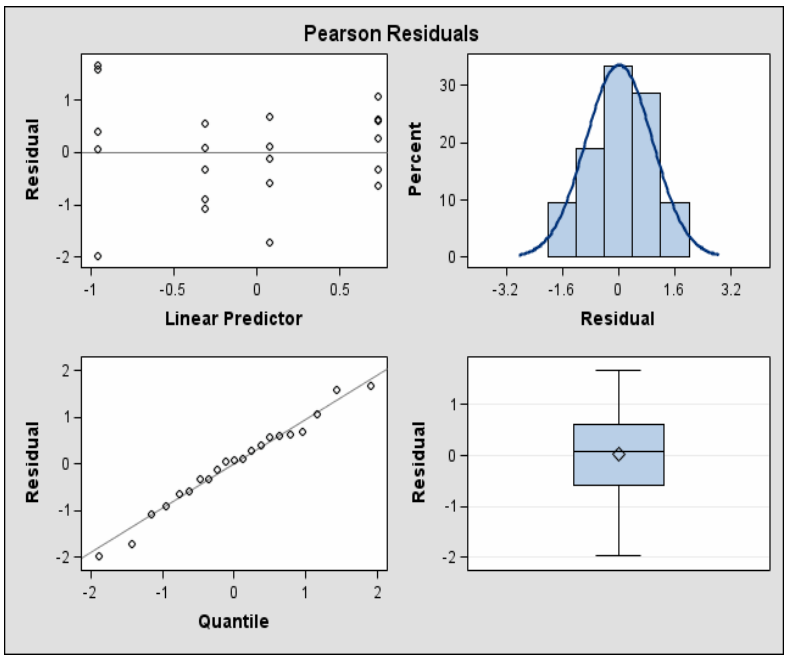
dan

$$\hat{\mu}_{.2} = \frac{1}{1+\exp(0.3996)} = 0.4014$$

Tabel 5. Rata-rata Kuadrat Terkecil untuk Variabel x_2

Kategori	**Penduga**	**Galat baku**	**db**	**Nilai-t**	**p-value**
Buncis	-0.6407	0.2118	18	-3.03	0.0073
Ketimun	0.3996	0.1964	18	2.03	0.0569

Melalui prosedur GLIMMIX dalam sistem SAS, kita juga dapat melakukan diagnosa mengenai asumsi distribusi dengan cara menentukan berbagai ukuran diagnostik secara grafik, dimana analisis residu yang dihitung adalah berdasarkan residu Pearson. Hasil diagnosa ini disajikan pada Gambar 1. Dari Gambar 1 terlihat bahwa pemilihan fungsi varians untuk menganalisis data ini adalah sudah tepat, yaitu menggunakan $V(\mu) = \mu(1 - \mu)$.



Gambar 1. Analisis Residu Pearson

5. KESIMPULAN

Pendekatan quasi-likelihood pertama kali dibentuk untuk mengatasi dua permasalahan pertama di atas dan mempunyai sifat-sifat yang perlu dicatat. Pertama, berbeda dengan pendekatan fungsi likelihood biasa, dalam fungsi quasi-likelihood tidak menyatakan struktur peluang tertentu, tetapi hanya memerlukan asumsi mengenai dua buah moment pertama. Hal ini pendekatan fungsi quasi-likelihood mempunyai fleksibilitas yang tinggi. Kedua, pendugaan untuk parameter regresi hanya untuk rata-rata. Untuk pendekatan berdasarkan likelihood untuk pendugaan parameter dispersi ϕ diperlukan beberapa teorema tambahan.

Wedderburn (1974) menurunkan beberapa sifat dari quasi-likelihood, tapi perlu dicatat bahwa teori ini mengasumsikan bahwa ϕ diketahui. Dengan asumsi ini terlihat bahwa quasi-likelihood merupakan log-likelihood sebenarnya jika dan hanya jika respons y_i berasal dari model keluarga eksponensial dengan parameter-satu (keluarga GLM dengan $\phi = 1$) dengan densitas-log.

Model pendekatan quasi-likelihood yang dibahas dalam makalah ini adalah untuk memodelkan data yang saling bebas, sedangkan menurut McCullagh dan Nelder (1983) model pendekatan quasi-likelihood dapat juga digunakan untuk memodelkan data yang tidak saling bebas. Sedangkan menurut Pawitan, et al. (2006) model pendekatan quasi-likelihood dapat juga dilakukan untuk membentuk model dispersi, model GLM bersama antara rata-rata dan dispersi, serta model GLM bersama untuk peningkatan kualitas model. Sementara itu, implementasi pendekatan quasi-likelihood dalam sistem SAS melalui prosedur GLIMMIX baru bisa akan bekerja pada sistem SAS versi 9.1 ke atas, artinya sistem SAS di bawah versi tersebut tidak akan bekerja dengan 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]. Jørgensen, B. (1987). Exponential dispersion models (with discussion). Journal of the Royal Statistical Society B, 49, 127-162.
[10]. Lawal, B. (2003) Categorical Data Analysis With SAS And SPSS Applications. London: Lawrence Erlbaum Associates.
[11]. Lee, Y., Nelder, J.A, and Pawitan, Y. (2006) Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. London: Chapman and Hall/CRC-Press.
[12]. McCullagh, P., and J.A. Nelder (1983). Generalized Linear Models. Second Edtion. New York: Chapman and Hall.
[13]. Myers, R.H. (1990). Classical and Modern Regression With Applications. Boston: PWS-KENT Publishing Company.
[14]. Nelder, J.A., and R.W.M. Wedderbun. (1972). Generalized Linear Models. Journal of Royal Statistical Society, Series A 153: 370-384.
[15]. Santner, T.J., and D.E. Duffy. (1989). The Statistical Analysis of Discrete Data. New York: Springer-Verlag.
[16]. Uusipaikka, E. (2009). Confidence Intervals in Generalized Regression Models. London: Chapman and Hall.
[17]. Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika, 61, 439-447.