Statistika, Vol. 12 No. 1, 9 – 18
Mei 2012

Penaksiran Parameter Model Regresi Beta untuk Memodelkan Data Proporsi

NUSAR HAJARISMAN

Jurusan Statistika, Universitas Islam Bandung
Jl. Purnawarman No. 63 Bandung 40116, Jawa Barat, Indonesia
nrisman@yahoo.co.uk

ABSTRAK.

Suatu variabel respons yang berbentuk proporsi yang nilainya ada dalam selang terbuka (0, 1) dapat dihubungkan dengan sejumlah variabel prediktor melalui model regresi beta. Model ini merupakan bagian dari model linear umum (generalized linear model, GLM) dimana variabel respons ini adalah megikuti distribusi beta yang merupakan anggota dari keluarga eksponensial. Parameter regresi dari model regresi beta dapat diinterpretasikan dalam bentuk rata-rata dari respons, dan ketika menggunakan fungsi hubung logit, maka parameter regresi ini diinterpretasikan sebagai odds rasio. Penaksiran parameter model menggunakan metode kemungkinan maksimum, dimana proses penaksirannya harus diselesaikan secara numerik. Dalam makalah ini akan dibahas mengenai penaksiran parameter model regresi beta melalui metode penskoran Fisher berdasarkan pada vektor skor dan matriks informasi Fisher. Terakhir, akan dibahas pula contoh penerapan dari pemodelan data proporsi melalui regresi beta ini.

Kata Kunci: distribusi beta, model linear umum, metode kemungkinan maksimum, odds rasio, fungsi logit, vektor skor, matriks informasi Fisher, dan proporsi.

1. PENDAHULUAN

Model regresi klasik yang mengasumsikan bahwa galat berdistribusi normal merupakan suatu model yang paling banyak diterapkan di berbagai bidang. Akan tetapi, penggunaan model regresi klasik ini tidak dapat sepenuhnya dapat diterapkan pada bidang penelitian tertentu. Misalnya, dalam penelitian tentang bioassay, biasanya variabel responnya bisa bervariasi dengan kovariat berbentuk dosis. Misalkan terdapat sekelompok hewan yang diamati (dalam hal ini adalah kumbang), dimana n_i menyatakan banyaknya kumbang yang diamati pada kelompok ke-i, dan y_i menyatakan banyaknya kumbang yang mati setelah diberi perlakuan. semacam zat carbon disulphide selama lima jam dengan berbagai macam konsentrasi (Dobson, 1983). Selanjutnya kita akan memodelkan p_i yang merupakan peluang matinya kumbang sebagai fungsi dari x_i atau dosis zat carbon disulphide. Masalah ini biasanya akan dimodelkan dengan cara meregresikan p_i pada x_i dengan menggunakan metode kuadrat terkecil biasa (*ordinary least square*, OLS). Tentu saja hal ini bukan merupakan solusi yang tepat untuk menyelesaikan masalah tersebut, dikarenakan oleh dua alasan, yaitu masalah non-linearitas, dimana model regresi linear biasa akan memberikan nilai taksiran p_i di luar wilayah (0, 1), serta masalah heteroskedastisitas, dimana varians dari p_i adalah tidak konstan.

Untuk mengatasi masalah tersebut biasanya diselesaikan dengan cara menggunakan pendekatan kemungkinan maksimum berdasarkan pada fungsi kemungkinan dari distribusi binomial sebagai dasar pada pemodelan regresi logistik. Namun selain menggunakan pendekatan pemodelan regresi logistik, pendekatan lain yang dapat digunakan adalah melalui pemodelan regresi beta. Model yang diusulkan dalam makalah ini tentu saja berdasarkan asumsi bahwa variabel respons mengikuti distribusi beta. Distribusi beta dikenal sebagai distribusi yang cukup fleksibel dalam memodelkan data yang responsnya berbentuk proporsi, karena densitasnya mempunyai pola yang berbeda bergantung pada nilai-nilai dari parameter yang ada dalam distribusi beta ini. Fungsi densitas dari variabel acak yang mengikuti distribusi beta diberikan oleh

$$f(y;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1},\quad 0<y<1 \qquad \ldots (1)$$

dimana $a > 0$, $b > 0$, dan $\Gamma(.)$ merupakan fungsi gamma. Rata-rata dan varians bagi y masing-masing diberikan oleh:

$$E(y) = \frac{a}{a+b} \quad \ldots (2)$$

dan

$$\text{var}(y) = \frac{ab}{(a+b)^2(a+b+1)} \quad \ldots (3)$$

Distribusi beta sangat fleksibel dan berbagai fenomena ketidakpastian dapat dimodelkan dengan menggunakan distribusi beta ini. Fleksibilitas ini mendorong berkembangnya penggunaan distribusi beta secara empiris dalam berbagai bidang aplikasi. Beberapa aplikasi dari distribusi beta dibahas oleh Johnson et al. (1995). Namun demikian apa yang dibahas dalam Johnson et al (1995) masih belum digunakan dalam menganalisis hubungan fungsional antara sejumlah variabel prediktor dengan satu variabel respons, dimana respons ini mengikuti distribusi beta. Selain itu, menurut Swearingen et al (2011) model regresi beta merupakan model yang memberikan penaksir parameter yang akurat dan efisien dibandingkan dengan metode kuadrat terkecil biasa, ketika variabel respons yang diamati distribusinya tidak simetris, atau pada saat terjadi masalah heteroskedastisitas. Oleh karena itu dalam makalah ini, akan dibahas mengenai pemodelan data proporsi dengan menggunakan model regresi beta.

Distribusi beta dicirikan oleh dua buah parameter bentuk, dimana melalui transformasi aljabar sederhana dari kedua parameter dapat ditunjukkan bahwa parameter dalam distribusi beta merupakan parameter rata-rata dan parameter presisi. Dengan cara tersebut, model regresi beta dapat memberikan penaksir parameter yang berhubungan dengan perubahan dalam rata-rata dan dispersi dari variabel respons, sekaligus dapat membuat inferensi mengenai hal tersebut. Dalam makalah ini juga akan dibahas mengenai inferensi untuk sampel besar yang didasarkan pada metode kemungkinan maksimum.

Pemodelan dan prosedur inferensi yang akan dibahas d sini akan lebih banyak mengadopsi dari konsep model linear umum yang dikembangkan oleh McCullagh dan Nelder (1989). Masalah komputasi untuk memodelkan regresi Beta ini akan menggunakan prosedur PROC NLMIXED dalam sistem SAS (SAS Institute, 2005) atau dapat juga menggunakan **R** (Cribari-Neto dan Zeileis, 2010). Terakhir, akan dibahas juga contoh numerik dari penggunaan model regresi beta untuk memodelkan data respons yang berbentuk proporsi.

2. MODEL REGRESI BETA

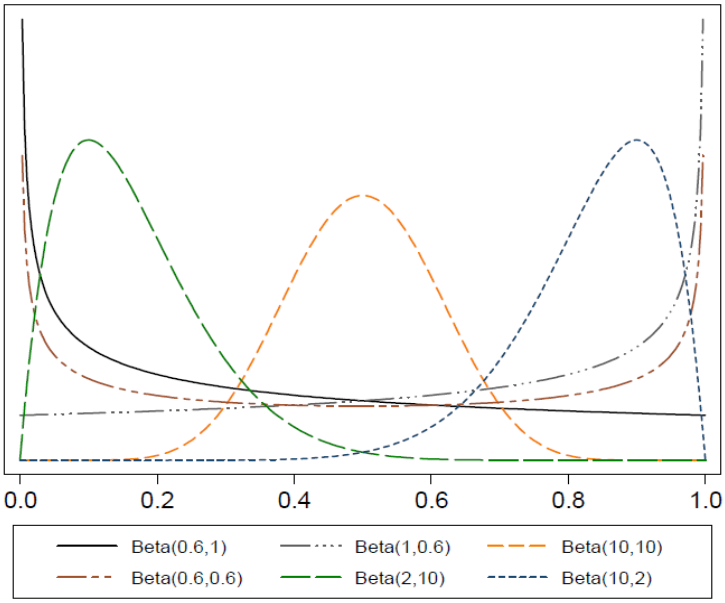
Pada bagian ini dibahas mengenai suatu model regresi untuk respons yang mengikuti distribusi beta. Fungsi densitas dari distribusi beta diberikan dalam Pers. (1), dengan parameter a dan b. Akan tetapi untuk keperluan pemodelan, maka biasanya akan sangat berguna apabila memodelkan rata-rata dari variabel respons. Selain itu pula biasanya perlu mendefinisikan model sedemikian rupa sehingga model ini berisi suatu parameter dispersi. Untuk membentuk model regresi beta dengan menyertakan rata-rata respons bersamaan dengan parameter dispersinya, maka perlu dilakukan reparameterisasi dari fungsi densitas beta. Misalkan $\mu = a/(a+b)$ dan $\kappa = a+b$, sehingga diperoleh bahwa $a = \mu\kappa$ dan $b = (1-\mu)\kappa$. Dengan demikian, berdasarkan Pers. (2) dan (3) diketahui bahwa

$$E(y) = \mu \quad \ldots (4)$$

dan

$$\text{var}(y) = \frac{\mu(1-\mu)}{\kappa+1} \quad \ldots (5)$$

dimana μ adalah rata-rata dari variabel respons dan κ dapat diinterpretasikan sebagai parameter presisi, dalam arti bahwa untuk μ tertentu, maka nilai dari κ lebih besar akan memberikan varians bagi variabel respons yang lebih kecil.



Gambar 1. Contoh distribusi beta untuk berbagai nilai parameter (μ, κ)

Setelah dilakukan reparameterisasi tersebut, maka fungsi dengan untuk variabel acak y yang mengikuti distribusi beta adalah sebagai berikut:

$$f\left(y;\mu,\kappa\right)=\frac{\Gamma\left(\kappa\right)}{\Gamma\left(\mu\kappa\right)\Gamma\left(\left(1-\mu\right)\kappa\right)}y^{\mu\kappa-1}\left(1-y\right)^{(1-\mu)\kappa-1},\quad 0<y<1 \qquad \ldots (5)$$

Parameterisasi semacam ini menyatakan bahwa $0< \mu < 1$ dan $\theta > 1$, dan juga dapat dengan mudah ditunjukkan bahwa $a = \mu\kappa > 0$ dan $b = \kappa(1 - \mu) > 0$. Gambar 1 menampilkan berbagai densitas beta yang berbeda menurut nilai-nilai dari parameter (μ, κ). Dari gambar tersebut dapat dilihat bahwa densitasnya akan memiliki pola yang berbeda bergantung pada nilai-nilai dari kedua parameter tersebut. Secara khusus dapat dilihat bahwa ketika $\mu = 0.5$, maka akan menampilkan bentuk densitas yang simetris, dan bentuk densitas yang tidak simetris ketika $\mu \neq 0.5$. Sebagai tambahan, dispersi dari distribusi untuk μ tertentu akan turun sebagaimana jika κ meningkat.

2.1 Pembentukan Model Regresi Beta

Untuk membentuk model regresi beta dilakukan melalui pendekatan model linear umum, dalam hal ini akan digunakan dua buah fungsi hubung. Satu fungsi hubung digunakan untuk parameter lokasi μ, dan fungsi hubung lainnya digunakan untuk parameter dispersi κ. Menurut Smithson dan Verkuilen (2005) bahwa fungsi ini merupakan fungsi non linear, halus (*smooth*), dan monoton yang memetakan dari ruang yang tidak berbatas (*unbounded*) dari prediktor linear ke dalam ruang sampel yang diamati, dalam hal ini terbatas pada interval terbuka (0, 1). Misalkan $\mathbf{X}$ dan $\mathbf{W}$ merupakan matriks kovariat (bisa identik), dengan $\mathbf{x}_i$ dan $\mathbf{w}_i$ merupakan vektor baris ke-i dari kedua matriks tersebut. Dimisalkan pula $\boldsymbol{\beta}$ dan $\boldsymbol{\delta}$ masing-masing adalah vektor koefisien regresi beta. Model linear umum untuk parameter lokasi adalah

$$g\left(\mu_i\right)=\mathbf{x}_i^T\boldsymbol{\beta}$$

dimana $g(\cdot)$ adalah fungsi monoton, suatu fungsi hubung yang mempunyai turunan. Bentuk umum yang menyatakan hubungan antara rata-rata dan varians adalah

$$\sigma_i^2=v\left(\mu_i\right)u\left(\kappa_i\right)$$

dimana v dan u masing-masing merupakan fungsi yang bersifat non-negatif. Parameter presisi κ_i diasumsikan sebagai suatu bentuk yang dapat dimodelkan sebagai

$$h(\kappa_i) = \mathbf{w}_i \delta$$

dimana h merupakan fungsi hubung yang lain. Mengikuti konsep yang diajukan oleh Smith (1989), diketahui bahwa fungsi likelihood bagi $\boldsymbol{\beta}$ ketika δ konstan akan terdefinisikan submodel lokasi. Sedangkan jika fungsi likelihood bagi $\boldsymbol{\delta}$ ketika β dalah konstan, maka akan terbentuk submodel dispersi.

Untuk variabel respons yang berdistribusi beta, maka rata-ratanya harus berada dalam selang terbuka, sehingga diperlukan suatu fungsi hubung yang dapat memenuhi kondisi tersebut. Salah satu pilihan fungsi hubung yang dapat digunakan adalah fungsi hubung logit, karena fungsi hubung ini mampu memetakan $\mu \in (0, 1)$ ke dalam ruang sampel yang sesuai dengan distribusinya. Fungsi hubung logit ini juga biasa digunakan sebagai fungsi hubung dalam model regresi logistik. Dengan demikian, model regresi beta dapat dikatakan sebagai bentuk umum dari model regresi logistik ketika variabel respons yang diamati berbentuk proporsi. Terdapat beberapa fungsi hubung lainnya yang dapat digunaka, seperti log-log komplementer atau probit. Collet (2003) telah membahas banyak tentang fungsi hubung untuk data biner, dimana hampir semuanya dapat digeneralisasi pada pemodelan regresi beta.

Selanjutnya, parameter presisi κ harus bernilai positif sebab varians tidak dapat bernilai negatif. Fungsi hubung log dapat memenuhi sifat tersebut, yaitu

$$\log(\kappa_i) = -\mathbf{w}_i \delta$$

Di sini menggunakan tanda negatif untuk membuat interpretasi mengenai koefisien δ menjadi lebih mudah. Oleh karena κ merupakan parameter presisi, maka suatu δ yang bernilai positif mengindikasikan varians yang lebih kecil, dan hal ini tentu saja menjadi sulit dalam interpretasinya. Akan lebih masuk akal untuk memodelkan dispersi daripada memodelkan presisi, sehingga bentuk di atas perlu diberi tanda negatif.

2.2 Penaksiran Parameter

Misalkan $y_1, \ldots, y_n$ adalah variabel acak saling bebas, dimana untuk setiap y_i, untuk $i = 1, \ldots, n$ mengikuti densitas yang diberikan dalam Pers. (5) dengan rata-rata μ_t dan parameter dispersi κ yang tidak diketahui. Model diperoleh dengan cara mengasumsikan bahwa rata-rata y_t dapat ditulis sebagai

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i \qquad \ldots (6)$$

dimana $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ adalah vektor dari parameter regresi, dan $x_{i1}, \ldots, x_{ik}$ merupakan data pengamatan pada k buah kovariat, untuk $k < n$, yang diasumsikan tetap (*fixed*) dan diketahui. Perlu dicatat bahwa varians dari respons y merupakan fungsi dari μ_t, dan akibatnya juga merupakan fungsi dari nilai kovariatnya. Dengan demikian, varians yang tidak konstan secara tidak langsung akan diakomodasikan ke dalam model (Ferrari dan Neto, 2004). Perhatikan bahwa parameter rata-rata dibatasi pada selang terbuka (0, 1), sehingga diperlukan suatu fungsi hubung yang akan memetakan parameter dari interval ke dalam ruang bilangan nyata.

Terdapat beberapa pilihan fungsi hubung yang dapat digunakan, misalnya menggunakan fungsi hubung logit $g(\mu) = \log\{\mu/(1-\mu)\}$, fungsi hubung probit $g(\mu) = \Phi^{-1}(\mu)$, dimana $\Phi(.)$ merupakan fungsi distribusi kumulatif dari variabel acak normal baku, serta fungsi hubung log-log komplementer $g(\mu) = \log\{-\log(1-\mu)\}$. Ketiga fungsi hubung tersebut biasa digunakan untuk menganalisis data biner melalui model regresi logistik (McCullagh dan Nelder, 1989). Lebih jauh, dari ketiga fungsi hubung tersebut, fungsi hubung logit merupakan fungsi hubung kanonik dan akan mengembalikan penaksir parameter ke dalam bentuk log odds rasio. Fungsi hubung logit $g(\mu)$ didefinisikan sebagai

$$g(\mu) = \text{logit}(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{x}_i^T\boldsymbol{\beta} \rightarrow \mu = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})} \qquad \ldots (7)$$

Dalam hal ini, parameter regresi mempunyai interpretasi yang penting. Misalkan bahwa nilai dari prediktor ke-i meningkat sebesar c unit, dan variabel prediktor lainnya dianggap tidak berubah, serta μ^* merupakan rata-rata dari variabel y di bawah suatu nilai kovariat yang baru, sedangkan μ menyatakan rata-rata y di bawah nilai kovariat yang asli, maka dapat ditunjukkan bahwa

$$\exp\left(c\beta_j\right)=\frac{\mu^*/\left(1-\mu^*\right)}{\mu/\left(1-\mu\right)} \quad \text{... (8)}$$

Artinya, $\exp(c\beta_j)$ adalah sama dengan odds rasio, sama dengan interpretasi dalam model regresi logistik (Collet, 2003).

Fungsi log-likelihood berdasarkan pada sampel dari n buah pengamatan yang saling bebas diberikan oleh

$$l\left(\beta,\kappa\right)=\sum_{i=1}^{n}l\left(\mu_i,\kappa\right) \quad \text{... (9)}$$

dimana

$$\begin{aligned} l\left(\mu_i,\kappa\right)&=\log\Gamma\left(\kappa\right)-\log\Gamma\left(\mu_i\kappa\right)-\log\Gamma\left(\left(1-\mu_i\right)\kappa\right)+\left(\mu_i\kappa-1\right)\log y_i \\ &\quad+\left\{\left(1-\mu_i\right)\kappa\right\}\log\left(1-y_i\right) \end{aligned} \quad \text{...(10)}$$

Misalkan $z_i=\log\left\{y_i/\left(1-y_i\right)\right\}$ dan $\mu_i^*=\psi\left(\mu_i\kappa\right)-\psi\left(\left(1-\mu_i\right)\kappa\right)$. Kemudian, vektor skor yang merupakan turunan pertama dari fungsi log-likelihood terhadap parameter $\boldsymbol{\beta}$ dan $\boldsymbol{\kappa}$, diberikan oleh

$$\begin{aligned} U\left(\beta,\kappa\right)=\frac{\partial l\left(\mu_i,\kappa\right)}{\partial\theta}&=\begin{pmatrix}\dfrac{\partial l\left(\mu_i,\kappa\right)}{\partial\beta}\\ \dfrac{\partial l\left(\mu_i,\kappa\right)}{\partial\kappa}\end{pmatrix} \\ &=\begin{pmatrix}\kappa X^T W\left(z_i-\mu_i^*\right)\\ \sum_{i=1}^{n}\left\{\mu_t\left(z_i-\mu_i^*\right)+\log\left(1-y_i\right)-\psi\left(\left(1-\mu_i\right)\kappa\right)+\psi\left(\kappa\right)\right\}\end{pmatrix} \end{aligned} \quad \text{... (11)}$$

dimana $\mathbf{X}$ adalah matriks data dari variabel prediktor berukuran $n \times k$ dengan unsur baris ke-t adalah x_i^{T} dan $\mathbf{W}=\operatorname{diag}\left(1/g'\left(\mu_1\right),....,1/g'\left(\mu_n\right)\right)$, dan vektor parameter $\boldsymbol{\theta}=(\beta,\kappa)$.

Tahap berikutnya adalah menentukan matriks informasi Fisher. Misalkan diketahui matriks $\mathbf{W}$ = diag$\{w_1, \ldots, w_n\}$, dengan unsur-unsur

$$w_i=\kappa\left\{\psi'\left(\mu_i\kappa\right)+\psi'\left(\left(1-\mu_i\right)\kappa\right)\right\}\frac{1}{\left\{g'\left(\mu_i\right)\right\}^2}$$

$c=\left(c_1,...,c_n\right)^T$, dengan $c_i=\kappa\left\{\psi'\left(\mu_i\kappa\right)\mu_i-\psi'\left(\left(1-\mu_i\right)\kappa\right)\left(1-\mu_i\right)\right\}$, dimana $\psi'(\square)$ merupakan fungsi trigamma. Dimisalkan pula $\mathbf{D}$ = diag$\{d_1, \ldots, d_n\}$, dengan

$$d_i=\psi'\left(\mu_i\kappa\right)\mu_i^2+\psi'\left(\left(1-\mu_i\right)\kappa\right)\left(1-\mu_i\right)^2-\psi'\left(\kappa\right).$$

Dengan demikian matriks informasi Fisher, yang merupakan turunan kedua dari fungsi log-likelihood yang diberikan dalam Pers. (10) terhadap parameter (β, κ) diberikan oleh

$$I(\beta,\kappa)=\begin{pmatrix}\dfrac{\partial l(\beta,\kappa)}{\partial\beta^2} & \dfrac{\partial l(\beta,\kappa)}{\partial\beta\partial\kappa}\\ \dfrac{\partial l(\beta,\kappa)}{\partial\kappa\partial\beta} & \dfrac{\partial l(\beta,\kappa)}{\partial\kappa^2}\end{pmatrix} \quad \ldots (12)$$

dimana $\dfrac{\partial l(\beta,\kappa)}{\partial\beta^2}=\kappa\mathbf{X}^T\mathbf{W}\mathbf{X}$, $\dfrac{\partial l(\beta,\kappa)}{\partial\beta\partial\kappa}=\mathbf{X}^T\mathbf{T}c$, dan $\dfrac{\partial l(\beta,\kappa)}{\partial\kappa^2}=\mathrm{tr}(\mathbf{D})$. Namun satu hal yang perlu dicatat di sini bahwa parameter β dan κ bersifat tidak ortogonal, dan hal ini berbeda dengan konsep model linear umum yang dibahas dalam McCullagh dan Nelder (1989).

Pada saat ukuran sampel besar, maka vektor penaksir parameter $\hat{\theta}`=\left(\hat{\beta},\hat{\kappa}\right)$ akan mengikuti pendekatan distribusi normal multivariat, atau dapat dinyatakan sebagai

$$\boldsymbol{\theta}\ \square\ N_p\left(\begin{pmatrix}\beta\\ \kappa\end{pmatrix},\left[I(\beta,\kappa)\right]^{-1}\right),$$

dimana $\hat{\beta}$ dan $\hat{\kappa}$ masing-masing adalah penaksir kemungkinan maksimum. Perlu ditambahkan juga bahwa $\left[I(\beta,\kappa)\right]^{-1}$ dapat digunakan untuk memperoleh galat baku asimptotik bagi penaksir kemungkinan maksimum.

Penaksir kemungkinan maksimum bagi β dan κ yang diperoleh dari $U(\beta,\ \kappa)$ = 0 bukan merupakan persamaan tertutup. Dengan demikian solusinya harus diselesaikan secara numerik dengan menggunakan algoritma optimisasi nonlinear, seperti metode penskoran Fisher. Prosedur numerik seperti itu memerlukan nilai awal yang digunakan dalam proses iteratif. Ferrari dan Neto (2004) menyarankan nilai awal untuk parameter β adalam menggunakan penaksir kuadrat terkecil biasa, yaitu dengan cara meregresikan $g(y_1)$, …, $g(y_n)$ pada matriks data **X**. Sementara itu, masih menurut Ferrari dan Neto (2004), diperlukan nilai awal untuk parameter κ. Sebagaimana yang dinyatakan sebelumnya diketahui bahwa varian bagi respons y diberikan dalam Pers. (5) yang berarti bahwa $\kappa=\mu_i(1-\mu_i)/\mathrm{var}(y_i)-1$. Perlu diketahui pula bahwa

$$\mathrm{var}\left(g(y_i)\right)\approx\mathrm{var}\left\{g(\mu_i)+(y_i-\mu_i)g`(\mu_i)\right\}=\mathrm{var}(y_i)\left[g`(\mu_i)\right]^2 \quad \ldots (13)$$

Yang berarti bahwa $\mathrm{var}(y_i)\approx\mathrm{var}\left\{g(y_i)\right\}\left\{g`(\mu_i)\right\}^{-2}$. Dengan demikian nilai awal untuk parameter κ yang disarankan adalah

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{\mu}_i\left(1-\mu_i^*\right)}{\tilde{\sigma}_i^2}-1 \quad \ldots (14)$$

dimana $\tilde{\mu}_i$ diperoleh dengan cara menerapkan $g^{-1}(\square)$ pada nilai taksiran ke-i dari regresi linear $g(y_1)$, …, $g(y_n)$ pada matriks data **X**, yaitu

$$\tilde{\mu}_i=g^{-1}\left(x_i^T\left(X^TX\right)^{-1}X^Tz\right)$$

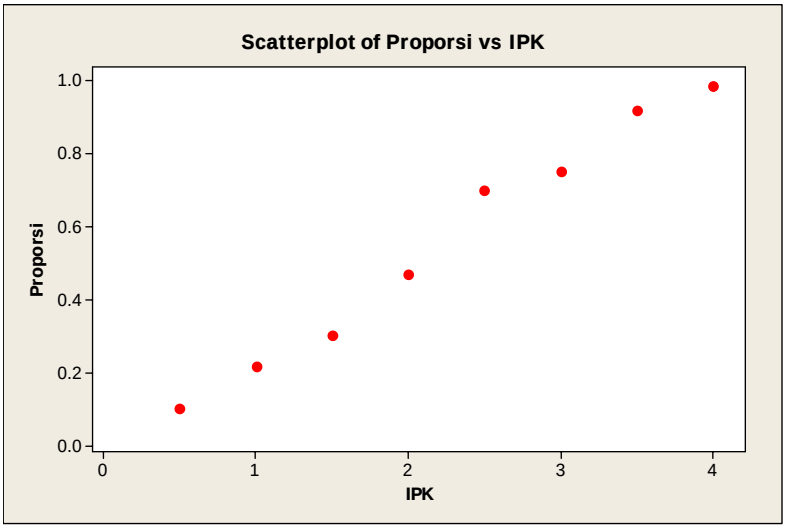
dan

$$\tilde{\sigma}_i^2=\tilde{e}^T\tilde{e}/\left[(n-k)\left\{g`(\tilde{\mu}_i)\right\}^2\right]$$

dimana $\tilde{e}=z-X\left(X^TX\right)^{-1}X^Tz$ yang merupakan vektor residu kuadrat terkecil yang diperoleh dari regresi linear dengan variabel respons yang sudah ditransformasi.

3. CONTOH NUMERIK

Berikut ini akan diberikan suatu contoh numerik mengenai implementasi pemodelan data dengan respons berbentuk proporsi dengan menggunakan regresi beta. Contoh ini mengadopsi penelitian pada bidang *bioassay*, dimana seringkali variabel responnya bisa bervariasi menurut kovariat yang berbentuk dosis. Dalam contoh ini dimisalkan proporsi kelulusan mahasiswa dalam mengikuti suatu mata kuliah tertentu diperoleh nilai indeks prestasi kumulatif (IPK) mahasiswa tersebut yang datanya disajikan dalam Tabel 1. Pada tabel tersebut berisi nilai IPK mahasiswa pada semester berjalan (x_i), banyaknya mahasiswa yang mengikuti mata kuliah tertentu (n_i), banyaknya mahasiswa yang lulus pada mata kuliah tertentu (r_i), serta proporsi mahasiswa yang lulus pada mata kuliah tersebut ($p_i = r_i/n_i$). Kemudian, hasil plot antara proporsi mahasiwa yang lulus dengan IPK ditampilkan pada Gambar 2.



Gambar 2. Plot antara IPK mahasiswa dengan proporsi mahasiswa yang lulus

Pada awalnya data tersebut dianalisis dengan menggunakan model regresi logistik, dengan fungsi hubung logit, yang didefinisikan sebagai

$$\pi_i = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Sehingga fungsi hubungnya adalah model logit yang didefinisikan sebagai logaritma dari odds ($\pi_i/1-\pi_i$), yaitu:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x$$

dimana parameter β_0 dan β_1 masing-masing adalah koefisien regresi yang ditaksir dengan menggunakan metode kemungkinan maksimum.

Tabel 1. Data proporsi kelulusan mahasiswa menurut nilai IPK

Nilai IPK Mahasiswa	Banyaknya mahasiswa (n_i)	Banyaknya mahasiswa yang lulus (r_i)	Proporsi mahasiswa lulus ($p_i = r_i/n_i$)
0.5	60	6	0.1000
1.0	60	13	0.2167
1.5	60	18	0.3000
2.0	60	28	0.4667
2.5	60	42	0.7000
3.0	60	45	0.7500
3.5	60	55	0.9167
4.0	60	59	0.9833

Berdasarkan hasil plot antara nilai IPK mahasiswa (x_i) dengan proporsi mahasiswa yang lulus (r_i/n_i) yang ditunjukkan pada Gambar 2, terlihat adanya pola yang membentuk model logistik, seperti kurva S. Namun demikian, dalam makalah ini akan dicoba memodelkan data yang tersaji pada Tabel 1 dengan menggunakan model regresi beta dengan tujuan untuk melihat bagaimana perilaku penaksir parameter, baik yang dihasilkan dari model regresi logistik maupun dari model regresi beta.

Data proporsi kelulusan mahasiswa tersebut akan dicocokan dengan menggunakan model regresi beta, dimana fungsi hubung yang digunakan adalah fungsi hubung logit. Dalam contoh ini yang dijadikan sebagai variabel prediktor adalah nilai IPK, dan akan dicocokan dalam bentuk model lokasi sebagai berikut:

$$\log\left[\mu/(1-\mu)\right] = \beta_0 + \beta_1 \text{IPK}$$

Dalam model ini diasumsikan bahwa varians dari variabel respons adalah konstan, sehingga parameter dispersinya dimodelkan dalam bentuk $\log\kappa = \delta$. Hasil dari pemodelan regresi beta ini disajikan pada Tabel 2 bersamaan dengan hasil pemodelan melalui regresi logistik. Berdasarkan hasil yang disajikan pada Tabel 2 terlihat bahwa model regresi beta dan model regresi logistik memberikan hasil yang tidak jauh berbeda dalam memodelkan proporsi kelulusan mahasiswa dalam mengambil satu mata kuliah tertentu menurut nilai IPK. Hal ini ditunjukkan bahwa nilai koefisien, galat baku koefisien, p-value, maupun 95% selang kepercayaan untuk kedua model adalah tidak jauh berbeda.

Tabel 2. Nilai koefisien regresi, galat baku, uji signifikansi untuk model regresi beta dan regresi logistik

Parameter	Koefisien	Galat baku	p-value	95% Selang kepercayaan	
				Batas bawah	Batas atas
Model regresi logistik					
β_0	-3.0012	0.2451	< 0.0001	-3.4977	-2.5356
β_1	1.4929	0.1080	< 0.0001	1.2885	1.7124
κ	0.8271	0.0000	< 0.0001	0.8271	0.8271
Model regresi beta					
β_0	-3.0170	0.2289	< 0.0001	-3.5449	-2.4892
β_1	1.5079	0.0978	< 0.0001	1.2824	1.7335
δ	-4.5942	0.4986	< 0.0001	-5.7440	-3.4443

Model regresi beta dapat dituliskan sebagai berikut:

$$\log\left[\mu/(1-\mu)\right] = -3.0170 + 1.5079x_i$$

Sedangkan nilai taksiran untuk parameter dispersinya adalah

$$\hat{\kappa} = \exp\left(-1\times -4.5942\right) = 98.9090$$

Nilai koefisien β_1 untuk model regresi beta adalah bernilai positif, yaitu β_1 = 1.5079. Dengan mengambil nilai exp(β_1) = 4.5, maka model regresi beta ini dapat diinterpretasikan bahwa bagi mahasiswa yang mempunyai nilai IPK lebih tinggi, maka kesempatan mahasiswa tersebut untuk lulus dalam mata kuliah tersebut adalah 4.5 kali lipat lebih besar dibandingkan dengan mahasiswa yang nilai IPK-nya lebih rendah.

Sementara itu, Tabel 3 manyajikan hasil perhitungan mengenai nilai taksiran untuk proporsi mahasiswa yang lulus dan nilai residu Pearson, baik yang dihasilkan melalui model regresi logistik maupun regresi beta. Terlihat pula bahwa kedua model tersebut memberikan hasil yang tentu saja tidak jauh berbeda. Residu chi-kuadrat Perason untuk regresi beta dihitung dengan menggunakan rumus:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i\left(1-\hat{\mu}_i\right)\left(\hat{\kappa}_i+1\right)^{-1}}}$$

Perhatikan bahwa suatu variabel acak yang mengikuti distribusi beta akan berada dalam interval (0, 1), sehingga residu chi-kuadrat Pearson tidak perlu harus mengikuti distribusi normal dengan rata-rata nol.

Tabel 3. Nilai proporsi, taksiran proporsi, dan residu Pearson

Proporsi	Model regresi logistik		Model regresi beta	
	Proporsi taksiran	Residu Pearson	Proporsi taksiran	Residu Pearson
0.1000	0.0949	0.1338	0.0942	0.1975
0.2167	0.1812	0.7135	0.1811	0.9239
0.3000	0.3182	-0.3033	0.3197	-0.4223
0.4667	0.4961	-0.4567	0.4997	-0.6606
0.7000	0.6750	0.4129	0.6798	0.4330
0.7500	0.8142	-1.2784	0.8186	-1.7789
0.9167	0.9024	0.3729	0.9056	0.3792
0.9833	0.9512	1.1548	0.9532	1.4249

Selanjutnya, statistik chi-kuadrat Pearson merupakan jumlah kuadrat dari residu Pearson untuk masing-masing model. Untuk model regresi logistik diperoleh statistik chi-kuadrat Pearson sebesar 4.1050, sedangkan untuk model regresi beta diperoleh sebesar 7.0334. Diketahui bahwa model regresi logistik mempunyai nilai statistik chi-kuadrat Pearson yang lebih kecil dibandingkan dengan model regresi beta. Namun selisih nilai chi-kuadrat Pearson antara kedua model tersebut adalah 7.0334 – 4.1050 = 2.9285, dan ini masih lebih kecil dibandingkan nilai kritis pada distribusi chi-kuadrat dengan α = 0.05 dan derajat bebas sama dengan 1 ($\chi^2_{(0.05;1)}$ = 3.8415). Artinya, kecocokan model terhadap data kelulusan mahasiswa, baik model regresi logistik maupun regresi beta mempunyai kecocokan yang sama secara statistik.

4. DISKUSI

Model regresi beta yang dibahas dalam makalah ini hanya memberikan pemodelan alternatif bagi mereka yang ingin memodelkan data dengan variabel respons yang berbentuk proporsi. Model yang dibahas pada makalah ini merupakan model regresi beta yang dasar, artinya model ini masih mengasumsikan bahwa varians dari variabel prediktornya adalah konstan, atau dengan kata lain tidak masalah heteroskedastisitas dalam data. Padahal sebagaimana yang telah diketahui bahwa data kelulusan mahasiswa seperti yang dibahas dalam makalah ini kemungkinan besar mempunyai masalah heteroskedastisitas. Kemampuan atau tingkat kelulusan mahasiswa dalam suatu mata kuliah tertentu akan bervariasi menurut nilai IPK-nya.

Beberapa model alternatif dengan mengkombinasikan masalah heteroskedastisitas dengan distribusi data yang miring (*skewness*) pada variabel dengan skala terbatas (*bounded*) pada kedua sisi. Salah satu model alternatif yang diusulkan adalah model yang dibahas oleh Kieschnick and McCullogh (2003). Variabel respons yang berbentuk proporsi ini juga dapat dimodelkan dengan menggunakan model regresi ordinal, sebagaimana yang cukup rinci dibahas dalam Long (1997) atau Powers dan Xie (2000). Selain itu, model regresi Tobit dengan adanya data tersensor pada kedua batas merupakan model alternatif lainnya yang mungkin tepat untuk memodelkan variabel respons yang berbentuk proporsi dengan wilayah dengan interval tertutup [0, 1], sedangkan model regresi beta yang dibahas dalam makalah ini adalah proporsi dengan wilayah dengan interval terbuka (0, 1), sebagaimana yang diungkapkan dalam Long (1997).

Terkait, pengembangan model regresi beta ini khususnya yang berkenaan dengan aspek penaksir kemungkinan maksimum untuk suatu variabel acak yang berdistribusi beta dapat dilakukan dengan merujuk pada apa yang dilakukan oleh Schonemann (1983), Papke dan Wooldridge (1996), Paolino (2001), serta Cribari-Neto dan Vasconcellos (2002). Sedangkan pengembangan model regresi beta yang berkaitan dengan masalah diagnostik model dapat merujuk pada Smithson dan Verkuilen (2006), atau Rocha dan Simas (2010). Sekali lagi, model regresi beta yang dibahas dalam makalah ini masih merupakan model dasar untuk memodelkan data yang berbentuk proporsi. Masih terbuka untuk dilakukan penelitian

bagaimana mengatasi masalah heteroskedastisitas ke dalam model, serta mengembangkan model regresi beta apabila datanya banyak mengandung nilai nol atau banyak mengandung nilai satu, atau bahkan keduanya.

DAFTAR PUSTAKA

[1]. Collet, D. (2003). *Modelling Binary Data*. Second Edition. London: Chapman and Hall.

[2]. Cribari-Neto F, and Zeileis A. (2010). "Beta Regression in R", *Journal of Statistical Software*, **34**(2), 1-24.

[3]. Cribari-Neto, F. and Vasconcellos, K. L. P. (2002). "Nearly unbiased maximum likelihood estimation for the beta distribution", *Journal of Statistical Computation and Simulation*, **72**, 107-118.

[4]. Dobson, A.J. (1983). *Introduction to Statistical Modelling*. London: Chapman and Hall.

[5]. Ferrari, S. L. P. and Cribari-Neto, F. (2004). "Beta Regression for Modeling Rates and Proportions", *Journal of Applied Statistics*, **10**, 1-18.

[6]. Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995) *Continuous Univariate Distributions, Volume 2*. Second Edition, New York: Wiley.

[7]. Kieschnick, R., & McCullogh, B. D. (2003). "Regression analysis of variates observed on (0,1): Percentages, proportions, and fractions", *Statistical Modeling*, **3**, 193–213.

[8]. Long, J. S. (1997). *Regression with Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

[9]. McCullagh, P., and J.A. Nelder (1983). *Generalized Linear Models*. Second Edition. New York: Chapman and Hall.

[10]. Nelder. J.A., and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of Royal Statiscal Society*, Series A, **153**: 370-384.

[11]. Paolino, P. (2001). "Maximum likelihood estimation of models with beta-distributed dependent variables", *Political Analysis*, **9**, 325-346.

[12]. Papke, L., & Wooldridge, J. (1996). "Econometric methods for fractional response variables with an application to 401(K) plan participation rates", *Journal of Applied Econometrics*, **11**, 619-632.

[13]. Powers, D. A., & Xie, Y. (2000). *Statistical Methods for Categorical Data*. San Diego, CA: Academic Press.

[14]. Rocha AV and Simas AB (2010). "Infuence Diagnostics in a General Class of Beta Regression Models." Test

[15]. SAS Institute (2005). *The GLIMMIX Procedure*. Cary, NC: SAS Institute.

[16]. Schonemann, P. H. (1983). "Some theory and results for metrics for bounded response scales", *Journal of Mathematical Psychology*, **27**, 311-324.

[17]. Smithson M, and Verkuilen J. (2006) "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables", *Psychological Methods*, **11**, pp 54-71.

[18]. Smyth, G. K. (1989). "Generalized linear models with varying dispersion", *Journal of the Royal Statistical Society*, Series B, **51**, 47–60.

[19]. Swearingen, C. J., Castro, M.S.M., and Bursac, Z. (2010). *Modeling Percentage Outcomes: The %Beta_Regression Macro*. SAS Global Forum 2011: Statistics and Data Analysis, Paper 335-2011.