Statistika, Vol. 9 No. 1, 55 – 64
Mei 2009

Pemeriksaan Ketepatan Fungsi Hubung dalam Analisis Data Biner

NUSAR HAJARISMAN
Program Studi Statistika Fakultas MIPA Unisba
Mahasiswa Sekolah Pascasarjana Institut Pertanian Bogor
nrisman@yahoo.co.uk

RINGKASAN

Dalam pemodelan data biner atau binomial ada sejumlah cara dimana model dugaan tidak layak. Yang paling penting dari semua itu adalah komponen sistematik linear dari model tidak dinyatakan dengan tepat. Sebagai contoh misalnya model tidak menyertakan suatu variabel penjelas yang memang seharusnya berada di dalam model, atau mungkin satu atau dua variabel penjelas perlu ditransformasi sebelumnya. Kedua, transformasi dari peluang respons yang digunakan mungkin tidak tepat; misalnya mungkin saja bahwa transformasi dari peluang respons yang telah digunakan adalah transformasi logistik padahal seharusnya menggunakan transformasi log-log komplementer. Ketiga, data mungkin berisi suatu data pencilan yang mengakibatkan data tidak cocok terhadap model dugaan. Teknik yang digunakan untuk memeriksa kelayakan model ini disebut juga sebagai proses diagnosa. Pada makalah ini pembahasan akan lebih difokuskan pada pemeriksaan ketepatan fungsi hubung dalam pemodelan data biner.
Kata Kunci: data biner, model linear terampat, devians, fungsi hubung, model logit, model probit, model log-log komplementer, log-likelihood.

1. Pendahuluan

Dalam berbagai bidang penelitian yang menggunakan prosedur statistika, seperti dalam bidang agronomi, pertanian, sosial dan ekonomi, politik, kesehatan, biologi, dan teknik, data yang diamati dibuat pada unit percobaan yang mengambil nilai salah satu dari dua kategori yang mungkin. Sebagai contoh, suatu benih akan berkecambah atau gagal berkecambah di bawah kondisi percobaan tertentu; suatu peralatan listrik yang diproduksi oleh sebuah pabrik elektronik dapat cacat atau tidak cacat; seorang pasien dalam percobaan klinis dapat dinyatakan sembuh atau sakit setelah diberi sejumlah perlakuan; atau serangga dapat dinyatakan bertahan hidup atau mati setelah diberi sejumlah dosis insektisida. Data semacam itu dikatakan sebagai data biner dan dua kategori yang mungkin untuk masing-masing observasi secara umum dinyatakan dengan istilah 'sukses' atau 'gagal'.

Dalam beberapa situasi, penelitian tidak hanya difokuskan pada respons dari satu unit percobaan tertentu (benih, pasien, alat listrik, dan serangga) tetapi pada segugus unit percobaan yang telah diberi perlakuan yang sama. Jadi, misalnya segugus benih dapat dipaparkan pada kondisi yang ditentukan oleh kelembaban dan suhu, kemudian proporsi dari benih yang berkecambah akan dicatat. Demikian juga bagi respons individu dari masing-masing pasien dalam percobaan klinis yang menerima perlakuan sama, serta mempunyai karakteristik yang mirip berdasarkan faktor-faktor demografis (umur atau jenis kelamin), dapat dikombinasikan untuk mendapatkan proporsi dari pasien yang dinyatakan sembuh. Data seperti ini disebut juga sebagai data biner terkelompok (*grouped binary data*) serta mewakili banyaknya peristiwa 'sukses' dari banyaknya unit percobaan yang dilakukan.

Data berbentuk proporsi seperti ini seringkali dimodelkan dengan menggunakan dengan menggunakan distribusi binomial sedangkan data biner itu sendiri diasumsikan mempunyai distribusi Bernoulli (Collet, 2003). Terdapat beberapa model yang dapat digunakan untuk memodelkan data respons binomial, diantaranya yaitu: model logistik, model probit, dan model log-log komplementer.

Setelah model dicocokan ke data pengamatan dari variabel respons biner atau binomial, maka langkah penting selanjutnya adalah melakukan pemeriksaan kelayakan model dugaan. Ada sejumlah cara dimana model dugaan tidak layak. Yang paling penting dari semua itu adalah komponen sistematik linear dari model tidak dinyatakan dengan tepat. Sebagai contoh

misalnya model tidak menyertakan suatu variabel penjelas yang memang seharusnya berada di dalam model, atau mungkin satu atau dua variabel penjelas perlu ditransformasi sebelumnya. Kedua, transformasi dari peluang respons yang digunakan mungkin tidak tepat; misalnya mungkin saja bahwa transformasi dari peluang respons yang telah digunakan adalah transformasi logistik padahal seharusnya menggunakan transformasi log-log komplementer. Ketiga, data mungkin berisi suatu data pencilan yang mengakibatkan data tidak cocok terhadap model dugaan. Teknik yang digunakan untuk memeriksa kelayakan model ini disebut juga sebagai proses diagnosa. Pada makalah ini pembahasan akan lebih difokuskan pada pemeriksaan ketepatan fungsi hubung dalam pemodelan data biner.

Sebagaimana yang telah diketahui bahwa di dalam pemodelan data biner atau binomial, suatu fungsi tertentu dari peluang respons, yang dikenal sebagai fungsi hubung, adalah menghubungkan ke kombinasi lienar dari variabel penjelas dalam model. Transformasi logistik adalah salah satu yang paling banyak digunakan, akan tetapi perlu diingat bahwa belum tentu transformasi logistik ini akan cocok untuk berbagai kasus. Dalam hal ini harus pula dipertimbangkan apakah perbedan transformasi akan membawa pada model yang lebih sederhana atau suatu model yang memberikan kecocokan yang lebih baik. Sebagai contoh, misalnya model logistik linear dengan segugus vaiabel penjelas bukan merupakan model yang cocok terhadap data yang diamati, tetapi bisa saja model probit atau log-log komplementer. Alternatifnya, komponen linear, katakan saja model log-log komplementer memerlukan bentuk yang lebih sederhana daripada komponen yang dalam model logistik.

Dalam keadaan demikian, maka akibatnya penentuan atau pemilihan fungsi hubung yang memadai dilakukan bersamaan dengan penentuan struktur linear dari model. Dengan demikian setiap studi yang berhubungan dengan kelayakan suatu fungsi hubung selalu didasarkan pada segugus variabel penjelas yang tetap (*fixed*). Dalam makalah ini akan dibahas mengenai berbagai metode untuk memeriksa kelayakan model yang difokuskan pada pemilihan suatu fungsi hubung dalam memodelkan data biner.

2. Model untuk Respons Biner

Pada bagian ini akan dibahas tentang model linier umum yang mana variabel-variabel responnya diukur dengan skala biner. Sebagai contoh, misalnya hidup atau mati, hadir atau tidak hadir, sehat atau sakit, dan lain-lain. Secara umum kejadian-kejadian itu dinyatakan dalam bentuk 'sukses' dan 'gagal' untuk dua buah kategori. Selanjutnya, akan didefiniskan variabel acak sebagai berikut:

$Y = 1$ jika variabel responnya menyatakan sukses,

$= 0$ jika variabel responnya menyatakan gagal,

dengan $\pi = P(Y = 1)$ dan $1 - \pi = P(Y = 0)$. Jika terdapat n variabel $Y_1, \ldots, Y_n$ yang saling bebas dengan $\pi_j = P(Y_j = 1)$, maka peluang bersamanya adalah:

$$\prod_{j=1}^{n} \pi_j^{y_j} (1-\pi_j)^{1-y_j} = \exp\left[\sum_{j=1}^{n} y_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log\left(1-\pi_j\right)\right] \quad \ldots (1)$$

dimana bentuk tersebut merupakan anggota dari keluarga distribusi eksponensial. Untuk kasus dimana π_j semuanya bernilai sama, maka akan didefiniskan $R = \sum_{j=1}^{n} Y_j$, yaitu banyaknya peristiwa sukses dalam n buah percobaan. Variabel acak R tersebut mempunyai distribusi binomial $b(n, \pi)$, yaitu dengan fungsi masa peluangnya sebagai berikut:

$$P(R = r) = \binom{n}{r} \pi^r (1-\pi)^{n-r}, \quad (r = 0, 1, \ldots, n) \quad \ldots (2)$$

Dengan demikian, maka $E(R) = n\pi$ dan $Var(R) = n\pi(1 - \pi)$.

Secara umum maka kita perhatikan N buah variabel yang saling bebas $R_1, R_2, \ldots, R_N$ menurut banyaknya peristiwa sukses dalam N sub kelompok atau strata yang berbeda (lihat Tabel 1). Jika $R \sim b(n_i, \pi_i)$, maka fungsi log-likelihoodnya adalah:

$$l(\pi_1,...,\pi_N;r_1,...,r_N)=\sum_{i=1}^{N} r_i \log\left(\frac{\pi_i}{1-\pi_i}\right)+n_i \log(1-\pi_i)+\log\binom{n_i}{r_i} \quad \text{... (3)}$$

Dalam hal ini distribusi dari Pers. (1) dan (2) merupakan kasus khusus dari Pers. (3).

Tabel 1. Frekuensi Untuk *N* Distribusi Binomial

	Sub kelompok atau strata			
	1	2	...	*N*
Sukses	R_1	R_2	...	R_N
Gagal	$n_1 - R_1$	$n_2 - R_2$	...	$n_N - R_N$
Total	n_1	n_2	...	n_N

Model-model yang dibahas dalam makalah ini merupakan kasus khusus dari model linier umum, suatu model yang diperkenalkan oleh Nelder dan Wedderburn (1972). Model linier umum ini dispesifikasikan oleh tiga buah komponen, yaitu: komponen acak, komponen sistematik, dan fungsi penghubung.

Komponen acak adalah suatu komponen yang mengidentifikasikan distribusi peluang dari variabel respon, dimana komponen ini akan berisi pengamatan tak bebas $\mathbf{Y} = (Y_1, ..., Y_N)'$ dari distribusi dalam keluarga eksponensial. Yaitu, masing-masing pengamatan Y_i mempunyai fungsi densitas peluang atau fungsi masa peluang dalam bentuk:

$$f(y_i;\theta_i) = a(\theta_i)\, b(y_i) \exp[y_i Q(\theta_i)] \quad \text{... (4)}$$

Keluarga ini menyangkut beberapa distribusi penting sebagai kasus khusus, termasuk distribusi binomial dan Poisson. Nilai parameter θ_i dalam Pers. (4) dapat bervariasi untuk $i = 1, 2,..., N$, bergantung pada nilai dari variabel-variabel penjelasnya. Sedangkan bentuk $Q(\theta)$ disebut sebagai parameter alamiah dari distribusi itu sendiri (Agresti, 1990).

Komponen sistematik dari model linier umum akan menghubungkan vektor $\eta = (\eta_1, ..., \eta_N)'$ kepada sekumpulan variabel penjelas melalui model linier:

$$g(\pi_i)=\eta_i=\mathbf{X}\boldsymbol{\beta} \quad \text{... (5)}$$

dimana $\mathbf{X}$ adalah matriks model (kadang-kadang disebut juga matriks rancangan) yang berisi nilai-nilai variabel-variabel penjelas untuk N buah pengamatan, dan $\boldsymbol{\beta}$ adalah vektor dari parameter-parameter di dalam model. Vektor η disebut sebagai prediktor linier.

Salah satu kekurangan dari model linear semacam ini adalah bahwa penduga dari π_i kadang-kadang akan berada diluar interval [0, 1]. Agar supaya masalah tersebut tidak terjadi, maka biasanya akan digunakan fungsi distribusi kumulatif:

$$F(x)=g^{-1}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)=\int_{-\infty}^{x} f(y)dy$$

Dimana $f(y)$ merupakan fungsi kepekatan peluang dari variabel acak y. Fungsi kepekatan peluang yang bias digunakan dalam menganalisis data biner adalah distribusi normal, logistik, serta log-log komplementer. Ketiga distribusi ini akan dibahas pada sub bagian berikut ini.

2.1 Model Probit

Jika distribusi normal digunakan sebagai fungsi kepekatan peluang, sehingga bentuk distribusi peluang kumulatifnya dinyatakan sebagai berikut:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]dy$$

$$=\Phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{... (6)}$$

dimana Φ menyatakan distribusi peluang kumulatif untuk normal baku $N(0, 1)$, sehingga diperoleh:

$$g(\pi_i) = \Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i \quad \ldots (7)$$

dimana $g(\pi_i)$ merupakan fungsi penghubungnya, serta $\beta_0 = -\mu/\sigma$ dan $\beta_1 = 1/\sigma$. Fungsi hubung g adalah invers fungsi peluang normal kumulatif Φ^{-1}. Pada saat distribusi kepekatan peluang yang digunakan adalah normal, maka model yang relevant untuk masalah ini disebut sebagai model probit. Probit dari peluang π didefinsikan untuk setiap x_i, $i = 1, 2, \ldots, k$ sebagai suatu nilai dari s sedemikian rupa sehingga

$$\pi_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \exp\left(-\frac{y^2}{2}\right) dy, \quad \text{dimana } s_i = \frac{X_i - \mu}{\sigma}$$
$$= \Phi(s_i)$$

2.2 Model Logit

Distribusi logistik mempunyai fungsi kepekatan sebagai berikut:

$$f(y) = \frac{\beta_1 \exp(\beta_0 + \beta_1 y)}{\left[1 + \exp(\beta_0 + \beta_1 y)\right]^2} \quad \ldots (8)$$

dan

$$\pi(x) = \int_{-\infty}^{x} f(y)dy = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \ldots (9)$$

Definisi alternatif dari $F(x)$ adalah

$$\pi(x) = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 x)\right]} \quad \ldots (10)$$

Yang akan menghasilkan bentuk logit sebagai berikut:

$$\ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] = \beta_0 + \beta_1 x_i = \lambda_i \quad \ldots (11)$$

Untuk $i = 1, 2, \ldots, k$, yang selanjutnya dirujuk sebagai fungsi logistik. Fungsi hubung yang bersesuaian yang diberikan oleh fungsi logit adalah:

$$g(\pi) = \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] \quad \ldots (12)$$

Dari persamaan di atas jelas bahwa odds untuk respons 'sukses' adalah

$$\frac{\pi(x_i)}{1 - \pi(x_i)} = \exp(\beta_0 + \beta_1 x_i) = e^{\beta_0} + \left(e^{\beta_1}\right)^{x_i} \quad \ldots (13)$$

Secara umum dapat dituliskan bahwa

$$\lambda_i = \mathbf{x}_i^T \boldsymbol{\beta} \quad \ldots (14)$$

dimana: $\mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. Pers. (14) digambarkan sebagai model regresi logistik linear sebab model tersebut seperti model regresi biasa untuk kasus dimana variabel penjelasnya berbentuk kuantitatif, selain itu Pers. (14) akan seperti model ANOVA jika varibel penjelasnya berbentuk kategorik. Dalam hal ini seringkali bentuk di atas disebut sebagai model logit.

2.3 Model Log-log Komplementer

Model lainnya yang juga dapat dipandang untuk memodelkan data dosis-respons adalah model log-log komplementer dan model log-log. Model log-log komplementer kadang-kadang disebut juga sebagai model nilai-ekstrem (Lawal, 2003) dan dicirikan oleh

$$f(y) = \beta_1 \exp\left[(\beta_0 + \beta_1 y) - \exp(\beta_0 + \beta_1 y)\right] \quad \ldots (15)$$

dan

$$\pi(x) = 1 - \exp\left[-\exp(\beta_0 + \beta_1 x)\right] \quad \ldots (16)$$

Suatu transformasi dalam bentuk

$$\ln\left[-\ln(1-\pi)\right] = \beta_0 + \beta_1 x \quad \ldots (17)$$

Akan mentransformasikan $\pi(x)$ ke dalam bentuk model linear. Fungsi hubung $\ln[-\ln(1 - \pi)]$ disebut juga sebagai fungsi log-log komplementer. Menurut Lawal (2003) model ini biasanya lebih banyak digunakan dibandingkan model probit dan logit untuk π yang bernilai mendekati 0 atau 1.

3. Pemeriksaan Model untuk Respons Biner

Pada bagian ini akan dibahas mengenai suatu prosedur yang dapat digunakan untuk menentukan apakah suatu transformasi tertentu mampu dengan layak menggambarkan peluang respons sebenarnya. Prosedur ini juga dapat digunakan untuk memberikan suatu indikasi bahwa transformasi tertentu merupakan transformasi yang paling 'tepat'.

Misalkan bahwa suatu fungsi hubung tertentu digunakan dalam memodelkan segugus data biner yang bergantung pada beberapa parameter, dimana suatu nilai yang berbeda dari parameter ini akan membawa pada fungsi hubung yang berbeda pula. Misalkan α_0 adalah nilai dari parameter tersebut yang digunakan dalam memodelkan segugus data biner, maka fungsi hubung untuk peluang respons ke-*i*, $i = 1, 2, \ldots, n$, dapat dinyatakan oleh

$$g(\pi_i; \alpha_0) = \eta_i \quad \ldots (18)$$

dimana η_i merupakan komponen linear dari model untuk pengematan ke-*i*. Dimisalkan pula bahwa fungsi hubung yang tepat (walaupun pada dasarnya belum diketahui) adalah $g(\pi_i, \alpha)$. Keluarga dari fungsi hubung yang diusulkan oleh Aranda-Ordaz (1981), dimana

$$g(\pi_i; \alpha) = \log\left\{\frac{(1-\pi_i)^{-\alpha} - 1}{\alpha}\right\} \quad \ldots (19)$$

adalah berguna untuk pemodelan data biner dan binomial. Pada saat $\alpha = 1$, maka diperoleh

$$g(\pi_i; \alpha) = \log\left\{\frac{\pi_i}{1-\pi_i}\right\}$$

yang merupakan transformasi logistik dari π_i. Kemudian jika $\alpha \to 0$, maka

$$\left\{(1-\pi_i)^{-\alpha} - 1\right\}/\alpha \quad \to \quad \log(1-\pi_i)^{-1}$$

sehingga diperoleh $g(\pi_i, \alpha) = \log\{-\log(1 - \pi_i)\}$ yang merupakan fungsi hubung log-log komplementer. Dalam banyak kasus, fungsi hubung yang dihipotesiskan, $g(\pi_i, \alpha_0)$ adalah fungshubung logit, yaitu dalam hal $\alpha_0 = 1$.

Fungsi $g(\pi_i, \alpha)$ dapat didekati oleh perluasan deret Taylor dari fungsi di sekitar α_0, yaitu:

$$g(\pi_i; \alpha) \approx g(\pi_i; \alpha_0) + (\alpha - \alpha_0)\left.\frac{\partial g(\pi_i; \alpha)}{\partial \alpha}\right|_{\alpha=0}$$

Model yang tepat kemudian dapat ditentukan melalui model:

$$g(\pi_i;\alpha_0) = \eta_i + \gamma z_i \quad \text{... (20)}$$

dimana $\gamma = \alpha_0 - \alpha$ dan $z_i = \partial g(\pi_i;\alpha)/\partial\alpha\big|_{\alpha=0}$. Model ini menggunakan fungsi yang dihipotesiskan dan termasuk nilai z_i sebagai variabel penjelas tambahan Z.

Sebelum model dalam Pers. (20) dicocokan, maka nilai Z harus ditentukan terlebih dahulu. Nilai Z ini bergantu pada π_i yang ditaksir oleh p_i, yaitu dugaan dari peluang respons untuk pengamatan ke-i dan diperoleh melalui pencocokan model dalam Pers. (18) dimana pemilihan fungsi hubung awal yang digunakan.

Untuk fungsi hubung yang digunakan dalam Pers. (19), nilai dari variabel z_i adalah

$$z_i = \frac{\log(1-p_i)}{(1-p_i)^{\alpha}-1} - \alpha^{-1}$$

dan dalam kasus khusus dimana $\alpha = 1$, yaitu jika fungsi hubung logit adalah yang dihipotesiskan, maka persamaan di atas menjadi:

$$z_i = -\left\{1 + p_i^{-1}\log(1-p_i)\right\} \quad \text{... (21)}$$

Apabila $\gamma = 0$, maka $\alpha = \alpha_0$ dan fungsi hubung yang dihipotesiskan adalah benar. Akibatnya, suatu uji hipotesis bahwa $\gamma = 0$ dalam Pers. (20) tidak lain adalah untuk menguji kelayakan fungsi hubung. Hipotesis ini dapat diuji dengan cara melihat pengurangan dalam devians dengan menambahkan Z ke dalam model. Apabila pengurangan dalam devians ini relatif besar dibandingkan dengan titik persentase dari distribusi χ^2 dengan derajat bebas satu, maka kita dapat memutuskan bahwa fungsi hubung awal yang dipilih tidak tepat. Prosedur ini dirujuk sebagai uji kecocokan fungsi hubung (Collet, 2003). Dalam prakteknya, uji kecocokan fungsi hubung ini mempunyai keterbatasan. Salah satunya adalah ketika menghadapi suatu gugus data yang besar, maka diperlukan suatu fungsi hubung lainnya, misalnya fungsi hubung probit.

4. Contoh Aplikasi

Dalam data bioassay, variabel responsnya dapat bercariasi sesuai dengan kovariat yang membentuk suatu dosis. Contoh sejenis yang melibatkan respons biner diberikan dalam Tabel 2, dimana R adalah banyaknya serangga yang mati setelah diberi obat pembasmi hama selama 5 jam dalam berbagai macam konsentrasi (data dari Dobson, 2002). Gambar 1 menunjukkan proporsi $p_i = r_i/n_i$ yang diplot terhadap dosis x_i.

Tabel 2. Data Kematian Serangga

Dosis x_i	Banyak serangga yang diamati, n_i	Banyaknya serangga yang mati, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Data tersebut kemudian akan dimodelkan melalui tiga jenis fungsi penghubung, yaitu logit, probit, dan log-log komplementer. Proses pendugaan parameter dilakukan secara iteratif dengan menggunakan metode Fisher-Scoring. Hasil analisis yang ditampilkan di sini adalah proses iterasi, nilai devians, dugaan untuk proporsi dan respons, penduga parameter, serta residu untuk masing-masing dari fungsi penghubung yang digunakan.

Untuk model logistik linear kita mengambil

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$$

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

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

Dari (8.3) diketahui fungsi log-kemungkinannya, yaitu:

$$l = \sum_{i=1}^{N}\left[r_i(\beta_0+\beta_1 x_i) - n_i \log\left(1+e^{\beta_0+\beta_1 x_i}\right) + \log\binom{n_i}{r_i}\right]$$

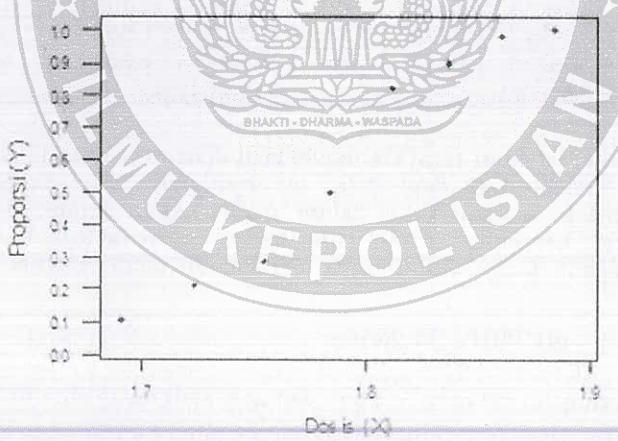
dan skor terhadap β_1 dan β_2 adalah

$$U_1 = \frac{\partial l}{\partial \beta_0} = \sum\left[r_i - n_i\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right] = \Sigma\,(r_i - n_i\pi_i)$$

$$U_2 = \frac{\partial l}{\partial \beta_1} = \sum\left[r_i x_i - n_i x_i\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right] = \Sigma\, x_i(r_i - n_i\pi_i)$$

Dengan cara yang sama akan diperoleh matriks informasi sebagai berikut:

$$\mathbf{I} = \begin{bmatrix} \sum n_i\pi_i(1-\pi_i) & \sum n_i x_i \pi_i(1-\pi_i) \\ \sum n_i x_i \pi_i(1-\pi_i) & \sum n_i x_i^2 \pi_i(1-\pi_i) \end{bmatrix}.$$



Gambar 1. Plot antara dosis dan proporsi kematian serangga

Penduga kemungkinan maksimum diperoleh melalui penyelesaian persamaan iteratif

$$\mathbf{I}^{(m-1)}\mathbf{b}^{(m)} = \mathbf{I}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

(dari (4.7)) dimana m menyatakan proses iterasi ke-m dan $\mathbf{b} = [b_0\ b_1]^T$ adalah vektor penduganya. Dengan nilai awal $b_0^{(0)} = b_1^{(0)} = 0$ hasil proses ietrasinya ditunjukkan dalam Tabel 8.4 bersama-sama dengan frekuensi taksiran $\hat{r}_i = n_i\hat{\pi}_i$, matriks penduga varians-kovarians $[\mathbf{I}(\mathbf{b})]^{-1}$ dan statistik rasio log-kemungkinan.

Galat baku penduga b_0 = -60.72 dan b_1 = 34.27 masing-masing diberikan oleh $(26.802)^{1/2}$ = 5.18 dan $(8.8.469)^{1/2}$ = 2.91. Di bawah hipotesis nol bahwa model logistik linear dapat menggambarkan data, maka D mempunyai pendekatan distribusi χ^2_6, sebab terdapat N = 8 kelompok dosis dengan p = 2 buah parameter. Tetapi diketahui bahwa dari tabel distribusi χ^2_6 dengan taraf kepercayaan 5% adalah 12.59 yang menunjukkan bahwa model tidak menggambarkan data dengan baik.

Tabel 3. Nilai penduga parameter untuk fungsi hubung logit, probit, dan log-log komplementer

Model		Estimate	Standard Error	Chi-square	p-value
Logit	B0	-60.7401	5.1819	137.40	< 0.0001
	B1	34.2859	2.9132	138.51	< 0.0001
Probit	B0	-34.9561	2.6413	175.15	< 0.0001
	B1	19.7410	1.4853	176.66	< 0.0001
Clog-log	B0	-39.5223	3.2229	150.38	< 0.0001
	B1	22.0148	1.7899	151.28	< 0.0001

Dengan menggunakan prosedur GENMOD dalam sistem SAS, model probit $\pi = \Phi(\beta_0 + \beta_1 x_i)$ dan model log-log komplementer $\pi = (1 - \exp(-\exp(\beta_0 + \beta_1 x_i))$ juga dipakai untuk mencocokan data. Hasil-hasil pendugaan model dan nilai devians untuk setiap fungsi hubung masing-masing disajikan dalam Tabel 3 dan Tabel 4. Sedangkan nilai dugaan untuk peluang respons dan dugaan untuk variabel responsnya untuk setiap fungsi hubung disajikan dalam Tabel 5.

Tabel 4. Nilai devians, chi-kuadrat Pearson, dan log-likelihood untuk model logit, probit, dan log-log komplementer

	LOGIT			PROBIT			CLOG-LOG		
	Df	value	Value/df	df	value	Value/df	df	value	Value/df
Devians	6	11.1156	1.8526	6	9.9870	1.6645	6	3.5143	0.5857
Pearson	6	9.9067	1.6511	6	9.3690	1.5615	6	3.3592	0.5599
Log-lik	-186.1771			-185.6128			-182.3765		

Diantara model-model tersebut ternyata model logit dengan bentuk linear tidak cocok terhadap data. Kecocokan model terhadap data ini mungkin bisa ditingkatkan dengan cara menambahkan bentuk kuadratik ke dalam model. Akan tetapi, dalam contoh ini akan diperhatikan apakah kecocokan model yang hanya berisi bentuk linear dapat ditingkatkan dengan cara mengganti fungsi hubungnya. Persamaan model regresi logit dugaan diberikan oleh

$$\text{logit}(\hat{\pi}_i) = -60.7401 + 34.2859 x_i$$

dimana $\hat{\pi}_i$ adalah dugaan peluang respons dan x_i adalah dosis obat untuk kelompok ke-i. Dari Tabel 4 terlihat bahwa devians untuk model ini adalah 11.1156 pada derajat bebas sebesar 6. Di sini fungsi hubung yang dihipotesiskan adalah fungsi logit, dan untuk mengetahui apakah fungsi hubung tersebut cukup tepat untuk menggambarkan peluang responsnya, maka variabel Z ditambahkan ke dalam model, dimana nilai ke-i dari Z diberikan dalam Pers. (21). Pada saat Z disertakan ke dalam model regresi logit, maka diperoleh nilai devians sebesar 6.4562 pada derajat bebas 5. Pengurangan devians sebesar 4.6594 setelah dimasukan variabe Z ke dalam model adalah signifikan pada taraf signifikansi sebesar 5%. Hal ini mempunyai makna bahwa fungsi hubung logit tidak tepat digunakan untuk data tersebut.

Tabel 5. Nilai dugaan peluang respons dan dugaan variabel respons

No.	x	n	y	LOGIT		PROBIT		CLOG-LOG	
				phi	Yhat	phi	yhat	phi	yhat
1	1.691	59	6	0.05938	3.5033	0.05774	3.4065	0.09582	5.6535
2	1.724	60	13	0.16367	9.8200	0.17811	10.6864	0.18803	11.2816
3	1.755	62	18	0.36162	22.4206	0.37804	23.4384	0.33777	20.9419
4	1.784	56	28	0.60491	33.8749	0.60328	33.7839	0.54178	30.3395
5	1.811	63	52	0.79440	50.0475	0.78665	49.5592	0.75684	47.6809
6	1.837	59	53	0.90406	53.3393	0.90459	53.3705	0.91844	54.1877
7	1.861	62	61	0.95547	59.2390	0.96262	59.6823	0.98575	61.1166
8	1.884	60	60	0.97926	58.7554	0.98732	59.2394	0.99914	59.9481

Koefisien z_i dalam model ini, $\hat{\gamma}$, adalah, 1.232, sehingga parameter dalam fungsi hubung yang umum yang diberikan dalam Pers. (19) diduga oleh:

$$\hat{\alpha} = 1 - \hat{\gamma} = -0.232$$

Yang tidak berbeda dengan nol. Hal ini mempunyai makna bahwa model dengan fungsi hubung log-log komplementer relatif lebih baik daripada model logit. Namun demikan, hasil di atas masih belum memuaskan karena belum dibandingkan dengan fungsi hubung lainnya, seperti fungsi hubung probit.

5. Kesimpulan

Setidaknya ada dua alasan penting mengapa model regresi logistik lebih banyak penggunaannya dibandingkan model probit dan model log-log komplementer untuk analisis data biner. Pertama, model logistik mempunyai interpretasi yang jelas dalam bentuk logaritma dari odds rasio. Interpretasi seperti ini akan sangat bermanfaat untuk analisis data dalam studi epidemiologi ataupun percobaan-percobaan klinis. Kedua, model yang berdasarkan pada transformasi logistik cukup tepat digunakan untuk analisis data yang dikumpulkan secara retrospektif sebagaimana dalam studi kasus-kontrol (*case-control study*). Namun demikian, tidak semua data mampu dicocokan dengan tepat melalui model logit ini. Oleh karena itu, pemeriksaan secara seksama terhadap model yang sedang dicocokan perlu dilakukan.

Berbagai metode untuk pemeriksaan dalam analisis data biner atau binomial telah banyak dibahas oleh Collet (2003). Namun dalam makalah ini hanya ditunjukkan pada pemeriksaan ketepatan fungsi hubung dalam analisis data biner. Lebih khusus lagi, pemeriksaan ketepatan fungsi hubung ini hanya dilakukan untuk fungsi hubung logit dan log-log komplementer. Perlu kiranya untuk mengembangkan metode ini pada kasus yang lebih umum, dimana semua fungsi hubung lainnya (seperti fungsi hubung identitas atau probit) dapat ditangani.

Daftar Pustaka

[1]. Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley and Sons.

[2]. Aitkin, M., D. Anderson, B. Francis, and J. Hinde. (1989). *Statistical Modelling in GLIM.* Oxford: Clorendeon Press.

[3]. Baker, R.J., and J.A. Nelder. (1978). *Generalized Linear Interactive Modeling (GLIM).* Release 3. Oxford: Numerical Algorithms Group.

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

[5]. Cox, D.R. (1970). *The Analysis of Binary Data.* London: Methuen.

[6]. Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

[7]. Dodson, 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.* 2nd Ed. New York: John Wiley and Sons.

[9]. Hosmer, D.W. and S. Lemeshow (1989). *Applied Logistic Regression.* New York: John Wiley and Sons.

[10]. Kleinbaum, D.G., (1994). *Logistic Regression: A Self-Learning Text,* New York: Springer-Verlag.

[11]. Lawal, B. (2003) *Categorical Data Analysis With SAS And SPSS Applications.* London: Lawrence Erlbaum Associates.

[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.