J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 6, No.2, 107-114, Januari 2010
ISBN:1858-1382

Aplikasi Model Regresi Logit dan Probit pada Data Kategorik

Georgina M. Tinungki *

Abstrak

Pembahasan dua model alternatif untuk data biner yaitu model Regressi Logit dan Probit. Regressi Logit digunakan untuk menggambarkan hubungan variabel dependen (variabel respon) dengan variabel independen (variabel prediktor) yang bersifat kategori, kontinu atau kombinasi keduanya. Sedangkan regresi Probit adalah suatu analisis regresi yang digunakan untuk menggambarkan hubungan antara variabel dependen dan variabel independen. Variabel dependen (variabel respon) biasa disimbolkan **Y** dengan skala pengukuran dikotomus (biner), dan variabel independen (variable prediktor) biasa disimbolkan **X** yang skala pengukuran bersifat dikotomus, polikotomus atau kontinu. Untuk pendugaan parameter pada model regresi dan regresi Logit digunakan Metode Maksimum Likelihood. Metode Maksimum likelihood (MLE) digunakan untuk mengestimasi parameter yang belum diketahui dengan menetapkan asumsi distribusi Bernoulli dan obyek pengamatan saling bebas, $E(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. Prinsip dari MLE untuk mendapatkan nilai taksiran β yang memaksimumkan fungsi likelihood. Aplikasi distribusi toleransi untuk log dosis adalah pendekatan normal dengan μ dan standar deviasi σ. Jika G adalah cdf dari distribusi normal, maka

$$\pi(x) = G(x) = \Phi[(x-\mu)/\sigma]$$

dimana Φ adalah standar normal cdf. Konsep dari *toleransi distribusi* menjadi dasar kebenaran untuk model. Sehingga, koresponden yang cocok untuk toleransi distribusi normal pada aplikasi toxikologi memiliki mean 1,77 dan standar deviasi 1/19,74 = 0,05.

Kata Kunci: *Regressi Logit, regressi Probit, data kategorik, metode maksimum Likelihood, toleransi distribusi.*

1. Pendahuluan

Langkah pertama yang dilakukan untuk memilih variabel yang layak adalah memeriksa korelasi antara variabel prediktor. Tujuannya adalah untuk mendapatkan variabel prediktor yang independen. Jika didapatkan variabel prediktor yang saling berkorelasi, maka hanya variabel prediktor yang mempunyai korelasi terbesar dengan variabel respon yang akan diikutkan dalam pembentukan model, karena variabel prediktor yang nilai korelasinya dengan variabel respon kecil sudah terwakili oleh variabel tersebut.

Selanjutnya membuat model univariat, yaitu dengan meregresikan masing-masing variabel prediktor terhadap variabel respon, untuk mengetahui apakah hubungan antara variabel respon dengan prediktor signifikan atau tidak. Variabel prediktor yang tidak mempunyai hubungan yang signifikan dengan variabel respon akan dikeluarkan dari pembentukan model. Kemudian memodelkan secara serentak (*multivariate*) variabel respon dan semua variabel prediktor yang signifikan pada pemodelan univariate.

* *Jurusan matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin, email: ina_matematika@yahoo.co.id*

2. Model Regresi Logit

Regresi logistik sering digunakan dalam menyelesaikan masalah klasifikasi pada metode parametrik. Metode ini digunakan untuk menggambarkan hubungan variabel dependen (variabel respon) dengan variabel independen (variabel prediktor) yang bersifat kategori, kontinu atau kombinasi keduanya.

Untuk menggambar kondisional mean dari Y (respon) terhadap X (prediktor) digunakan hitungan $\pi(x) = E(Y|x)$. Bentuk dari model logistik adalah sebagai berikut

$$\pi(x) = \frac{\exp(\beta_0 + \beta' x)}{1 + \exp(\beta_0 + \beta' x)}, \qquad (1)$$

dengan $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ dan $X = (x_1, x_2, \ldots, x_p)$, p = jumlah parameter.
Dengan menggunakan transformasi logit dari $\pi(x)$, maka model logistik dapat disebut Model Logit yang ditunjukkan oleh:

$$g(x) = \ln\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \sum_{j=1}^{p} \beta_j x_j \qquad (2)$$

Pada regresi linier berganda diasumsikan bahwa $Y = \pi(x) + \varepsilon$ dimana ε adalah galat error dan menunjukkan selisih obyek pengamatan terhadap nilai harapannya. Galat diasumsikan berdistribusi normal dengan rataan nol variansi tetap terhadap variabel pengamatannya. Sehingga dengan respon yang biner dinyatakan, jika

$Y = 0$, maka $\varepsilon = -\pi(x)$ dengan probabilitas $1 - \pi(x)$,
$Y = 1$, maka $\varepsilon = 1 - \pi(x)$ dengan probabilitas $\pi(x)$.

Dan dapat dinyatakan bahwa ε memiliki $E(\varepsilon) = 0$ dan $\text{var}(\varepsilon) = \pi(x).[1 - \pi(x)]$ yang mengikuti distribusi binomial (Hosmer,1989).

Metode pendugaan yang digunakan untuk mengestimasi parameter yang belum diketahui adalah pendugaan maksimum likelihood (*Maximum Likelihood Estimation* atau MLE) dengan menetapkan asumsi distribusi Bernoulli dan obyek pengamatan saling bebas, $E(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

Pada pasangan pengamatan (x_i, y_i), fungsi likelihood yang dimaksimumkan adalah

$$\varsigma(x_i) = \pi(x_i)^{y_i}\left[1 - \pi(x_i)\right]^{1-y_i}. \qquad (3)$$

Karena pengamatan diasumsikan bersifat independen, maka likelihood pengamatan mrupakan perkalian dari fungsi likelihood masing-masing, misal dinyatakan dengan

$$I(\beta) = \prod_{i=1}^{I} \varsigma(x_i)$$
$$= \prod_{i=1}^{I} \left[\frac{\pi(x_i)}{(1-\pi(x_i))}\right]^{y_i} \left\{1 - \pi(x_i)\right\}^{n_i} \qquad (4)$$

dimana y_i adalah variabel random binomial yang saling bebas, $i = 1, 2, \ldots, I$ dengan $E(y_i) = n_i . \pi(x_i)$ dan $\sum n_i = N$.

Karena $z = \exp(\ln z)$, maka untuk $z = \left[\frac{\pi(x_i)}{(1-\pi(x_i))}\right]^{y_i}$ persamaan (4) dapat ditulis menjadi

$$\left[\prod_{i=1}^{I}\{1+\pi(x_i)\}^{n_i}\right]\exp\left[\sum y_i \ln\left\{\frac{\pi(x_i)}{\{1-\pi(x_i)\}}\right\}\right] \quad (5)$$

kemudian dilakukan transformasi logit terhadap model regresi logistik pada persamaan (5). sehingga didapatkan model :

$$\left[\sum_{i=1}^{I}\left\{1+\exp\left(\sum_j \beta_j x_{ij}\right)\right\}^{-n_i}\right]\left[\exp\left\{\sum_j (y_i x_{ij})\beta_j\right\}\right] \quad (6)$$

Prinsip dari MLE untuk mendapatkan nilai taksiran β adalah dengan memaksimumkan fungsi likelihood. Nilai maksimum dari fungsi persamaan (6) dapat diperoleh melalui transformasi log. Karena fungsi logaritmanya monoton naik, maka

$$\begin{aligned} L(\beta) &= \ln(I(\beta)) \\ &= \left\{\sum_j\left(\sum_i y_i x_{ij}\right)\beta_j\right\} - \sum_i n_i \ln\left\{1+\exp\left(\sum_j \beta_j x_{ij}\right)\right\} \end{aligned} \quad (7)$$

Nilai β diperoleh melalui turunan parsial pertama $L(\beta)$ terhadap β yang disamakan dengan nol, sehingga persamaan (7) akan menjadi :

$$\frac{\partial L(\beta)}{\partial L\beta_a} = \sum_i y_i x_{ia} - \sum_i n_i x_{ia}\pi_i = 0 \quad (8)$$

dimana $a = 0,1,2,\ldots,p$ dengan

$$\pi_i = \left[\frac{\exp\left(\sum_j \beta_i x_{ij}\right)}{1+\exp\left(\sum_j \beta_j x_{ij}\right)}\right]$$

menyatakan taksiran likelihood maksimum dari $\pi(x_i)$. Sedang terhadap varians diperoleh dari turunan parsial kedua dari persamaan (8) yang hasilnya menurut Hogg & Craig (1995) diberikan dalam bentuk berikut,

$$\frac{\partial^2 L(\beta)}{\partial\beta_a \partial\beta_b} = -\sum_i n_i x_{ia} x_{ib}\left[\frac{\exp\left(\sum_j \beta_i x_{ij}\right)}{1+\exp\left(\sum_j \beta_j x_{ij}\right)}\right]$$

$$= -\sum_i n_i x_{ia} x_{ib} \pi_i (1 - \pi_i) \qquad (9)$$

dimana $a, b = 0, 1, 2, \ldots, p$.

3. Model Regresi Probit

Regresi Probit adalah suatu analisis regresi yang digunakan untuk menggambarkan hubungan antara variabel dependen dan variabel independen. Variabel dependen (variabel respon) biasa disimbolkan *Y* dengan skala pengukuran dikotomus (biner), dan variabel independen (variable prediktor) biasa disimbolkan *X* yang skala pengukuran bersifat dikotomus, polikotomus atau kontinu. Jika

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

dan diketahui

	X
$Y = 1$	$P(y = 1 \mid x)$
$Y = 0$	$P(y = 0 \mid x)$

maka

$$P(y = 1/x) = \int^{\beta_0 + \beta_1 x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt = \Phi[\beta_0 + \beta_1 x]. \qquad (10)$$

dengan Φ[.] adalah fungsi kumulatif distribusi normal standar.

$$\Phi^{-1}[P(y = 1|x)] = \Phi^{-1}[\Phi[\beta_0 + \beta_1 x]]$$
$$\Phi^{-1}[P(y = 1|x)] = \beta_0 + \beta_1 x$$

atau $Z = \beta_0 + \beta_1 x_1$ dengan $Z = \Phi^{-1} P(y = 1 \mid x)$.
Dengan cara yang sama jika variabel bebas lebih dari satu, maka:

$$\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{e} \qquad (11)$$

dengan *Z* merupakan variabel yang tidak diobservasi, dan observasinya adalah

$Y = 1$ Jika $Z > 0$
$Y = 0$ jika $Z \leq 0$

dengan ε adalah residual yang diasumsikan berdistribusi normal dengan mean nol (0) dan varians satu (1).
Probabilitas $Y_i = 1$ dari persamaan (1) adalah :

$$P(y_i = 1 \mid x) = \Phi(X_i\beta) \qquad (12)$$
$$P(y_i = 0 \mid x) = 1 - \Phi(X_i\beta) \qquad (13)$$

4. Aplikasi

Model respon biner digunakan pada toxikologi untuk menggambarkan efek dari dosis bahan kimia racun apakah yang dapat menyebabkan kematian. Pada aplikasi ini, konsep dari *toleransi distribusi* menjadi dasar kebenaran untuk model.

Misalkan *x* menunjukkan dosis (atau log dosis) untuk bahan kimia toxic. Untuk menyeleksi subjek secara acak, dengan memisalkan *Y* =1 jika subjek mati, maka andaikan sebuah subjek memiliki toleransi *T* untuk dosis, dengan (*Y*=1) ekuivalen untuk ($T \leq x$). Dalam hal ini, kemungkinan didapatkan serangga yang dapat bertahan dari semprotan racun jika dosis *x* kurang dari *T*, dan mati jika dosis terkecilnya *T*. Merubah toleransi di antara subjek, dan misalkan *G*(*t*)= ($T \leq t$) menunjukkan fungsi distribusi bersyarat (*cumulatif distribution function* atau cdf) untuk distribusi populasinya. Untuk dosis *x* tertentu, peluang penyeleksian subjek yang mati secara acak adalah

$$P(Y = 1) = \pi(x) = P(T \leq x) = G(x) \tag{14}$$

Jika F adalah cdf dari transformasi linier T, sebagaimana standar cdf untuk distribusi keluarga eksponensial jika *G* adalah anggota, maka peluangnya adalah $F(\alpha + \beta x)$.

Dalam beberapa eksperimen toxikologi, distribusi toleransi untuk log dosis adalah pendekatan normal dengan mean μ dan standar deviasi σ. Jika *G* adalah cdf dari distribusi normal, maka

$$\pi(x) = G(x) = \Phi[(x - \mu)/\sigma] \tag{15}$$

dengan Φ adalah standar normal cdf. Dengan mensubtitusi $F = \Phi, \alpha = -\mu/\sigma$ dan $\beta = 1/\alpha$ pada persamaan (15), maka diperoleh

$$\Phi^{-1}[\pi(x)] = \alpha + \beta \tag{16}$$

yang merupakan **model regresi *probit.***

Untuk model probit, grafik respon untuk $\pi(x)$(atau untuk $1 - \pi(x)$, dimana $\beta < 0$) terlihat normal dengan mean $\mu = -\alpha/\beta$ dan standar deviasi $\sigma = 1/|\beta|$. Terdapat 68% dari distribusi normal yang berada dalam standar deviasi dari mean, $1/|\beta|$ adalah jarak antara nilai *x* dimana $\pi(x)$ = 0,16 atau 0,84, dan dimana $\pi(x)$ = 0,50. Nilai peubah di $\pi(x)$ pada nilai partikular *x* adalah $\frac{\partial \pi(x)}{\partial x} = \beta\phi(\alpha + \beta x)$, dimana ϕ adalah fungsi kepadatan normal standar. Nilai tertinggi ketika $+\beta x = 0$ ($x = -\alpha/\beta$), dimana ini akan sama dengan $\beta/(2\pi)^{1/2} = 0{,}48$ dan titik $\pi(x) = 1/2$.

Dengan perbandingan, kurva respon $\pi(x)$ untuk **model regresi logistik** dengan parameter β cocok untuk logistik cdf dengan standar deviasi $\pi/|\beta|\sqrt{3}$. Nilai peubah di $\pi(x)$ pada $x = -\alpha/\beta$ adalah 0,25β. Nilai peubah saat $\pi(x)$=1/2 adalah sama untuk corresponding cdf pada kurva probit dan logistik ketika logistik β adalah 0,40/0,25 = 1,6 kali dari nilai probit β. Standar deviasi akan sama ketika logistik β adalah $\pi/\sqrt{3}$ =1,8 kali probit β. Ketika kedua modelnya pas, parameter penduga pada model logistik adalah berkisar antara 1,6-1,8 kali dari model probit.

Estimasi ML untuk model probit dapat diperoleh dengan menggunakan algoritma scoring Fisher untuk GLM.

Logit dan probit links simetris berkisar 0,5, dari pengertiannya maka

$$link(\pi) = -link(1 - \pi)$$

yang diperoleh dari

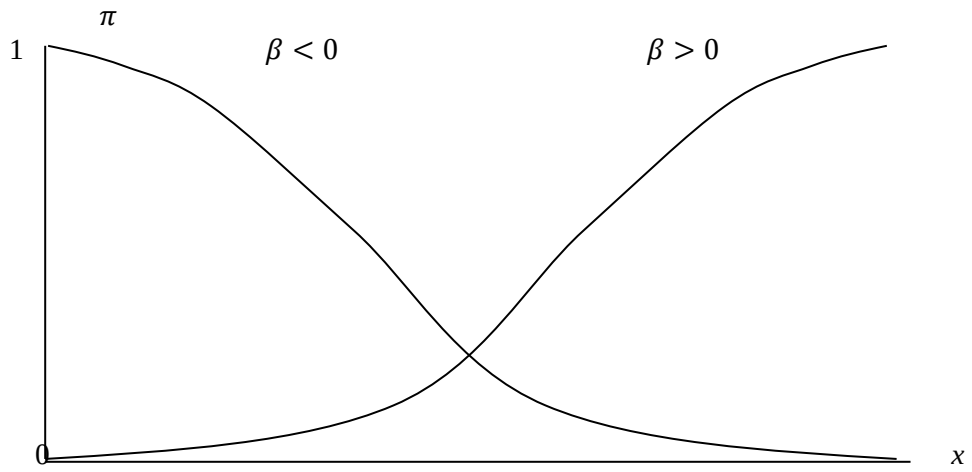
$$logit(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = -\log\left[\frac{1-\pi}{\pi}\right] = -logit(1 - \pi).$$

Secara tidak langsung kurva respon untuk $\pi((x))$ mendekati simetri yang berkisar pada titik 0,5. Faktanya, $\pi(x)$ mendekati 0 dan pada nilai yang sama mendekati 1. Model logit dan probit tidak cocok ketika $\pi(x)$ meningkat dari 0 secara lambat, tapi terkadang tepat pada saat mendekati 1. Kurva dari respon

$$\pi(x) = 1 - \exp[-\exp(\alpha + \beta\ )] \tag{17}$$

yang bentuknya ditunjukkan pada Gambar 1. Gambar ini terlihat tidak simetris, $\pi(x)$ menyimpang dari 1 dan lebih tajam menyimpang dari 0. Untuk model ini,

$$\log[-\log(1 - \pi(x))] = \alpha + \beta x \tag{18}$$



Gambar 1. Kurva Respon Model Logit danProbit

Link untuk GLM disebut *complementary log-log* link. Untuk menginterpretasikan model (18), maka dapat ditentukan nilai x_1 dan x_2, yaitu

$$\log[\ -\log(1 - (\pi(x_2)\ )] - [\ -\log(1 - (\pi(x_1)\ )] = \beta(x_1 - x_2)$$

diperoleh

$$\frac{\log(1-(\pi(x_1)\ )}{\log(1-(\pi(x_2)\ )} = \exp[\beta(x_1 - x_2)]$$

dan

$$1 - \pi(x_2) = (1 - \pi(x_1))^{\exp[\beta(x_1 - x_2)]}.$$

Peluang untuk "kegagalan" pada x_2 sama dengan peluang kegagalan pada x_1 yang ditingkatkan untuk nilai $\exp(\beta)$ yang lebih tinggi untuk setiap kenaikan unit pada jarak $x_2 - x_1$.

Model yang berhubungan adalah

$$\pi(x) = \exp[-\exp(\alpha + \beta\ )] \tag{19}$$

untuk setiap $\pi(x)$ yang berjalan lambat dari 1, namun mendekati 0 dengan tajam. Sebagaimana kenaikan *x*, terjadi perpotongan kurva ketika $\beta > 0$ dan $\beta < 0$. Model ini menggunakan *log-log* link

$$\log[-\log(\pi(x))] = \alpha + \beta x$$

$$G(x) = \exp\{-\exp[-(x - a)/b]\}\ )].$$

Untuk parameter $b > 0$ dan $-\infty < a < \infty$, dan memiliki mean a+0,557b dan standar deviasi $\pi b/\sqrt{6}$. Model log-log link juga cocok menggunakan algoritma scoring Fisher untuk GLM.

Bliss (1935) melaporkan jumlah dari pejuang terbunuh setelah 5 jam pembukaan gas Karbon Disulfide pada konsentrasi yang beragam, sebagaimana yang diberikan pada Tabel 1.

Tabel 1. Jumlah Beetles yang Mati Setelah Pemberian Karbon Disulfide.

			Fitted Values		
Log Dose	No. Beetles	No. Killed	Comp. Log-Log	Probit	Logit
1,691	59	6	5,7	3,4	3,5
1,724	60	13	11,3	10,7	9,8
1,775	62	18	20,9	23,4	22,4
1,784	56	28	30,3	33,8	33,9
1,811	63	52	37,7	49,6	50,0
1,837	59	53	54,2	53,4	53,3
1,861	62	61	61,1	59,7	59,2
1,884	60	60	59,9	59,2	58,8

Sumber: Data dari Bliss(1935).

Model dengan log-log link berimbang memiliki dugaan ML sama untuk $\hat{\alpha} = -39{,}52\ dan\ \hat{\beta} = 22{,}01$. Pada log dosis 1,7, peluang cocok untuk sisa adalah $1 - \hat{\pi} = \exp\{-\exp[-39{,}52 + 22{,}02(1{,}7)]\} = 0{,}885$, sedangkan untuk log dosis 1,8 diperoleh 0,332, sedangkan pada log dosis 1,9 diperoleh 5×10^{-5}. Peluang untuk bertahan hidup adalah exp(22,01×0,1) = 9,03 lebih tinggi untuk setiap 0,1 kenaikan pada log dosis.

5. Kesimpulan

Untuk setiap nilai log dosis, dengan mengalikan peluang estimasi terbunuh dengan jumlah dari pejuang pada level itu, diperoleh nilai laporan yang cocok pada Tabel 1. *Statistic goodness-of-fit* atau kesesuaian model adalah G^2 = 3,5, dengan df=6.

Model logit dan probit cocok digunakan, dimana nilai G^2 adalah 11,1 untuk model logit dan 10,0 untuk model probit. Untuk probit model $\hat{\pi}(x) = \Phi(-34{,}96 + 19{,}74x)$. Yang mana $\hat{\pi} = 0{,}5$ pada $34{,}96/19{,}74 = 1{,}77$. Ini adalah koresponden yang cocok untuk toleransi distribusi normal yang memiliki mean 1,77 dan standar deviasi 1/19,74=0,05.

Daftar Pustaka

[1] Agresti, A., 2002, *Categorical Data Analysis* 2nd *Edition*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] Santosa, B. dan Hanum, D. R., 2007, Studi komparasi metode klasifikasi dua kelas, http://www.elearning.unej.ac.id [26 januari 2009].

[3] Diggle, P.J., Liang, K.Y., dan Zeger, S.L., 1994, *Analysis of Longitudinal Data*, Oxford Science Publication.

[4] Dobson, A.J., 1983, *Introduction To Statistical Modelling*, London, Chapman & Hall. 1st Edition.

[5] Hosmer, D.W. dan Lemeshow, S., 2000, *Applied Logistic Regression* 2nd *Edition*, John Wiley & Sons, Inc.

[6] Hogg, R. V. dan Craig, A. T., 1995, *Introduction to Mathematical Statistics*, Prentice Hall, 5th edition.

[7] Kenward, K.M. dan Smith, D.M., 1995, Computing the generalized estimating equations for repeated measurement, *Genstat Newsletter*.

[8] Kurtner, M.H., 2004, *Applied Linear Regression Models* 4th, Mc Graw Hill Companies.

[9] Montegomery, D.C. dan Peck, E.A, 1992, *Introduction To Linier regression Analysis*, John Wiley & Sons, NIC, 2nd Edition.

[10] McCullagh, P. dan Nelder, J.A., 1989, *Generalized Linear Models*, London: Chapman and Hall, 2nd Edition.

[11] McCullagh, P. dan Nelder, J.A., 1992, *Generalized Linear Models*, London: Chapman and Hall, 3th Edition.