Jurnal EducatiO
Vol. 4 No. 2, Desember 2009, hal. 119-135

REGRESI LOGISTIK UNIVARIAT DENGAN DATA RESPON TIDAK SEIMBANG

Surya Mayadi
STKIP Hamzanwadi Selong

ABSTRACT

In binary logistic regression problem it is common for the two classes to be imbalanced. One case is very rare compared to the other. This paper focus on determining Maximum Likelihood Estimator on univariate logistic regression for infinitely unbalanced data response. By finding the differential of log-likelihood equation that is resulted, its maximum depend on points x given $Y = 1$ through their average $\bar{x}$. Thus we can substitute each x given $Y = 1$ with $\bar{x}$ or substituting all x given $Y = 1$ with one $\bar{x}$. If we use normal distribution for x given $Y = 0$ then the slope β can be determined by simple formula $\hat{\beta} = (\bar{x} - \mu)/\sigma^2$ where μ and σ^2 each an average and variance x for which $Y = 0$.

Keywords : Logistic Regression, MLE, Infinitely Unbalanced Data Response

PENDAHULUAN

Regresi logistik adalah regresi yang menggunakan dua nilai yang berbeda untuk menyatakan variabel responnya (Y). Biasanya digunakan nilai 0 untuk menyatakan kegagalan dan nilai 1 untuk menyatakan kesuksesan.

Dalam aplikasi regresi logistik bisa terjadi salah satu dari dua kelas sangat jarang atau sedikit dibandingkan kelas yang lain. Kejadian peperangan, peristiwa yang berkaitan dengan orang yang menderita penyakit jarang seperti flu babi, jumlah siswa yang tidak lulus di sebuah sekolah yang maju misalnya, di modelkan sebagai kejadian yang jarang. Sedangkan padanannya yaitu keadaan damai, keadaan tidak menderita flu babi dan banyaknya siswa yang lulus disekolah maju, dianggab sebagai kejadian yang umum. Dalam hal ini banyaknya data untuk $Y = 0$ tidak seimbang dibandingkan dengan banyaknya data untuk $Y = 1$.

Jika $Y \in \{0,1\}$ melambangkan satu respon acak dari variabel dependen Y, akan dinotasikan $Y = 1$ sebagai kasus yang jarang dan $Y = 0$ sebagai kasus umum. Andaikan N adalah banyaknya observasi x untuk $Y = 0$ dan n adalah banyaknya observasi x untuk $Y = 1$. Fokus penelitian ini adalah mengungkapkan bagaimana menentukan estimasi maksimum likelihood atau MLE regresi logistik jika data respon tidak seimbang yaitu N jauh lebih besar dibandingkan n.

LANDASAN TEORI

2.1 Distribusi Bernoulli

Definisi 2.1 (Dudewicz & Mishra, 1988)

Suatu variabel random X dikatakan berdistribusi Bernoulli bila untuk suatu p, bernilai $0 \leq p \leq 1$

$$P(X = x) = \begin{cases} P^x(1-P)^{1-x} untuk\ x = 0,1 \\ 0\ untuk\ x\ yang\ lain \end{cases}$$

Suatu peubah acak binom dapat dipandang sebagai jumlah n peubah acak Bernoulli yaitu sebagai banyaknya yang berhasil dalam n usaha Bernoulli.

Definisi 2.2 (Dudewicz & Mishra,1988)

Suatu variabel random X mempunyai distribusi Binomial jika untuk suatu bilangan bulat positif n dan suatu p dengan $0 \leq p \leq 1$:

$$P(X = x) = \begin{cases} \binom{n}{x} p^x(1-p)^{(n-x)}, & x = 0,1,\ldots n \\ 0 \quad untuk\ x\ yang\ lain \end{cases}$$

2.2 Distribusi Normal

Definisi 2.3 (*Bain & Engelhardt, 1992)*

Suatu variabel random X mengikuti distribusi normal dengan mean μ dan variansi σ^2 jika mempunyai fungsi densitas

$$f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

untuk $-\infty < x < \infty$ *dimana* $-\infty < \mu < \infty$ *dan* $0 < \sigma < \infty$. *Ini dituliskan dengan* $X \sim N(\mu, \sigma^2)$.

2.3 Regresi Logistik

Regresi logistik merupakan salah satu solusi yang dapat digunakan untuk menganalisis kasus-kasus penelitian dengan tujuan untuk mencari pola hubungan antara sekumpulan variabel prediktor dengan suatu variabel respon bertipe katagorik atau kualitatif secara simultan.

Regresi logistik biner khususnya, adalah regresi dimana variabel respon hanya memiliki dua kemungkinan nilai (dikotomis) misalnya ya atau tidak, sukses atau gagal, sehat atau sakit dan sebagainya.

Jika probabilitas suatu peristiwa untuk terjadi adalah $\pi(\boldsymbol{x})$, maka probabilitas suatu peristiwa itu untuk tidak terjadi adalah $1-\pi(\boldsymbol{x})$, dan odds adalah $\pi(\boldsymbol{x})\ /(1-\pi(\boldsymbol{x}))$. Secara khusus :

$$logit\left(\pi(\boldsymbol{x}_i)\right) = ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \qquad (2.3.1)$$

dengan $\pi(\boldsymbol{x}_i) = E(Y = 1|\boldsymbol{x}_i)$

β_0 = konstanta, yang lazim disebut intersep

$\beta_1, \ldots, \beta_p$ = koefisien regresi variabel prediktor yang lazim disebut slope

$x_1, \ldots, x_p$ = variabel prediktor yang pengaruhnya akan diteliti

(*diacu dari* ***Alfred DeMaris, 2004***)

Model di atas menganggap logit dari peristiwa tersebut linear terhadap x_i. Probabilitas suatu peristiwa terjadi jika diberikan nilai-nilai yang spesifik dari x_i bisa dihitung dengan rumus :

$$\pi(\boldsymbol{x}_i) = \frac{exp\left(f(\boldsymbol{x}_i)\right)}{1 + exp\left(f(\boldsymbol{x}_i)\right)} \qquad (2.3.2)$$

2.4. Regresi Logistik Univariat

Fokus dari penelitian ini adalah regresi logistik biner dengan satu variabel bebas yang kadang-kadang disebut regresi logistik sederhana atau regresi logistik univariat. Bentuk persamaan regresinya menjadi lebih sederhana yaitu :

$$\pi(x_i) = \frac{exp(\alpha + x_i\beta)}{1 + exp(\alpha + x_i\beta)} \qquad (2.4.1)$$

(*diacu dari* ***Gary King*** *and* ***Lanche Zeng, 2000***)

2.5. Metode Maksimum Likelihood

Metode untuk mengestimasi parameter regresi logistik adalah dengan menggunakan metode *maximum likelihood* . Metode ini memperoleh dugaan maksimum likelihood bagi $\boldsymbol{\beta}$ dengan iterasi Newton-Raphson.

Definisi 2.4 (Bain and Engelhardt,1992)

Fungsi densitas bersama dari n variabel $X_1, X_2, \ldots, X_n$ *adalah* $f(x_1, x_2, \ldots, x_n ; \theta)$ *disebut fungsi likelihood* $L(\theta)$. *Jika* $X_1, X_2, \ldots, X_n$ *adalah representasi sampel random dari* $f(x_i; \theta)$ *maka* $L(\theta) = f(x_1; \theta) f(x_2; \theta) \ldots f(x_n; \theta)$.

Sedangkan *maximum likelihood estimate* (MLE) didefinisikan atas dasar definisi fungsi likelihood sebagai berikut.

Defenisi 2.5 (Bain and Engelhardt,1992)

Bila $f(x_1, x_2, \ldots, x_n ; \boldsymbol{\beta})$, $\boldsymbol{\beta} \in S$ *sebagai fungsi distribusi probabilitas bersama dari* $x_1, x_2, \ldots, x_n$ *maka nilai* $\tilde{\boldsymbol{\beta}}$ *dalam* $\boldsymbol{\beta}$ *dimana* $L(\boldsymbol{\beta})$ *maksimum disebut estimasi maksimum likelihood dari* β *yang dinyatakan dengan :*

$$f(x_1, x_2, \ldots, x_n \,|\tilde{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta} \in S} f(x_1, x_2, \ldots, x_n \,|\boldsymbol{\beta}) \qquad (2.5.1)$$

Nilai $\boldsymbol{\beta}$ yang memaksimalkan $L(\boldsymbol{\beta})$ juga akan memaksimalkan log-likelihood $ln\, L(\boldsymbol{\beta}) = \boldsymbol{\ell}(\boldsymbol{\beta})$. Untuk memperoleh $\tilde{\boldsymbol{\beta}}$ yang memaksimumkan $\ell\,(\boldsymbol{\beta})$ diperoleh dengan menderivatifkan $L(\boldsymbol{\beta})$.

Teorema 2.6

Misalkan suatu sampel terdiri dari n observasi dari pasangan (X_i, Y_i) *dan model regresi logistik adalah:*

$$\pi(\boldsymbol{x}_i) = \frac{exp(\beta_0 + x_{1i}\beta_1 + \cdots + x_{pi}\beta_p)}{1 + exp(\beta_0 + x_{1i}\beta_1 + \cdots + x_{pi}\beta_p)}$$

maka penduga $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_n)$ *dengan menggunakan metode maksimum likelihood adalah penyelesaian dari persamaan likelihood :*

$$\sum_{i=1}^{n}(Y_i - \pi(\boldsymbol{x}_i)) = 0 \qquad (2.5.2)$$

dan

$$\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ji}(Y_i - \pi(\boldsymbol{x}_i)) = 0 \qquad (2.5.3)$$

Bentuk $\frac{\partial \boldsymbol{\ell}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ pada (2.5.2) dan (2.5.3) dapat ditulis :

$$\frac{\partial \boldsymbol{\ell}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{x}(\boldsymbol{y} - \boldsymbol{\pi}_i) \qquad (2.5.4)$$

Definisi 2.7 (Budhi, 2001)

Misalkan $f(x_1, x_2, \ldots, x_n)$ *fungsi n variabel yang mempunyai turunan parsial kedua dan misalkan* $x_0 = (h_1, h_2, \ldots h_n)$ *merupakan titik kritis fungsi* f *yaitu titik yang memenuhi sistem persamaan*

$$\frac{\partial f(x)}{\partial x_1} = 0, \quad \frac{\partial f(x)}{\partial x_1} = 0, \quad \ldots, \frac{\partial f(x)}{\partial x_n} = 0$$

Kemudian bentuk matriks Hessian fungsi f

$$H = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial {x_1}^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \cdots & \cdots & & \cdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial {x_n}^2} \end{pmatrix} \qquad (2.5.5)$$

yaitu matriks dengan $h_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ *. Jika matriks* H *definit positif maka titik* x_0

merupakan titik minimum dan jika H *definit negatif maka* x_0 *merupakan titik maksimum.*

2.6. Metode Newton-Raphson

Metode Newton-Raphson adalah suatu metode yang menggunakan pendekatan deret Taylor untuk menyelesaikan persamaan non linier. Misalkan penaksir awalnya $\boldsymbol{\beta}^{(0)}$, akan dicari estimasi $\boldsymbol{\beta}$ sebagai parameter tunggal yaitu $\tilde{\boldsymbol{\beta}}$ yang memaksimumkan fungsi $L(\boldsymbol{\beta})$. Jika digunakan pendekatan deret Taylor orde ke-2 pada $\boldsymbol{\beta}^{(0)}$ akan diperoleh persamaan:

$$L(\boldsymbol{\beta}) = L(\boldsymbol{\beta}^{(0)}) + L'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}L''(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^2 + \frac{1}{6}L'''(\boldsymbol{\beta}^{*})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^3 \qquad (2.6.1)$$

untuk suatu $\boldsymbol{\beta}^{*}$ yang bernilai antara $\boldsymbol{\beta}^0$ dan $\boldsymbol{\beta}$.

Nilai maksimum diperoleh jika $L'(\boldsymbol{\beta}) = 0$ maka dengan demikian:

$$L'(\boldsymbol{\beta}^0) + L''(\boldsymbol{\beta}^0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \frac{1}{2}L'''(\boldsymbol{\beta}^{*})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^2 = 0 \qquad (2.6.2)$$

Sehingga apabila $\boldsymbol{\beta}^0$ cukup dekat ke $\tilde{\boldsymbol{\beta}}$ maka persamaan (2.5.2) dapat didekati dengan menyelesaikan persamaan $L'(\boldsymbol{\beta}^0) + L''(\boldsymbol{\beta}^0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = 0$ sehingga didapat penyelesaian persamaan :

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^0 - \frac{L'(\boldsymbol{\beta}^0)}{L''(\boldsymbol{\beta}^0)} \qquad (2.6.3)$$

dengan $\boldsymbol{\beta}^{(1)}$ pendekatan untuk $\tilde{\boldsymbol{\beta}}$. Dari persamaan (2.5.3) akan diperoleh nilai-nilai pendekatan berikutnya penaksir awal secara iterasi.

Jika $\boldsymbol{\beta}^{(t)}$ menyatakan pendekatan ke$-$t untuk $\tilde{\boldsymbol{\beta}}$ maka:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \frac{L'(\boldsymbol{\beta}^{(t)})}{L''(\boldsymbol{\beta}^{(t)})} \qquad (2.6.4)$$

Lemma 2.8

Nilai penaksir $\boldsymbol{\beta}$ *dengan menggunakan metode Newton–Raphson pada langkah ke* $-t$ *adalah :*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left\{\boldsymbol{x}Diag\left(\pi_i{}^{(t)}(1 - \pi_i{}^{(t)})\boldsymbol{x}^T\right\}^{-1}\boldsymbol{x}(\boldsymbol{y} - \pi_i{}^{(t)})\right. \qquad (2.6.5)$$

2.7 Masalah Kekonvekan

Masalah kekonvekan ini memegang peranan cukup penting dalam regresi logistik terutama menyangkut penyelesaian MLE regresi logistik yang diperoleh yang mengharuskan bahwa *konveks hull* dari kumpulan data $\boldsymbol{x}$ untuk $Y = 1$ harus beririsan dengan *konveks hull* dari kumpulan data $\boldsymbol{x}$ untuk $Y = 0$.

Definisi 2.9 (Bazaraa dkk., 1993)

Kombinasi konveks adalah kombinasi linier dari titik-titik $x_1, x_2, \dots, x_k$ *yang berbentuk* $x = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k$ *dengan* $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$ *dan* $\lambda_i \geq 0$ *untuk semua* i *dari 1 sampai* k.

Definisi 2.10 (Bazaraa dkk., 1993)

Himpunan C *dalam ruang vektor* $\mathbb{R}^n$ *dikatakan konveks jika garis lurus penghubung sembarang dua titik pada* C *terletak pada himpunan* C *itu sendiri. Dengan perkataan lain himpunan* C *dalam* $\mathbb{R}^n$ *dikatakan konveks jika* $x, y \in C$ *dan* $0 \leq \lambda \leq 1$ *maka* $\lambda x + (1 - \lambda) y \in C$.

Definisi 2.11 (Bazaraa dkk., 1993)

Konveks Hull S *dilambangkan dengan* $H(S)$ *adalah himpunan dari semua kombinasi konveks dari* S. *Dengan kata lain,* $\boldsymbol{x} \in H(S)$ *jika dan hanya jika* $\boldsymbol{x}$ *dapat direpresentasikan sebagai* $\boldsymbol{x} = \sum_{j=1}^{k} \lambda_j x_j$ *dimana* $\sum_{j=1}^{k} \lambda_j = 1$.

Definisi 2.12 (Leon, 1998)

Diberikan A *matriks simetris real berukuran* $n \times n$.

1. *Matriks* A *dikatakan semidefinit positif jika* $\boldsymbol{x}^T A \boldsymbol{x} \geq 0$ *untuk setiap vektor tak nol* $\boldsymbol{x} \in \mathbb{R}^n$ *dan sekurang-kurangnya terdapat satu* $x \neq 0$ *sehingga* $\boldsymbol{x}^T A \boldsymbol{x} = 0$.
2. *Matriks* A *dikatakan definit positif jika* $\boldsymbol{x}^T A \boldsymbol{x} > 0$ *untuk setiap vektor tak nol* $\boldsymbol{x} \in \mathbb{R}^n$.

Teorema berikut digunakan untuk menyelidiki matriks semidefinit positif dan matriks definit positif.

Teorema 2.13 (Leon, 1998)

Diberikan A *matriks simetris real berukuran* $n \times n$

1. *Matriks* A *semidefinit positif jika dan hanya jika nilai eigen* $\lambda_i \geq 0$ *untuk* $i = 1,2,\dots,n$.
2. *Matriks* A *definit positif jika dan hanya jika nilai eigen* $\lambda_i > 0$ *untuk* $i = 1,2,\dots,n$.

Definisi 2.14 (Bazaraa dkk.,1993)

Misal C *adalah himpunan konveks tidak kosong dalam* $\mathbb{R}^n$ *dan* $f: C \to \mathbb{R}$. *Fungsi* f *dikatakan fungsi konveks pada* C *jika* $x, y \in C$ *dan* $0 \leq \lambda \leq 1$ *maka*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)\,. \qquad (2.7.1)$$

Fungsi f dikatakan fungsi konveks sempurna jika tanda $\leq$ pada (2.7.1) diganti dengan tanda $<$. Selanjutnya suatu fungsi f dikatakan konkav pada C jika dan hanya jika fungsi $-f(x)$ adalah konveks pada C. Dengan kata lain fungsi f dikatakan konkav sempurna jika $f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda) f(y)$.

Teorema 2.15

Diberikan $f: K \subseteq \mathbb{R}^n \to \mathbb{R}$, $K \neq \emptyset$, K *Himpunan konveks dan* $f \in C^2$. *Fungsi f konveks pada K jika dan hanya jika matriks Hessian f semidefinit positif untuk setiap* $x \in K$.

(Bukti, lihat Edwin K.P. Chong dan Stanislaw, 2008)

Teorema 2.16

Misalkan f(x,y) fungsi dua variabel yang mempunyai turunan ketiga yang kontinyu. Misalkan (a,b) merupakan titik kritis f dan $\Delta = f_{xx}(a,b) f_{yy}(a,b) - \left[f_{yy}(a,b)\right]^2$ *disebut diskriminan fungsi f. Jika* $\Delta > 0$ *dan*

i. $f_{xx}(a,b) > 0$, *maka titik (a,b) merupakan titik minimum lokal.*

ii. $f_{yy}(a,b) < 0$, *maka titik (a,b) merupakan titik maksimum lokal.*

(Bukti lihat Budi, 2001)

Teorema (2.24) menyatakan bahwa jika matrik Hessian f definit positif maka titik kritisnya merupakan titik minimum lokal sedangkan jika matrik Hessian f definit negatif maka titik kritisnya merupakan titik maksimum lokal.

Teorema 2.17

Diberikan $f: K \subseteq \mathbb{R}^n \to \mathbb{R}$, $K \neq \emptyset$, K *himpunan konveks dan* $f \in C^2(K, \mathbb{R})$. *Jika matriks Hessian H*$(\bar{x})$ *definit positif untuk setiap* $x \in K$ *maka f fungsi konveks tegas.*

(Bukti, lihat Bazaraa, dkk, 1993)

Teorema 2.18

Diberikan $f: K \subseteq \mathbb{R}^n \to \mathbb{R}$, $K \neq \emptyset$, K *himpunan konveks. Jika* f *fungsi konveks tegas pada* K *maka titik minimum lokal* $\bar{x}$ *dari* f *sekaligus merupakan titik minimum global dan tunggal.*

(Bukti, lihat Bazaraa, dkk, 1993)

Dari teorema (2.24), (2.25) dan (2.26), dapat dikatakan bahwa jika suatu fungsi konveks tegas dan mempunyai titik minimum lokal maka titik tersebut merupakan titik minimum global. Analog dengan itu, jika suatu fungsi konkav tegas dan mempunyai titik maksimum lokal maka titik tersebut merupakan titik maksimum global.

2.8. Kondisi Overlap

Seperti di ketahui bahwa MLE yang digunakan pada regresi logistik menjadi gagal jika nilai-nilai $\boldsymbol{x}$ untuk $Y = 1$ dapat dipisah secara linier dari nilai-nilai $\boldsymbol{x}$ untuk $Y = 0$. Keberadaan dan ketunggalan MLE untuk regresi logistik linier dikarakterisasi oleh Silvapulle(1981) yang mengemukakan bahwa Jika $H_0 \cap H_1 \neq \emptyset$

maka MLE regresi logistik yang tunggal dan berhingga yaitu $(\hat{\alpha}, \hat{\beta}')$ ada. Tetapi jika $H_0 \cap H_1 = \emptyset$ maka MLE tidak ada. Disini H_0 dan H_1 adalah K*onveks hull* dari himpunan data masing-masing untuk $Y = 0$ dan $Y = 1$.

Mengingat hasil Silvapulle tersebut maka diperlukan untuk menganggab adanya overlap antara data $x_1, x_2, \dots, x_n$ diberikan $Y = 1$ dan distribusi F_0 dari $\boldsymbol{x}$ diberikan Y = 0 untuk mendapatkan penyelesaiannya. Dengan membuat N mendekati tak terhingga keadaannya berbeda tapi masih memerlukan kondisi overlap tersebut. (*pernyataan tentang kondisi overlap diacu dari* ***Owen, 2007****)*

PEMBAHASAN

3.1 MLE Regresi Logistik dengan Data Respon Tidak Seimbang

Dalam penelitian ini data yang terdiri dari satu variabel bebas dinyatakan dengan pasangan (x, y) dimana $x \in \mathbb{R}$ dan $y \in \{0,1\}$. Andaikan ada n observasi untuk $Y = 1$ dan N observasi untuk $y = 0$. Nilai–nilai x untuk $Y = 1$ misalkan adalah $x_{11}, x_{12}, \dots, x_{1n}$ sedangkan nilai-nilai x untuk $Y = 0$ adalah $x_{01}, x_{02}, \dots, x_{0N}$.

Sesuai dengan definisi (2.1) setiap titik (x_i, y_i) mempunyai fungsi peluang $(\pi(x_i))^{y_i}(1 - \pi(x_i))^{1-y_i}$ sehingga fungsi likelihoodnya adalah :

$$L(\alpha,\beta) = \Pi(\pi(x_i))^{y_i}(1 - \pi(x_i))^{1-y_i}$$

$$= \frac{e^{\sum_{i=1}^{n} \alpha + x_{1i}\beta}}{\prod_{i=1}^{n}(1 + e^{\alpha + x_{1i}\beta}) \prod_{i=1}^{N}(1 + e^{\alpha + x_{0i}\beta})}$$

selanjutnya fungsi log-likelihoodnya adalah :

$$\boldsymbol{\ell}(\alpha,\beta) = \log(L(\alpha,\beta))$$
$$= \sum_{i=1}^{n}(\alpha + x_{1i}\beta) - \sum_{i=1}^{n} log(1 + e^{\alpha + x_{1i}\beta}) - \sum_{i=1}^{N} log(1 + e^{\alpha + x_{0i}\beta})$$
$$= \sum_{i=1}^{n}\left((\alpha + x_{1i}\beta) - log(1 + e^{\alpha + x_{1i}\beta})\right) - \sum_{i=1}^{N} log(1 + e^{\alpha + x_{0i}\beta}) \quad (3.1.1)$$

Jika x berdistribusi kontinyu maka kita dapat mengganti $\sum_{i=1}^{N} log\left(1+e^{\alpha+x_{0i}\beta}\right)$ pada persamaan (3.1.1) dengan $N\int log\left(1+e^{\alpha+x\beta}\right)f_0(x)dx$ dimana f_0 adalah suatu fungsi distribusi peluang $\boldsymbol{x}$ untuk $Y=0$. Karena beberapa atau semua komponen $\boldsymbol{x}$ kemungkinannya adalah diskrit, F_0 akan digunakan sebagai fungsi distribusi peluang $\boldsymbol{x}$ untuk $Y=0$ tersebut.

Dengan sedikit perubahan pada hasil terakhir yang didapat jika dilakukan pemusatan disekitar rata-rata untuk nilai variabel prediktor $\boldsymbol{x}$ untuk $Y=1$, maka log likelihoodnya akan menjadi :

$$\boldsymbol{\ell}(\alpha,\beta) = n\alpha - \sum_{i=1}^{n} log\left(1+e^{\alpha+(x_{1i}-\bar{x})\beta}\right) - N\int log\left(1+e^{\alpha+(x-\bar{x})\beta}\right)dF_0(x) \qquad (3.1.2)$$

3.2 Lemma-Lemma Pendukung

Lemma 3.1

Untuk $\alpha, z \in \mathbb{R}$ *maka pertidaksamaan berikut ini berlaku :*

$$e^{\alpha+z} \geq log(1+e^{\alpha+z}) \geq \left[log(1+e^{\alpha}) + \frac{ze^{\alpha}}{(1+e^{\alpha})}\right]_+ \geq \left[\frac{ze^{\alpha}}{(1+e^{\alpha})}\right]_+ = \frac{z_+e^{\alpha}}{(1+e^{\alpha})} \qquad (3.2.1)$$

Lemma 3.2

Misalkan $n \geq 1$ *dan* $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$. *Jika diasumsikan bahwa distribusi* F_0 *mengelilingi* $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ *dan bahwa* $0 < N < \infty$ *maka log-likelihood* $\boldsymbol{\ell}(\alpha,\beta)$ *mempunyai pembuat maksimum berhingga* $(\hat{\alpha},\hat{\beta})$.

3.3 Hasil Utama

Teorema 3.3

Jika $n \geq 1$ *dan* $x_1, x_2, \ldots, x_n \in \mathbb{R}$ *tertentu dan anggab bahwa* F_0 *memenuhi tail condition (3.2.3) dan mengelilingi* $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ *, maka pembuat maksimum* $(\hat{\alpha}, \hat{\beta})$ *dari* ℓ *pada persamaan (3.1.2) memenuhi :*

$$\lim_{N \to \infty} \frac{\int e^{x\hat{\beta}} x \, dF_0(x)}{\int e^{x\hat{\beta}} \, dF_0(x)} = \bar{x} \qquad (3.3.1)$$

Bukti

Dengan membuat $\partial \boldsymbol{\ell} / \partial \beta = 0$ persamaan (3.1.2) menjadi :

$$\frac{\partial \boldsymbol{\ell}}{\partial \beta} = -\sum_{i=1}^{n} \frac{(x_{1i} - \bar{x}) e^{\hat{\alpha} + (x_{1i} - \bar{x})\hat{\beta}}}{1 + e^{\hat{\alpha} + (x_{1i} - \bar{x})\hat{\beta}}} - N \int \frac{(x - \bar{x}) e^{\hat{\alpha} + (x - \bar{x})\hat{\beta}}}{1 + e^{\hat{\alpha} + (x - \bar{x})\hat{\beta}}} dF_0(x) = 0$$

kemudian dilakukan pembagian dengan $Ne^{\hat{\alpha} - \bar{x}\hat{\beta}}$ diperoleh :

$$\int \frac{(x - \bar{x}) e^{x\hat{\beta}}}{1 + e^{\hat{\alpha} + (x - \bar{x})\hat{\beta}}} \, dF_0(x) = -\frac{1}{N} \sum_{i=1}^{n} \frac{e^{x_{1i}\hat{\beta}} (x_{1i} - \bar{x})}{1 + e^{\hat{\alpha} + (x_{1i} - \bar{x})\hat{\beta}}} \qquad (3.3.2)$$

Jika $N \to \infty$ akan membuat sisi kanan persamaan (3.3.2) diatas menjadi nol karena menurut lemma (3.3), $\|\hat{\beta}\|$ terbatas jika N mendekati tak terhingga. Oleh karena itu MLE – nya memenuhi persamaan $\int \frac{\left(x e^{x\hat{\beta}} - \bar{x} e^{x\hat{\beta}}\right)}{1 + e^{\hat{\alpha} + (x - \bar{x})\hat{\beta}}} \, dF_0(x) = 0$ atau $\int \frac{x e^{x\hat{\beta}} dF_0(x)}{1 + e^{\hat{\alpha} + (x - \bar{x})\hat{\beta}}} = \int \bar{x} e^{x\hat{\beta}} dF_0(x)$ yaitu $\frac{\int e^{x\hat{\beta}} x \left[1 + e^{\hat{\alpha} + (x - \bar{x})\hat{\beta}}\right]^{-1} dF_0(x)}{\int e^{x\hat{\beta}} \, dF_0(x)} = \bar{x}$. Sehingga untuk $N \to \infty$ berlaku $\frac{\int e^{x\hat{\beta}} x \, dF_0(x)}{\int e^{x\hat{\beta}} \, dF_0(x)} = \bar{x}$ ■

3.4 Distribusi Normal F_0

Pada bagian ini akan digunakan distribusi normal sebagai F_0 pada persamaan (3.3.1). Misalkan bahwa $F_0 = N(\mu, \sigma^2)$ maka :

$$\int e^{x\hat{\beta}} \, dF_0(x) = e^{\mu\hat{\beta} + \frac{1}{2}\hat{\beta}^2\sigma^2} \frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}z^2} dz$$

$$= e^{\mu\hat{\beta} + \frac{1}{2}\hat{\beta}^2\sigma^2} \qquad (3.4.1)$$

Selanjutnya :

$$\int e^{x\hat{\beta}} x \, dF_0(x) = e^{\mu\hat{\beta}+\frac{1}{2}\hat{\beta}^2\sigma^2} \frac{1}{\sqrt{2\pi}} \int (\sigma z + \mu + \hat{\beta}\sigma^2) e^{-\frac{1}{2}z^2} dz$$

$$= (\mu + \hat{\beta}\sigma^2) e^{\mu\hat{\beta}+\frac{1}{2}\hat{\beta}^2\sigma^2} \quad (3.4.2)$$

Dengan mensubstitusikan persamaan (3.4.1) dan (3.4.2) ke persamaan (3.3.1) diperoleh :

$$\hat{\beta} = \frac{\bar{x} - \mu}{\sigma^2} \quad (3.4.3)$$

4. Data Simulasi

4.1 Hasil Data Simulasi

Tabel 1. slope(β) regresi logistik

N	Data	Slope (β)	Intersep (α)	Persen Simpangan terhadap β dan α Data Asli	
				β	α
100	Asli	-1.21235	0.615333		
	Ganti setiap x untuk $Y = 1$ (x_{1i}) dengan $\bar{x}$	-1.65363	1.613998	36.3983	162.2967
	Ganti semua x untuk $Y = 1$ (x_{1i})dengan satu $\bar{x}$	-1.3798	-1.31026	13.8118	312.9351
500	Asli	-1.05664	-1.28075		
	Ganti setiap x untuk $Y = 1$ (x_{1i}) dengan $\bar{x}$	-1.10532	-1.16515	4.6076	9.0259
	Ganti semua x untuk $Y = 1$ (x_{1i})dengan satu $\bar{x}$	-1.06154	-3.57243	0.4646	178.9326
1000	Asli	-1.01149	-2.10594		
	Ganti setiap x untuk $Y = 1$ (x_{1i}) dengan $\bar{x}$	-1.03297	-2.05505	2.1231	2.4168
	Ganti semua x untuk $Y = 1$ (x_{1i})dengan satu $\bar{x}$	-1.01373	-4.4036	0.2211	109.1034
5000	Asli	-0.98946	-3.7344		
	Ganti setiap x untuk $Y = 1$ (x_{1i}) dengan $\bar{x}$	-0.99317	-3.7254	0.3757	0.2408

	Ganti semua x untuk $Y = 1$ (x_{1i})dengan satu $\bar{x}$	-0.98865	-6.0389	0.0817	61.7103
10000	Asli	-0.99783	-4.40694		
	Ganti setiap x untuk $Y = 1$ (x_{1i}) dengan $\bar{x}$	-0.99974	-4.40232	0.1915	0.1048
	Ganti semua x untuk $Y = 1$ (x_{1i})dengan satu $\bar{x}$	-0.99737	-6.71061	0.046	52.2736

Tabel 1 menunjukkan bahwa slope(β) regresi logistik dari data asli cukup dekat dengan slope regresi logistik jika dilakukan penggantian setiap x_{1i} dengan $\bar{x}$ dan penggantian semua x_{1i} dengan satu nilai $\bar{x}$. Berbeda dengan slope, intersep cukup dekat ke intersep data asli jika dilakukan penggantian setiap x_{1i}.

Dari karakteristik nilai β dan α yang dihasilkan pada tabel tersebut nampak bahwa β lebih dekat dengan β data asli jika dilakukan penggantian semua x_{1i} dengan satu $\bar{x}$. Sedangkan α lebih dekat dengan α data asli jika dilakukan penggantian setiap x_{1i} dengan $\bar{x}$.

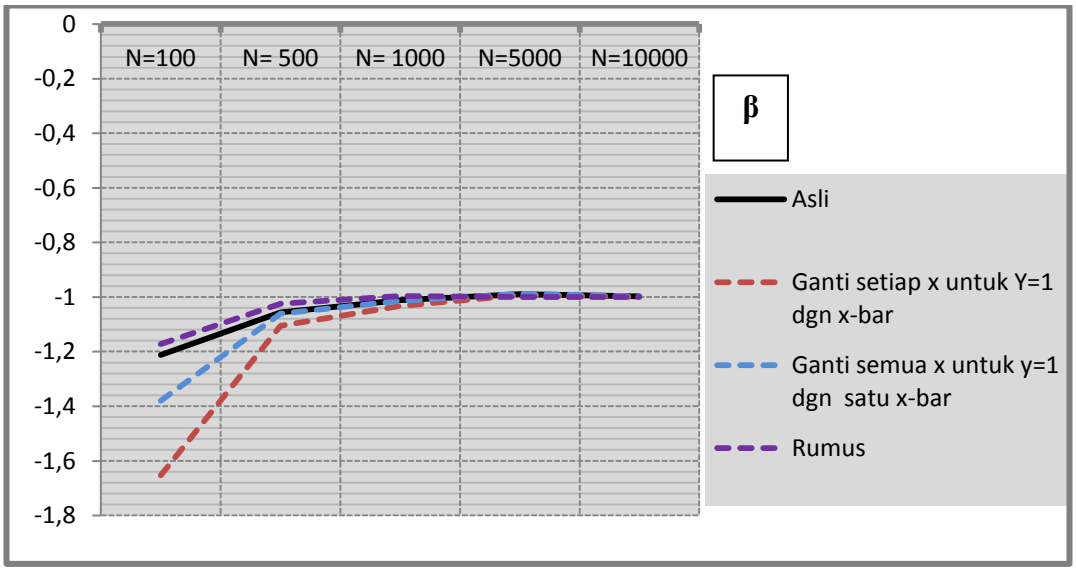
Hasil yang diperoleh lebih lanjut disajikan dalam tabel 2 berikut :

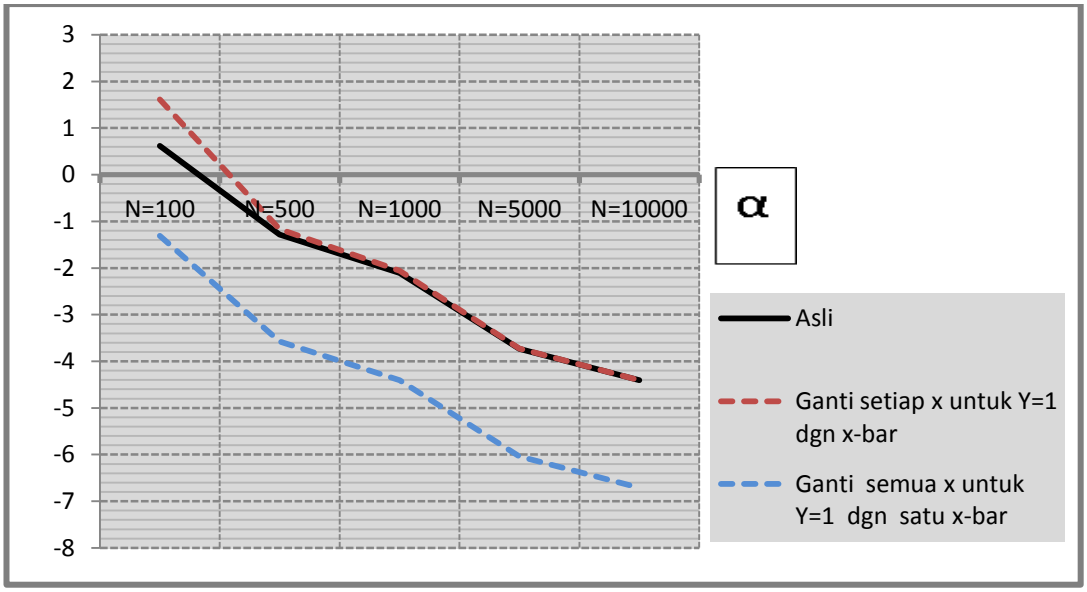
N	Nilai β		Persen Penyimpangan
	Asli	Rumus	
100	-1.21235	-1.17218	3.3136
500	-1.05664	-1.02485	3.0084
1000	-1.01149	-0.99676	1.4563
5000	-0.98946	-0.99961	1.0265
10000	-0.99783	-0.99989	0.2060

\

Dari tabel 2 terlihat bahwa jika digunakan rumus $\hat{\beta} = \frac{(\bar{x}-\mu)}{\sigma^2}$ slope(β) cukup dekat dengan β data asli. Terlihat bahwa persen penyimpangan semakin kecil dengan bertambah besarnya N.

Hasil data simulasi di atas dapat diilustrasikan dengan grafik dibawah ini:





Gambar 1. Grafik Nilai intersep untuk setiap data untuk N yang bertambah besar

KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Estimasi maksimum likelihood (MLE) pada regresi logistik dengan data respon sangat tidak seimbang ditentukan dengan mencari nilai maksimum persamaan $\ell(\alpha,\beta) = n\alpha - \sum_{i=1}^{n} log\left(1 + e^{\alpha+(x_{1i}-\bar{x})\beta}\right)$

$-N\int log\left(1+e^{\alpha+(x-\bar{x})\beta}\right)\,dF_0(x)$ yang penyelesaiannya dalam bentuk persamaan $\lim_{N\to\infty}\frac{\int e^{x\hat{\beta}}x\,dF_0(x)}{\int e^{x\hat{\beta}}\,dF_0(x)}=\bar{x}$.

2. Koefisien persamaan regresi logistik dengan data respon sangat tidak seimbang dapat ditentukan dengan mengganti setiap $\boldsymbol{x}$ untuk $Y=1$ (x_{1i}) masing-masing dengan $\bar{x}$ atau dengan mengganti semua $\boldsymbol{x}$ untuk $Y=1$ (x_{1i}) dengan satu $\bar{x}$.
3. Koefisien regresi logistik dengan data respon sangat tidak seimbang dapat ditentukan dengan rumus $\hat{\beta}=\frac{\bar{x}-\mu}{\sigma^2}$ dimana $\bar{x}$ adalah rata-rata $\boldsymbol{x}$ untuk $Y=1$, μ adalah rata-rata $\boldsymbol{x}$ untuk $Y=0$ dan σ^2 adalah variansi $\boldsymbol{x}$ untuk $Y=0$. Semakin besar N, nilai β yang diperoleh dengan rumus makin dekat dengan β regresi logistik data asli.

5.2. Saran

Perlu pengkajian yang lebih mendalam terhadap pembahasan ini terutama menyangkut regresi logistik multivariat dengan variabel respon dengan dua katagori (biner) maupun lebih dari dua katagori (multinomial). Juga menyangkut kemungkinan menggunakan F_o yang lain misalnya distribusi Cauchy yang mempunyai ekor yang lebih besar dari ekor distribusi normal.

Pengkajian yang lebih mendalam juga dapat dilakukan dengan mencari seberapa besar persentase penyimpangan koefisien ditentukan dari besarnya perbandingan antara banyaknya data untuk $Y=1$ dan banyaknya data untuk $Y=0$ sehingga nantinya untuk setiap nilai perbandingan dihasilkan nilai persentase penyimpangan.

DAFTAR PUSTAKA

Alfred DeMaris. (2004). *Regression with Social Data. John Wiley & Sons*, New York.

Art B. Owen. (2007). Infinitely Imbalanced Logistic Regression. *Journal of Machine Learning Research* 8(2007) 761-773.

Bain L. J dan Engelhardt M..(1992). *Introduction to Probability and Mathematical Statistics*, Duxbury Press: Belmont, California.

Budhi W.S.. (2001). *Kalkulus Peubah banyak dan penggunaannya*, ITB, Bandung

Draper,N.. (1992). *Analisis Regresi Terapan.* Gramedia Pustaka Utama, Jakarta.

Edward J. Dudewicz/Satya N Mishra. (1988). Modern Mathematical Statistics, Alih bahasa : RK Semebiring,1995, ITB Bandung.

Edwin K.P. Chong dan Stanislaw H. Zak. (2008). *An Intoduction to Optimization third edition*, John Wiley & Sons, Inc.,Hoboken, New Jersey

Gary King and Lanche Zeng. (2000). *Logistic Regression in Rare Events Data*, The Global Burden of Disease 2000 in Aging Populations. Research Paper No. 2.

Leon, J.S.. (1998). *Aljabar Linear dan Aflikasinya*, Edisi kelima, Alih bahasa oleh Bondan, A. Erlangga, Jakarta.

Mokhtar S. Bajaraa, Hanif D. Sherali and C.M. Shetty. (1993). *Non Linear Programming*, ,John Wiley & Sons, New York.

Montgomery,D.. (1961). *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York.

Munir R.. (2003). *Metode Numerik*, Informatika, Bandung.

Searle S.R. (1970) *Linear Models*, John Wiley & Sons, New York .

Setya Budi, Wono. (2001). *Kalkulus Peubah Banyak dan Penggunaannya*, ITB Bandung.

Simonof, J.. (2003). *Analyzing Categorical Data*, Springer-Verlag, New York.

Walpole,Ronald E. dan Myers, Raymond H. (1995). *Probability and Statistics for Engineers and Scientists, fourth edition*, Alih bahasa: RK. Sembiring, ITB Bandung.