Analisis Regresi Logistik Eksak Menggunakan Metode *Penalized Maximum Likelihood Estimation*

[1]Disa Fauzia Nur Azizah, [2] Teti Sofia Yanti, [3]Nusar Hajarisman
[1,2]Prodi Statistika, FakultasMatematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Jl. Tamansari No.1 Bandung 40116
email: [1]dissydis25@gmail.com,[2]tetisofiayanti@gmail.com,[3]nusarhajarisman@yahoo.com

Abstrak. Di dalam statistika terdapat satu pendekatan model matematis yang digunakan untuk menganalisis hubungan beberapa faktor dengan sebuah variabel yang bersifat dikotomik (biner), pendekatan terebut merupakan regresi logistik biner. Terdapat beberapa macam regresi logistik biner yang salah satunya adalah Analisis Regresi Logistik Eksak. Analisis ini digunakan ketika metode asimptotik biasa tidak dapat digunakan karena sampel kecil, datanya jarang, juga terdapat banyaknya data kembar, terkadang suatu metode hanya valid pada situasi tertentu saja. Penaksiran parameter dilakukan menggunakan metode *Penalized Maximum Likelihood Estimation*, selanjutnya dilakukan pengujian parameter menggunakan pengujian eksak klaster satu tahap, dan yang terakhir model yang telah didapat diinterpretasikan menggunakan odd rasio. Data yang digunakan adalah data primer mengenai mahasiswa statistika Unisba angkatan 2013 yang masih aktif kuliah yang diuji antara lama belajar mandiri perminggu dengan ketepatan menyelesaikan studi sampai bulan Agustus 2017. Penerapan Regresi logistik eksak pada penelitian ini menunjukkan hasil bahwa semakin lama belajar, maka peluang untuk menyelesaikan studi sampai bulan Agustus 2017 semakin besar.

Kata Kunci: Regresi Logistik, Regresi Logistik Eksak, Metode *Penalized Maximum Likelihood Estimation*, Uji Eksak Klaster Satu Tahap, Odds Rasio

A. Pendahuluan

Dalam suatu penelitian di lapangan seringkali terdapat beberapa permasalahan yaitu data yang dikumpulkan terlalu sedikit yang terjadi akibat proses penelitian yang cukup panjang sehingga banyak kasus yang diteliti mengalami *drop out* atau karena populasinya yang kecil. Bagaimanapun metode asimptotik mungkin tidak sesuai ketika ukuran sampelnya kecil, atau datanya jarang, menceng, dan terdapat banyaknya data kembar, karena suatu metode terkadang hanya valid pada situasi tertentu saja (Deer, 2000).

Jika variabel respon (Y) memiliki skala biner dan terdapat satu atau lebih variabel prediktor yang berskala kontinu atau kategori, maka untuk mencari hubungan antar kedua variabel tersebut digunakan metode regresi logistik (Hosmer, Lemeshow, 1989). Metode tersebut biasanya digunakan ketika sampel berukuran besar. Ketika sampelnya kecil atau datanya jarang, akurasi perkiraan asimptotik tidak dapat diandalkan. Ketika ukuran sampel kecil, datanya jarang, atau menceng, metode eksak sangat diperlukan untuk mengambil kesimpulan. Dalam statistik nonparametrik pada tahun 1934 dikenal tes kemungkinan yang eksak dari Fisher (Siegel, 1988).

Metode yang banyak digunakan dalam menduga parameter pada regresi logistik yaitu Metode Kemungkinan Maksimum (*Maximum Likelihood Method).* Namun metode tersebut kurang bagus untuk digunakan, karena hasil estimasi akan menunjukkan model yang rumit dan tidak menggeneralisasi dengan baik. Apabila hasil estimasi parameter dari *maximum likelihood* kurang bagus, maka digunakan metode lain, yaitu metode *penalized maximum likelihood estimation* (PMLE).

Dari keadaan tersebut, penggunaan prosedur menggunakan eksak dengan memakai metode PMLE dapat menjadi alternatif untuk menyelesaikan permasalahan-permasalahan seperti ukuran sampel yang kecil.

B. Landasan Teori

Regresi Logistik

Regresi logistik adalah sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear atau yang biasa disebut dengan istilah *Ordinary Least Squares (OLS) regression*. Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel respon yang berskala dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah. Bentuk umum peluang regresi logistik adalah sebagai berikut:

$$\pi_i(x) = \frac{\exp\left(\beta_{0+} \sum_{p=1}^{q} \beta_p x_p\right)}{1 + \exp\left(\beta_{0+} \sum_{p=1}^{q} \beta_p x_p\right)}$$

$$p = 1, \dots, q \quad , \quad i = 1, \dots, n$$

dimana β_p menyatakan parameter-parameter regresi, x_p adalah pengamatan variabel prediktor ke-*p* dari sejumlah q variabel prediktor (Hosmer dan Lemeshow (1989) dalam Tiro (2000)).

Regresi Logistik Eksak

Regresi logistik eksak merupakan alat yang sangat berguna untuk model biner dengan ukuran sampel yang kecil dan menjadi solusi ketika metode asimptotik biasa tidak dapat diandalkan. Model eksak ini digunakan ketika ukuran sampel terlalu kecil untuk regresi logistik yang biasa, dan ketika terdapat sel yang bernilai nol. Dalam penelitian ini juga dilakukan pemeriksaan analisis data ukuran sampel kecil ketika data sampel dikumpulkan berdasarkan pengamatan yang berkorelasi. Misalkan data sampel didasarkan pada hasil klastering satu tahap, klastering dua tahap, atau klastering yang lebih tinggi. Model yang dihasilkan akan sama dengan model regresi logistik yang biasa, dimana peneliti hanya menggunakan satu variabel prediktor, sehingga model dapat ditulis sebagai:

$$Log \frac{\pi_i}{1 - \pi_i} = \beta_o + \beta x$$

Metode *Penalized Maximum Likelihood Estimation*

Parameter model regresi logistik ditaksir dengan metode *maximum likelihood estimation* (MLE) dan metode iteratif Newton-Raphson. Namun, ada saat di mana metode MLE tidak lagi dapat digunakan karena ukuran sampel teralu kecil atau apabila terdapat pemisahan pada data sehingga penaksir menjadi tidak konvergen. Untuk mengatasi hal ini, maka digunakan pendekatan metode PMLE yang pertama kali diusulkan oleh Firth (1993). PMLE dapat diberikan sebagai berikut:

$$L(\beta)^* = L(\beta)|I(\beta)|^{1/2}$$

Dimana $L(\beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i(x)^{y_i}[1 - \pi_i(x)]^{1-y_i}$ dan fungsi tereduksi $I(\beta)$

adalah *invariant jeffrey*:

$$I(\beta) = X^T W X$$

Dengan *corresponding log-likelihood:*

$$\ln L(\beta)^* = \ln L(\beta) + 0.5\ln|I(\beta)|$$

Prinsip dasar metode PMLE adalah memodifikasi fungsi skor likelihood menjadi fungsi skor penalized likelihood. PMLE akan diperoleh dengan cara membagi atau memilih banyaknya observasi *i* ke dalam dua observasi baru yang memiliki nilai respon y_i dan $1 - y_i$ dengan pembobot masing-masing $1 + \frac{h_i}{2}$ dan $\frac{h_i}{2}$ sehingga didapat persamaan skor:

$$U(\beta)^* \equiv \frac{\partial lnL(\beta)^*}{\partial(\beta)} = \sum_{i=1}^{n}(y_i - \pi_i)x_i\left(1 + \frac{h_i}{2}\right) + \sum_{i=1}^{n}(1 - y_i - \pi_i)x_i\left(\frac{h}{2}\right)$$
$$= 0$$
$$= \sum_{i=1}^{n}\left\{y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\}x_i = 0$$

Dimana h_i merupakan elemen diagonal ke-*i* dari matriks topi H, dengan $H = W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{1/2}$, dan $W = diag\{\pi_i(1 - \pi_i)\}$. Kini metode PMLE pada $\hat{\beta}$ dapat diperoleh dari proses iterasi hingga konvergensinya diperoleh sebagai berikut:

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + I^{-1}\left(\hat{\beta}^{(s)}\right)U(\beta^{(s)})^*$$

Ulangi proses tersebut sampai dengan konvergen, artinya nilai $\hat{\beta}^{(s+1)}$ sangat mendekati $\hat{\beta}^{(s)}$. Biasanya nilai taksiran dikatakan konvergen jika nilai selisih antara $\hat{\beta}^{(s+1)}$ dan $\hat{\beta}^{(s)}$ sebesar 10^{-6}, dimana s merupakan banyaknya iterasi.

Pengujian Eksak Klaster Satu Tahap

Teknik pengklasteran dapat ditempuh melalui dua cara yaitu dengan satu tahap atau dengan dua tahap. Jika semua kelompok yang ada dalam populasi diambil sebagai sampel, maka pengambilan klaster sampling hanya satu tahap. Tapi jika tidak semua kelompok pada populasi yang diambil, melainkan hanya beberapa kelompok saja, maka prosedurnya menggunakan klaster sampling dua tahap.

Dalam pemodelan klaster satu tahap, semua unit yang berada di dalam klaster terpilih dimasukan sebagai anggota. Misalkan y_i adalah jawaban responden, y=1 untuk sukses dan y=0 untuk gagal. Dan x_i adalah variabel prediktor. Pengujian eksak klaster satu tahap diuji dengan hipotesis:

$$H_0: \beta = 0 \qquad H_1: \beta > 0$$

Dibawah hipotesis H_0, distribusi bersyarat dari Z=z mereduksikan:

$$\Pr(Z_k = z_k | x_k^*; s_1, s_2) = \frac{\prod_{k=1}^{K}\binom{x_k^*}{z_k}}{\Sigma_{z^*\epsilon\Gamma(s_1,s_2)}\prod_{k=1}^{K}\binom{x_k^*}{z_k^*}}$$

karena rasio likelihood menurun terhapad t, p-value satu tahap untuk uji ini adalah

$$\Pr(t > t_{obs}|H_o, x, s) = \Sigma_{z^* \epsilon \Gamma(s_1,s_2):t(z^*)\geq t} \left[\frac{\prod_{k=1}^{K} \binom{x_k^*}{z_k}}{\Sigma_{z^* \epsilon \Gamma(s_1,s_2)} \prod_{k=1}^{K} \binom{x_k^*}{z_k^*}} \right]$$

Kriteria uji: Tolak H_o apabila p-value lebih kecil dari α.

Interpretasi Model

Model yang telah diuji dapat diinterpretasikan menggunakan odds rasio dengan perhitungan sebagai berikut:

Tabel Nilai Model Regresi Logistik

	X=1	X=0
Y=1 (Sukses)	$\pi(1) = \frac{exp^{\beta_0+\beta_1}}{1+exp^{\beta_0+\beta_1}}$	$\pi(0) = \frac{exp^{\beta_0}}{1+exp^{\beta_0}}$
Y=0 (Gagal)	$1-\pi(1) = \frac{1}{1+exp^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+exp^{\beta_0}}$

Nilai odds untuk Y=1 terhadap X=1 dinyatakan dengan $\pi(1)/[1-\pi(1)]$, sedangkan nilai odds untuk Y=0 terhadap X=0 adalah $\pi(0)/[1-\pi(0)]$. Sehingga:

$$OR = \frac{\pi(1)/[1-\pi(1)],}{\pi(0)/[1-\pi(0)]}$$

$$= \frac{\pi(1)[1-\pi(0)]}{\pi(0)[1-\pi(1)]}$$

$$= \frac{\left(\frac{exp^{\beta_0+\beta_1}}{1+exp^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+exp^{\beta_0}}\right)}{\left(\frac{exp^{\beta_0}}{1+exp^{\beta_0}}\right)\left(\frac{1}{1+exp^{\beta_0+\beta_1}}\right)}$$

$$= \frac{exp^{\beta_0+\beta_1}}{exp^{\beta_0}} = exp^{\beta_1}$$

Nilai odds ratio akan berada dalam batasan 0 sampai dengan tak hingga. Pada saat $OR < 1$ terdapat asosiasi yang rendah antara variabel baris dengan variabel kolom. Sedangkan pada saat $OR = 1$, maka tidak ada asosiasi antara variabel baris dengan variabel kolom. Dan pada saat $OR > 1$ artinya terdapat asosiasi yang tinggi antara variabel baris dengan variabel kolom.

C. Hasil Penelitian dan Pembahasan

Deskripsi Data

Data yang digunakan adalah data primer mengenai mahasiswa Statistika Unisba aktif angkatan 2013. Variabel prediktor dalam penelitian ini adalah lama belajar mandiri dalam jam perminggu. Dimana terdapat 37 mahasiswa, 23 orang menyelesaikan studi sampai bulan Agustus 2017 dan 14 orang lainnya tidak menyelesaikan studi sampai bulan Agustus 2017. Dengan lama belajar mandiri perhari adalah 1 sampai 8 jam.

Model Regresi Logistik Eksak

Data pengamatan yang akan diteliti yaitu lama belajar mandiri dalam satu minggu (x). Penaksiran parameter yang digunakan menggunakan metode maksimum *likelihood* dan metode iterasi Newton-Raphson. Penaksiran parameter menggunakan metode *penalized maximum likelihood estimation* dan metode iterasi Fisher's Scoring.

Berdasarkan hasil perhitungan Regresi Logistik Eksak dengan bantuan *software* SAS dapat diperoleh nilai-nilai taksiran parameter untuk Model Regresi Logistik Eksak. Untuk model dengan nilai taksiran parameter β_0 adalah -1.8183, dan nilai taksiran parameter β_1 adalah 0.5934.

Model logit eksak menggunakan *penalized maximum likelihood estimation* dapat ditulis sebagai berikut:

$$g(x) = ln\left(\frac{\pi_1(x)}{1-\pi_1(x)}\right) = -1.8183 + 0.5934x$$

Pengujian Parameter Model Regresi Logistik Eksak

Selanjutnya akan dilakukan pengujian parameter secara eksak dengan menggunakan pengujian eksak klaster satu tahap, dengan hipotesis sebagai berikut:

$$H_0: \beta = 0, \qquad H_1: \beta > 0$$

Pertama hitung $\sum z_k$ atau jumlah mahasiswa yang menyelesaikan studi sampai bulan Agustus 2017 untuk menentukan s_1. Setelah itu hitung s_2 dengan rumus $\sum z_k\,(n_k - z_k)$ dimana n_k adalah banyaknya responden dalam setiap klaster. Lalu setelah itu cari nilai t awal dengan cara menjumlahkan z_k dikali dengan x_k^*. Dimana nilai s_1 dan s_2 merupakan nilai tetap. Jika nilai s_1 dan s_2 telah ditetapkan maka z_k^* dapat dicari dan dihitung. Hasil perhitungan dengan menggunakan persamaan:

$$\Pr(Z_k = z_k | x_k^*; s_1, s_2) = \frac{\prod_{k=1}^{K} \binom{x_k^*}{z_k}}{\Sigma_{z^* \epsilon \Gamma(s_1,s_2)} \prod_{k=1}^{K} \binom{x_k^*}{z_k^*}}$$

Berdasarkan output *software* SAS didapat hasil peluang sebesar 0.000035. Setelah itu lakukan perhitungan menggunakan persamaan (2.15) untuk mencari nilai p-value. Dimana nilai z_k^* yang dimasukkan ke dalam rumus bersyarat nilai t harus lebih besar dari t awal dan nilai s_1 dan s_2 sesuai dengan yang telah ditetapkan. Hasil perhitungan p-value untuk uji eksak klaster satu tahap dengan menggunakan persamaan:

$$\Pr(t > t_{obs}|H_o, x, s) = \Sigma_{z^* \epsilon \Gamma(s_1,s_2):t(z^*)\geq t} \left[\frac{\prod_{k=1}^{K} \binom{x_k^*}{z_k}}{\Sigma_{z^* \epsilon \Gamma(s_1,s_2)} \prod_{k=1}^{K} \binom{x_k^*}{z_k^*}} \right]$$

Berdasarkan output *software* SAS didapatkan hasil p-value = 0.0001. Kriteria pengujiannya adalah tolak H_0 jika nilai p-value kurang dari α, dimana nilai α = 0,05. Karena nilai p-value = 0.0001 lebih kecil dari α = 0,05 maka H_0 ditolak artinya parameter regresi logistik eksak berpengaruh secara signifikan dalam model.

Interpretasi Model Regresi Logistik Eksak

Model Regresi Logistik Eksak dapat diinterpretasikan menggunakan odds rasio, hasil perhitungan dari odds rasio dengan hipotesis sebagai berikut

$$H_0: OR = 1, \qquad H_1: OR \neq 1$$

Dengan menggunakan persamaan dan hasil dari *software* SAS didapat nilai odds rasio adalah $e^{\beta_1} = e^{0.5934} = 1.810$ dan p-value = 0.0002. Kriteria pengujiannya adalah tolak H_0 jika p-value lebih besar dari α, dimana nilai $\alpha = 0.05$. Karena p-value = 0.0002 lebih kecil dari $\alpha = 0.05$ maka H_0 ditolak yang artinya odd rasio tidak mengandung nilai 1 artinya terdapat perbedaan lamanya belajar dengan ketepatan mahasiswa menyelesaikan studi. Misalkan seseorang yang belajar 0 jam akan berbeda dengan yang belajar 1 jam, 2 jam, dan lainnya. Dengan nilai odd rasio sebesar 1.810 dapat disimpulkan bahwa perbandingan peluang seseorang untuk menyelesaikan studi sampai bulan Agustus 2017 dengan belajar mandiri lebih lama 1 jam dari yang lainnya dalam seminggu adalah 1.810 kali atau 2 kali lebih banyak dari yang lainnya.

Selanjutnya akan dilakukan perhitungan nilai peluang untuk masing-masing nilai x. Berikut merupakan perhitungan peluang untuk lamanya belajar mandiri dalam menyelesaikan studi sampai bulan Agustus 2017 dengan menggunakan persamaan:

$$\pi_i(x) = \frac{\exp\left(\beta_{0+} \sum_{p=1}^{q} \beta_p x_p\right)}{1 + \exp\left(\beta_{0+} \sum_{p=1}^{q} \beta_p x_p\right)}$$

Tabel Nilai Peluang Lama Belajar Mandiri dalam Menyelesaikan Studi sampai Bulan Agustus 2017 Menggunakan Model Regresi Logistik Eksak

Lama Belajar (jam) Perminggu (X)	Perhitungan	Peluang
0	$\pi(0) = \frac{\exp(-1.8183 + 0.5934(0))}{1 + \exp(-1.8183 + 0.5934(0))}$	0.1396
1	$\pi(1) = \frac{\exp(-1.8183 + 0.5934(1))}{1 + \exp(-1.8183 + 0.5934(1))}$	0.2271

Lanjutan Tabel

Lama Belajar (jam) Perminggu (X)	Perhitungan	Peluang
2	$\pi(2) = \frac{\exp(-1.8183 + 0.5934(2))}{1 + \exp(-1.8183 + 0.5934(2))}$	0.3472
3	$\pi(3) = \frac{\exp(-1.8183 + 0.5934(3))}{1 + \exp(-1.8183 + 0.5934(3))}$	0.4905
4	$\pi(4) = \frac{\exp(-1.8183 + 0.5934(4))}{1 + \exp(-1.8183 + 0.5934(4))}$	0.6354
5	$\pi(5) = \frac{\exp(-1.8183 + 0.5934(5))}{1 + \exp(-1.8183 + 0.5934(5))}$	0.7593
6	$\pi(6) = \frac{\exp(-1.8183 + 0.5934(6))}{1 + \exp(-1.8183 + 0.5934(6))}$	0.8510
7	$\pi(7) = \frac{\exp(-1.8183 + 0.5934(7))}{1 + \exp(-1.8183 + 0.5934(7))}$	0.9118
8	$\pi(8) = \frac{\exp(-1.8183 + 0.5934(8))}{1 + \exp(-1.8183 + 0.5934(8))}$	0.9493

Dari hasil diatas dapat disimpulkan bahwa seseorang yang belajarnya semakin lama memiliki peluang semakin besar untuk menyelesaikan studi pada bulan Agustus 2017.

D. Kesimpulan

Berdasarkan pembahasan pada bab sebelumnya, didapat kesimpulan bahwa untuk mengatasi permasalahan data dengan sampel kecil yaitu dengan menggunakan analisis regresi logistik eksak. Dari perhitungan odds rasio diketahui terdapat perbedaan lamanya belajar dengan ketepatan mahasiswa dalam menyelesaikan studi, dan dihitung peluang untuk masing-masing lamanya belajar, sehingga dapat disimpulkan bahwa semakin lama belajar dalam satu minggu, maka peluang untuk menyelesaikan studi tepat waktu atau sampai bulan Agustus 2017 akan semakin besar.

E. Saran

Ketika akan menggunakan analisis regresi logistik eksak, pastikan data tersebut merupakan data dengan sampel kecil dengan menghitung jumlah sampel atau menghitung nilai ekspektasi dari masing-masing sel. Bagi yang akan menganalisis metode yang sama sebaiknya menambahkan variabel prediktor lebih dari satu. Model Regresi Logistik Eksak selain dengan menggunakan SAS dapat juga menggunakan aplikasi *software* yang lain seperti C+ dan R.

Daftar Pustaka

Agresti, A. (1990). *Categorical Data Analysis,* Canada: John Wiley & Sons, Inc.

Agresti, A. (2002). Categorical Data Analysis. New York: Inc. John Wiley and Sons.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*. New York: John Wiley and Sons.

Allison, P. D. (1999). Logistic Regression Using SAS System: Theory and Application. Cary, NC: SAS Institute Inc.

Christopher Zorn; 2005; A Solution to Separation in Binary Response Models.

Cox, D. R. and Snell, E. J. (1989). Analysis of Binary Data, Second Edition. Boca Raton: Chapman and Hall.

Deer, Robert E, (2000) Performing Exact Logistic Regression With the SA System.

Firth, D., (1993), Bias Reduction of Maximum Likelihood Estimates, *Biometrika*, 80, hal. 27-38

Hajarisman, Nusar. (2009). *Analisis Data Kategorik*. Bandung: Prodi Statistika Universitas Islam Bandung.

Hosmer, D.W. dan Lemeshow. (1989). Applied Logistic Regression. New York: John Wiley and Sons.

Hosmer, D.W. dan Lemeshow, S. (2000). *Applied Logistic Regresion*. Second Edition. N ew York: John Wiley and Sons, Inc.

Marsisno, Waris (1999). Penerapan Anaalisis Regresi Logistik Dalam Analisis Risiko Perioperative Stroke Pada Operasi Jantung Forum Statistik Tahun XVII no 3 : 155-165.

McCullagh, P. (1980). Regression Models for Ordinal Data. Journal of the Royal Statistical Society. Series B (Methodological), 42, 109-142.

Jeffreys, H. (1946). "An Invariant Form of the Prior Probability in Estimation Problems." Proceedings of the Royal Society of London, Series A 186:453–461.

Seigel, S. dan Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Science*, 2nd edn, New York.