Regresi Logistik pada Data *Rare Event*

[1]Rundy Rumi Ari Wistara, [2]Suliadi, [3]Abdul Kudus
[1,2,3] *Statistika, Fakultas MIPA, Universitas Islam Bandung,*
Jl. Rangga Malela No. 1 Bandung 40116
e-mail: [1] rundy_ra@ymail.com, [2] suliadi@gmail.com, [3] akudus69@ahoo.com

Abstrak. Regresi logistik merupakan salah satu metode statistika yang digunakan untuk menganalisis hubungan beberapa faktor dengan sebuah variabel respon. Pada regresi logistik variabel respon terdiri dari dua kategori yaitu "sukses" dan "gagal" yang dinotasikan dengan y = 1 (sukses) dan y = 0 (gagal). Regresi logistik baik digunakan jika persentase y = 0 dan y = 1 tidak jauh berbeda. Dalam kasus kredit macet dapat dijumpai kondisi di mana persentase kredit macet jauh lebih kecil dibandingkan dengan persentase kredit lancar pada variabel respon. Kondisi seperti itu disebut dengan *rare event.* Pada data *rare event* akan menyebabkan Pr(Y = 1) *underestimates* sedangkan untuk Pr(Y = 0) *overestimates.* Masalah sampel terbatas (*finite sample*) dapat menyebabkan (i) model yang terbentuk akan menghasilkan penaksir parameter yang bias; (ii) kesalahan baku bagi penaksir yang lebih kecil (*underestimates*) dan (iii) dapat menyebabkan P(Y = 1) *underestimates*. Skripsi ini membahas bagaimana mengoreksi penduga parameter yang bias dan koreksi peluang pada regresi logistik jika data respon jarang terjadi (*rare event*). Model regresi logistik *rare event* akan diterapkan pada data kasus kredit bank di Amerika dengan Y = 1 jika nasabah mengalami kredit macet lebih dari 90 hari. Hasil koreksi bias terhadap koefisien regresi adalah bahwa bias pada $\hat{\beta}_0$ lebih besar di bandingkan dengan yang lainnya. Namun untuk hasil bias pada keseluruhan taksiran parameter kecil, hal ini karena sampel yang cukup besar yaitu sebanyak 12013. Hasil kesalahan baku penaksir terlihat bahwa kesalahan baku penaksir parameter terkoreksi lebih kecil dibandingkan dengan kesalahan baku pada penaksir parameter regresi logistik. Oleh karena itu regresi logistik pada data *rare event* lebih baik digunakan daripada regresi logistik. Sedangkan untuk koreksi taksiran peluang ($\tilde{\pi}$) lebih kecil dibandingkan dengan taksiran peluang jika tidak menggunakan koreksi pada regresi logistik.

Kata kunci: Regresi Logistik, *Rare Event*, Bias, Peluang.

A. Pendahuluan

Dalam kehidupan sehari-hari semua orang pasti memiliki kebutuhan. Kebutuhan ada yang bersifat mendesak dan ada yang tidak. Kebutuhan yang mendesak menuntut untuk segera dipenuhi. Namun pemenuhan tersebut tidak terlepas dari masalah biaya atau dana. Dana yang diperlukan biasanya tidak sedikit jumlahnya, sementara dana yang tersedia acapkali tidak mencukupi.

Kebanyakan orang dalam menghadapi kekurangan dana salah satu jalan keluar yang dapat dilakukan adalah dengan berutang kepada pihak bank. Para nasabah yang telah memperoleh fasilitas kredit dari bank tidak seluruhnya dapat mengembalikan utangnya dengan lancar sesuai dengan waktu yang telah diperjanjikan. Akibat nasabah tidak dapat membayar lunas utangnya, maka akan tergambar perjalanan kredit menjadi macet atau terhenti.

Salah satu metode yang dapat dipergunakan untuk memetakan nasabah ke dalam kategori kredit macet dan lancar yaitu metode regresi logistik. Regresi logistik merupakan salah satu metode statistika yang digunakan untuk menganalisis hubungan antara satu variabel respon (Y) dengan satu atau lebih variabel bebas (X_i). Dimana variabel respon terdiri dari dua kategori yaitu "sukses" dan "gagal" yang dinotasikan dengan Y = 1 (sukses) dan Y = 0 (gagal). Sebagai contoh pada kasus kartu kredit, Y = 0

jika variabel responnya menyatakan kredit lancar dan Y = 1 jika variabel responnya menyatakan kredit macet.

Regresi logistik baik digunakan jika persentase Y = 0 dan Y = 1 tidak jauh berbeda. Dalam kasus kredit macet dapat dijumpai kondisi di mana persentase kredit macet jauh lebih kecil dibandingkan dengan persentase kredit lancar pada variabel respon. Kondisi seperti itu disebut dengan *rare event.* Pada data *rare event* akan menyebabkan Pr(Y = 1) *underestimates* sedangkan untuk Pr(Y = 0) *overestimates.* Masalah sampel terbatas (*finite sample*) dapat menyebabkan (i) model yang terbentuk akan menghasilkan penaksir parameter yang bias; (ii) kesalahan baku bagi penaksir yang lebih kecil (*underestimates*) dan (iii) dapat menyebabkan P(Y = 1) *underestimates*. Skripsi ini membahas bagaimana mengoreksi penduga parameter yang bias dan koreksi peluang pada regresi logistik jika data respon *rare event* dan diaplikasikan pada kasus kredit macet.

1. Rumusan Masalah

Berdasarkan uraian dari latar belakang yang telah diungkapkan, maka masalah yang dapat diidentifikasi adalah:

1. Bagaimana perbandingan penaksir parameter regresi logistik dengan regresi logistik pada data *rare event* ?
2. Bagaimana perbandingan kesalahan baku penaksir parameter regresi logistik dengan regresi logistik pada data *rare event* ?
3. Bagaimana perbandingan taksiran peluang regresi logistik dengan regresi logistik pada data *rare event* ?

2. Tujuan Penelitian

Berdasarkan identifikasi masalah maka tujuan dalam penulisan skripsi ini adalah:

1. Membandingkan penaksir parameter regresi logistik dengan regresi logistik pada data *rare event*.
2. Membandingkan kesalahan baku penaksir parameter regresi logistik dengan regresi logistik pada data *rare event*.
3. Membandingkan taksiran peluang regresi logistik dengan regresi logistik pada data *rare event*.

B. Tinjauan Pustaka

1. Regresi Logistik

Menurut Hosmer dan Lemeshow (1989) model regresi logistik yang dipengaruhi oleh *k* variabel bebas dapat dinyatakan sebagai nilai harapan dari *Y* dengan diberikan nilai *x*.

$$E(Y \mid x) = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)} \tag{2.1}$$

Untuk mempermudah dalam menaksir parameter regresi, maka π_i pada persamaan (2.1) ditransformasikan dengan menggunakan transformasi logit. Sehinga dapat ditulis sebagai berikut:

$$logit\ (\pi_i) = g(x) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.2}$$

2. Penaksiran Parameter Model

Metode penaksiran parameter yang biasa digunakan dalam regresi logistik adalah metode MLE (*Maximum Likelihood Estimation*). Variabel respon Y memiliki sebaran Bernoulli dengan parameter π_i dan fungsi sebaran peluangnya adalah:

$$P(y_i \mid x_i) = \begin{cases} \pi_i^{\,y_i}\left[1-\pi_i\right]^{1-y_i} & ,untuk\, y_i = 0\, atau\, 1 \\ 0 & ,untuk\, y_i\ yang\, lain \end{cases}$$

Menurut Hosmer dan Lemeshow (1989), fungsi *likelihood* distribusi Bernoulli untuk *n* sampel bebas adalah

$$l(\beta) = \prod_{i=1}^{n} \pi_i^{\,y_i}\left(1-\pi_i\right)^{1-y_i} \qquad (2.3)$$

Untuk memudahkan mencari nilai $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ yang memaksimumkan fungsi *likelihood* digunakan bentuk logaritma natural dari fungsi *likelihood*, yang disebut sebagai fungsi *log-likelihood*. Logaritma natural fungsi peluang bersamanya dapat ditulis sebagai berikut:

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^{n} \left\{ y_i \left(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p\right) + \ln\left[1-\pi_i\right] \right\} \qquad (2.4)$$

Selanjutnya dihitung turunan pertama dari $L(\beta)$ masing-masing terhadap $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ kemudian disyaratkan sama dengan nol.

$$\frac{dL(\beta)}{d\beta_0} = \frac{d}{d\beta_0} \sum_{i=1}^{n} \left\{ y_i \left(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p\right) + \ln\left[1-\pi_i(\beta)\right] \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i - \pi_i \right\} \qquad (2.5)$$

$$\frac{dL(\beta)}{d\beta_1} = \frac{d}{d\beta_1} \sum_{i=1}^{n} \left\{ y_i \left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right) + \ln\left[1-\pi_i(\beta)\right] \right\}$$

$$= \sum_{i=1}^{n} x_{1i}\left[y_i - \pi_i \right] = 0 \qquad (2.6)$$

.
.
.

$$\frac{dL(\beta)}{d\beta_p} = \frac{d}{d\beta_p} \sum_{i=1}^{n} \left\{ y_i \left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right) + \ln\left[1-\pi_i(\beta)\right] \right\}$$

$$= \sum_{i=1}^{n} x_{ki}\left[y_i - \pi_i \right] = 0 \qquad (2.7)$$

Dari persamaan (2.7), (2.8), dan (2.9) masih terkandung y_i, dari turunan pertama di atas sulit untuk dihitung secara manual oleh sebab itu digunakan bantuan *software*.

Selanjutnya akan dihitung turunan kedua, turunan kedua ini akan dilihat apakah ada solusi atau tidak.

Bentuk umum dari turunan parsial kedua fungsi *log-likelihood* adalah:

$$\frac{\delta^2 L(\beta)}{\delta \beta_k^2} = -\sum_{i=1}^{n} x_{ki}{}^2 \pi_i (1-\pi_i) < 0$$

$$\frac{\delta^2 L(\beta)}{\delta \beta_r \delta \beta_k} = -\sum_{i=1}^{n} x_{ki.ri} \pi_i (1-\pi_i) < 0$$

dimana *i, j = 0, 1, 2, ..., k*. Dan penaksir matriks variansnya adalah

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \pi_i (1-\pi_i) x_i' x_i \right]^{-1} \quad (2.8)$$

dimana π_i adalah peluang sukses, 1-π_i adalah peluang gagal dan x_i adalah variabel bebas dengan *i* = 1, 2, ..., *k*.

3. Regresi Logistik pada Data *Rare Event*

Misalkan variabel respon $Y_1, Y_2, ...Y_i, ...Y_n$ merupakan sampel acak yang berdistribusi Bernoulli dengan $\pi_i = P(Y = 1)$ dan $1 - \pi_i = P(Y = 0)$ untuk i = 1, 2, ..., *n*. Dalam model regresi logistik peluang π_i adalah fungsi distribusi kumulatif logistik padapersamaan 2.1. Transformasi logit sebagaimana dijelaskan pada bagian regresi logistik yaitu persamaan 2.2.

1) Koreksi Bias terhadap Koefisien

Untuk mengoreksi bias $\hat{\beta}$ dapat ditaksir oleh *weighted least-squared*:

$$bias(\hat{\beta}) = (X'WX)^{-1} X'W\xi \quad (2.9)$$

dimana

$$\xi_i = 0.5 Q_{ii} ((1+\overline{w}_1)\hat{\pi}_i - \overline{w}_1) \quad (2.10)$$

$$Q_{ii} = elemen\ diagonal\ utama \left\{ X \left(X'WX \right)^{-1} X' \right\} \quad (2.11)$$

$$W = diag\{\hat{\pi}_i (1-\hat{\pi}_i) w_i\} \quad (2.12)$$

Dengan vektor pembobot w_i sebagai berikut,

$$w_i = \overline{w}_1 Y_i + \overline{w}_0 (1 - Y_i) \quad (2.13)$$

dimana $\overline{w}_1 = \frac{\tau}{\bar{y}}$ sebagai pembobot untuk nilai satu dan $\overline{w}_0 = \frac{1-\tau}{1-\bar{y}}$ sebagai pembobot nilai nol. Sedangkan τ adalah proporsi kejadian sukses dalan populasi dan $\bar{y}$ adalah proporsi kejadian sukses dalan sampel. Metode WLS pada regresi logistik rare event mudah untuk diterapkan karena komponennya sama dengan metode WLS pada regresi logistik. Dengan ξ sebagai variabel respon, X sebagai variabel bebas dan W sebagai pembobot. Sedangkan untuk penaksir koreksi biasnya yaitu,

$$\tilde{\beta} = \hat{\beta} - bias(\hat{\beta}) \quad (2.14)$$

Untuk mendapatkan prediksi peluang maka bisa dilakukan dengan memasukkan koreksi penaksir bias ($\tilde{\beta}$) ke dalam persamaan logit sebagai berikut:

$$\tilde{\pi} = \Pr(\hat{y}_i = 1 | \tilde{\beta}) = \frac{\exp(x_i \tilde{\beta})}{1 + \exp(x_i \tilde{\beta})} \quad (2.15)$$

Namun, hal ini tidak optimal karena mengabaikan ketidakpastian pada $\tilde{\beta}$. Oleh karena itu perlu dilakukan koreksi ulang terhadap $\tilde{\pi}_i$. Bentuk koreksi peluangnya sebagai berikut:

$$\Pr(y_i = 1) \approx \tilde{\pi}_i + C_i \qquad (2.16)$$

dengan faktor koreksinya adalah

$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)x_i V(\tilde{\beta})x_i' \qquad (2.17)$$

dimana

$$V(\tilde{\beta}) = \left(\frac{n}{(n+k)}\right)^2 V(\hat{\beta})$$

dan matriks varians $V(\hat{\beta})$ sebagaimana dijelaskan pada bagian penaksiran parameter model regresi logistik.

C. Bahan dan Metode Penelitian

1. Bahan

Data yang digunakan untuk mengaplikasikan analisis regresi logistik *rare event* ini berasal dari "*Give me some kredit*" yang diluncurkan dalam situs Kaggle. Data tersebut berisi tentang nasabah yang memiliki fasilitas kredit (debitur). Variabel respon yang digunakan adalah status kredit yaitu:

$$y_i = \begin{cases} 0\text{ , tidak mengalami kredit macet lebih dari 90 hari} \\ 1\text{ , mengalami kredit macet lebih dari 90 hari} \end{cases}$$

Sedangkan variabel bebas ada sebanyak 10 variabel yang terdiri dari Jumlah Saldo Kartu Kredit, Usia Debitur, Frekuensi Mengalami Kredir Macet 30-59 Hari, Pembayaran Utang Bulanan, Pendapatan Bulanan, Jumlah Pinjaman Terbuka dan Kredit, Frekuensi Mengalami Kredir Macet >90 Hari, Jumlah Kredit KPR dan Properti, Frekuensi Mengalami Kredir Macet 60-89 Hari, Jumlah Tanggungan Keluarga.

2. Metode

Metode dan tahap-tahap penelitian yang dilakukan untuk mencapai tujuan penulisan adalah sebagai berikut:

1. Melakukan penaksiran koefisien parameter regresi logistik.
2. Melakukan penaksiran model regresi logistik pada data *rare event* dengan langkah-langkah sebagai berikut:
   a) Menghitung nilai proporsi kejadian sukses dalam populasi (τ).
   b) Menghitung nilai proporsi kejadian sukses dalam sampel ($\bar{y}$).
   c) Masukan langkah 1 dan 2 ke dalam persamaan (2.15) lalu hitung vektor pembobot (w_i).
   d) Menentukan vektor pembobot W pada persamaan (2.14).
   e) Kemudian tentukan nilai Q_{ii} pada persamaan (2.13).
   f) Lalu tentukan vektor dari ξ_i pada persamaan (2.12).
   g) Menghitung $bias(\hat{\beta})$ dengan menggunakan persamaan (2.11).
   h) Menghitung penaksir terkoreksi ($\tilde{\beta}$) pada persamaan (2.16).

   Langkah-langkah di atas dilakukan dengan menggunakan perintah 'relogit' pada *package* 'Zelig' *software* R.

3. Mengoreksi kesalahan baku bagi penaksir pada regresi logistik pada persamaan (2.20).
4. Koreksi terhadap P(Y = 1) model regresi logistik pada persamaan (2.18) dengan langkah-langkah sebagai berikut:
   a) Menghitung prediksi peluang dengan memasukkan penaksir terkoreksi ke dalam persamaan (2.17).
   b) Menghitung faktor koreksi (C_i) pada persamaan (2.19)

   Langkah 4.a dan 4.b dilakukan dengan menggunakan *software* SAS IML

D. Pembahasan

Penaksiran parameter untuk model regresi logistik dilakukan dengan menggunakan metode *maximum likelihood*. Variabel respon yang digunakan adalah status kredit yaitu:

$$y_i = \begin{cases} 0\text{ , tidak mengalami kredit macet lebih dari 90 hari} \\ 1\text{ , mengalami kredit macet lebih dari 90 hari} \end{cases}$$

dimana banyaknya nilai Y = 1 yaitu sebesar 827 dan banyaknya nilai Y = 0 sebesar 11186. Namun demikian data yang mengandung fenomena *rare event* dimana persentase Y = 1 hanya sebesar $\frac{827}{12013} = 6{,}88\%$. Hal tersebut akan mengakibatkan *underestimated* pada P(Y = 1) yang artinya terdapat bias pada penaksir parameternya. Dengan demikian harus dilakukan koreksi terhadap koefisien parameter regresi logistik. Bias pada penaksir parameter kecil, hal ini dikarenakan sampel yang cukup besar yaitu sebanyak 12013. Sedangkan bias pada $\hat{\beta}_0$ lebih besar di bandingkan dengan yang lainnya.

Kesalahan baku pada penaksir parameter sangat penting digunakan dalam suatu analisis salah satunya pengujian hipotesis. Kesalahan baku bagi penaksir parameter terkoreksi lebih kecil dibandingkan dengan kesalahan baku bagi penaksir parameter regresi logistik. Ketika data *rare event* koreksi terhadap kesalahan baku lebih baik digunakan daripada regresi logistik.

Koreksi peluang pada data *rare event* dilakukan dengan menggunakan penaksir terkoreksi untuk mendapatkan prediksi peluang $\tilde{\pi}_i$. Koreksi peluang ($\tilde{\pi}$) lebih kecil dibandingkan dengan taksiran peluang jika tidak menggunakan koreksi pada regresi logistik.

E. Kesimpulan

Kesimpulan dari skripsi ini adalah:

1. Hasil koreksi bias terhadap koefisien regresi adalah bahwa bias pada $\hat{\beta}_0$ lebih besar di bandingkan dengan yang lainnya. Namun untuk hasil bias pada keseluruhan taksiran parameter kecil, hal ini karena sampel yang cukup besar yaitu sebanyak 12013.
2. Hasil kesalahan baku penaksir terlihat bahwa kesalahan baku penaksir parameter terkoreksi lebih kecil dibandingkan dengan kesalahan baku pada penaksir parameter regresi logistik. Oleh karena itu regresi logistik pada data *rare event* lebih baik digunakan daripada regresi logistik.

3. Sedangkan untuk koreksi taksiran peluang ($\tilde{\pi}$) lebih kecil dibandingkan dengan taksiran peluang jika tidak menggunakan koreksi pada regresi logistik.

DAFTAR PUSTAKA

Agresti, Alan. 2002. *Categorical Data Analysis*. New York: Inc. John Wiley and Sons.

Bahsan, M. 2012. *Hukum Jaminan dan Jaminan Kredit Perbankan Indonesia*. Jakarta: PT. Raja Grafindo Persada.

D.W. Hosmer, dan S. Lemeshow. 1989. *Applied Logistik Regression*. New York: Inc. John Wiley and Sons.

Hajarisman. N. 2009. *Analisis Data Kategorik*. Bandung: Tidak diterbitkan.

King Gary, dan Zeng Langche. 2001. *Logistic Regression in Rare Events Data*. http://gking.harvard.edu/files/0s.pdf.

Nasution, Aatiqah. 2012. *Regresi Logistik untuk Menentukan Peluang yang Mempunyai Kartu Yogya dan yang Tidak Mempunyai Kartu Yogya Berdasarkan Kepuasan Konsumen*. Bandung: tidak diterbitkan.

Supramono, Gatot. 2009. *Perbankan dan Masalah Kredit – Suatu Tinjauan di Bidang Yuridis*. Jakarta: PT. Rineka Cipta.