J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 11, No. 2, 122-128, Januari 2015
ISBN:1858-1382

Estimasi Parameter Model Regresi Logistik Biner Komponen Utama Non Linear dengan Maksimum Likelihood

Anna Islamiyati

Abstrak

Regresi logistik biner digunakan pada data respon yang mengandung dua kategorik, dan ketika terjadi multikolinearitas pada variabel prediktor yang berskala campuran, maka pendekatan statistika yang dapat digunakan untuk mengatasi masalah tersebut adalah analisis komponen utama non linear. Estimasi parameter regresi logistik komponen utama non linear dilakukan melalui metode maksimum likelihood dan diperoleh hasil taksiran parameter yang implisit sehingga dilanjutkan dengan metode iterasi Newton Raphson untuk mendapatkan hasil estimasi yang konvergen.

Kata kunci: analisis komponen utama non linear, maksimum likelihood, multikolinearitas, regresi logistik biner.

1. Pendahuluan

Penggunaan regresi selalu dibatasi dengan beberapa asumsi, salah satunya adalah tidak terjadi multikolinearitas, yaitu korelasi kuat antara variabel prediktor. Asumsi ini tidak hanya berlaku pada regresi linear yang kuantitatif tetapi juga pada regresi logistik. Regresi logistik digunakan pada saat data variabel respon berbentuk kategorik, dan jika datanya hanya terdiri dari dua kategori maka regresi logistik yang digunakan disebut regresi logistik biner. Penggunaan regresi logistik biner sudah banyak digunakan dalam beberapa bidang aplikasi, misalnya di bidang kesehatan, industri, kependudukan, dan lain-lainnya.

Ketika terjadi pelanggaran asumsi multikolinearitas, salah satu pendekatan yang dapat digunakan adalah analisis komponen utama, yang akan mereduksi variabel prediktor menjadi beberapa komponen utama. Pendekatan ini pun juga sudah banyak dikembangkan dan digunakan dalam beberapa bidang aplikasi. Namun, yang banyak digunakan adalah penggunaan pada data yang berbeda untuk model regresi logistik dan analisis komponen utama, atau dalam penelitian yang berbeda. Ternyata dalam perkembangan data pada saat sekarang ini, banyak data yang variabel responnya kategorik, memiliki variabel prediktor yang multikolinearitas. Sehingga perlu pendekatan untuk mengatasi masalah tersebut untuk mendapatkan taksiran model yang baik.[1]

Makalah ini akan mengkaji tentang penggunaan analisis komponen utama non linear pada data respon kategorik dimana variabel prediktornya mengalami korelasi yang kuat. Analisis komponen utama non linear merupakan pengembangan dari analisis komponen utama, yang digunakan pada saat variabel prediktor yang digunakan dalam jumlah banyak dan jenis datanya yang bervariasi (Kroonenberg, 1997). Penggunaan analisis komponen utama non linear telah digunakan pada penelitian tentang prestasi belajar mahasiswa di Unhas (Islamiyati dan Talangko, 2010). Penggunaan regresi logistik komponen utama sudah dikaji oleh Aguilera (2000), namun dalam makalah ini akan dikaji mengkhusus mengenai estimasi parameter dalam regresi logistik

[1,2,3] *Jurusan Matematika FMIPA Universitas Hasanuddin Makassar, Jl. Perintis Kemerdekaan Km.10 Makassar*

biner yang menggunakan pendekatan analisis komponen utama non linear melalui metode maksimum likelihood.

2. Analisis Komponen Utama Non Linear

Analisis komponen utama merupakan salah satu metode perampatan dalam regresi untuk memperoleh hasil yang lebih baik jika terjadi pelanggaran asumsi, khususnya dalam masalah kolinearitas. Peubah bebas pada regresi komponen utama berupa hasil kombinasi linear dari peubah asal*Z*, yang disebut komponen utama*W*. Koefisien penduga dari metode ini diperoleh melalui penyusutan dimensi komponen utama, dimana subset komponen utama yang dipilih harus tetap mempertahankan keragaman yang sebesar-besarnya (Jolliffe, 1989). Analisis komponen utama menggunakan struktur matriks variansi-kovariansi dari suatu himpunan variabel melalui beberapa kombinasi linier dari himpunan variabel tersebut.

Analisis komponen utama non linear merupakan pengembangan dari analisis komponen utama yang menggunakan pendekatan *alternating least squares* (Gifi, 1990). Analisis komponen utama non linear merupakan suatu metode yang mampu menganalisis data berskala campuran (nominal, ordinal, rasio dan interval) secara simultan yaitu dengan cara mengelompokkan variabel-variabel yang korelasi linearnya sejalan menjadi satu komponen utama, sehingga dari *p* variabel akan didapat *m* komponen utama yang saling independen yang masih dapat mewakili keseluruhan persoalan $(m \le \mathrm{p})$ (Kroonenberg, 1997).

Penormalisasian **X** pada analisis komponen utama non linear harus memenuhi syarat $\mathbf{X^T X}$ = **I**, yang dinormalisasi oleh:

$$\mathbf{X}^T\mathbf{M}\mathbf{X} = pn\mathbf{I} \text{ dan } \mu'\mathbf{M}\mathbf{X} = 0,$$

yang berimplikasi bahwa **X** adalah skor yang standar jika $\mathbf{M}_j = \mathbf{I}$.

Cara penghapusan komponen utama dimulai dari prosedur seleksi akar ciri dari suatu persamaan $\left|\mathbf{Z}'\mathbf{Z} - \lambda\mathbf{I}\right| = 0$, dimana ***Z*** adalah hasil pembakuan dari peubah ***X***, yaitu:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}},\ Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}},\ Z_3 = \frac{(X_3 - \mu_3)}{\sqrt{\sigma_{33}}},\ldots, Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}.$$

Jika data multivariabel tanpa missing data, maka penaksir akar karakteristik diperoleh melalui $m^{-1}\mathbf{R}(\mathbf{Q})$ dimana $\mathbf{R}(\mathbf{Q})$ adalah matriks korelasi antara skor linear gabungan dari semua himpunan matriks **Q** pada semua dimensi (Gifi, 1990).

Jika **R** adalah matrik korelasi sampel berordo *m xm* dengan pasangan penaksir akar karakteristik dan penaksir vektor karakteristik yaitu:

$$(\hat{\lambda}_1, \hat{v}_1),\ (\hat{\lambda}_2, \hat{v}_2),\ \ldots,\ (\hat{\lambda}_m, \hat{v}_m)$$

dan $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ observasi yang distandarkan dengan matrik korelasi **R**, maka penaksir komponen utama ke-*k* adalah:

$$\hat{W}_k = \hat{\mathbf{v}}_k^T\mathbf{z} = \hat{v}_{1k}\, z_1 + \hat{v}_{2k} z_2 + \cdots + \hat{v}_{3k} z_{\mathrm{p}}, \quad k = 1, 2, \ldots, m$$

dimana $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \ldots \ge \hat{\lambda}_m \ge 0$, dan $\hat{\lambda}_{\mathrm{k}}$ adalah *penaksir akar karakteristik* dari matrik **R** dan mempunyai sifat tak bias yaitu $E(\hat{\lambda}_k) = \lambda_{\mathrm{k}}$.

Analisis komponen utama non linearberdasarkan pada *Teori Meet Loss*, dengan *loss function* homogenitas yaitu:

$$\sigma_{\mathrm{M}}(\mathbf{X},\mathbf{Y}) = \mathrm{m}^{-1}\sum_{k=1}^{\mathrm{m}}(\mathbf{X}-\mathbf{G}_{\mathrm{k}}\mathbf{Y}_{\mathrm{k}})^{T}(\mathbf{X}-\mathbf{G}_{\mathrm{k}}\mathbf{Y}_{\mathrm{k}})$$

Jika akar ciri λ diurutkan dari nilai terbesar sampai nilai terkecil, maka pengurutan komponen utama W_{k} berpadanan dengan pengurutan λ_k . Ini berarti bahwa komponen-komponen tersebut menerangkan proporsi keragaman terhadap respons *Y* yang semakin lama semakin kecil (Draper dan Smith, 1981).Strategi untuk mengeliminasi komponen utama dengan memperhatikan proporsi total varians komponen utama non linear yang diperoleh yaitu:

$$\frac{\lambda_q}{\lambda_1+\lambda_2+\ldots+\lambda_m} > 0{,}80.$$

Komponen utama W_{k}saling ortogonal sesamanya dan dibentuk melalui suatu hubungan:

$$W_k = v_{1k}Z_1 + v_{2k}Z_2 + v_{3k}Z_3 + \ldots + v_{pk}Z_p$$

Dimana $v_{1k}, v_{2k}, v_{3k}, \ldots, v_{pk}$ merupakan elemen dari vektor eigen yang berhubungan dengan λ_k dan **Z** merupakan variabel baku.

3. Regresi Logistik Biner

Model regresi logistik termasuk juga dalam model linier umum, dimana komponen acak tidak harus mengikuti sebaran normal, tapi harus termasuk dalam sebaran keluarga eksponensial. Sebaran Bernoulli termasuk dalam salah satu dari sebaran keluarga eksponensial, adalah sebaran dari peubah acak yang hanya mempunyai dua kategori yaitu bernilai 0 dan 1. Regresi logistik adalah pendekatan model matematik yang dapat digunakan untuk menggambarkan hubungan variabel prediktor (*X*) pada variabel respon (*Y*) yang biner (Hosmer dan Lemeshow, 1989). Model regresi logistik digunakan untuk melihat probabilitas terjadinya suatu keadaan dengan memperhitungkan faktor-faktor yang mempengaruhinya, dan membandingkan resiko munculnya suatu keadaan akibat suatu faktor setelah memperhitungkan faktor-faktor lain yang ada dalam model (Agresti, 1990).

Fungsi probabilitas distribusi Bernoulli adalah sebagai berikut.

$$f\left(Y_i\right) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, Y_i = 0,1,$$

dimana, $\pi_i$adalah probabilitas kejadian ke-i dan peubah acak ke-i. Jika Y_i = 0 maka$f$ (Y_i) = $(1-\pi_i)$, dan jika Y_i = 1 maka$f$ (Y_i) = π_i.

Fungsi model logit terletak antara range 0 dan 1 yang diperoleh dengan menggunakan fungsi logistik:

$$\pi(X_i) = \frac{1}{1+e^{-g(X_i)}},$$

dengan $g(X_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$.

Apabila tedapat pasangan data (X, Y) dengan *X* adalah variabel prediktor dan *Y* adalah variabel respon, dapat disajikan model regresi logistik dengan bentuk umum sebagai berikut:

$$\pi(X_i) = P(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi})}{1+\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi})}, \quad (1)$$

dimana $\pi(X_i) = P(Y=1)$ menyatakan proporsi skor atau nilai $Y=1$ di dalam populasi diantara semua dengan skala pasangan data yang mungkin. Besaran π merupakan nilai $P(Y=1)$, sering kali dinyatakan sebagai peluang peristiwa atau kasus yang ditentukan oleh skor $Y=1$. Peluang $\pi(X) = P(Y=1)$ tergantung pada skor atau nilai variabel prediktor X.

Transformasi logit dari model (1), yang linear dalam parameter-parameternya, yaitu:

$$\begin{aligned} g(X_i) &= \ln\left[\frac{\pi(X_i)}{1-\pi(X_i)}\right] \\ &= \ln \pi(X_i) - \ln(1-\pi(X_i)) \\ &= \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} \\ &= \boldsymbol{\beta}^T \mathbf{X} \end{aligned}$$

Dengan $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ dan $\mathbf{X} = (1, X_1, \ldots, X_k)^T$

(Agresti,2000)

4. Estimasi Parameter Model Regresi Logistik Komponen Utama Non Linear

Model regresi logistik komponen utama non linear diberikan oleh:

$$\pi(W) = \frac{\exp(\alpha_0 + \sum_{k=1}^{m} \alpha_k W_k)}{1+\exp(\alpha_0 + \sum_{k=1}^{m} \alpha_k W_k)} + \varepsilon, \quad (2)$$

dimana :

$\pi(W)$	:	probabilitas kejadian sukses pada Y = 1
α_0	:	konstanta
α_k	:	koefisien regresi logistik
W_k	:	Komponen utama yang terbentuk dari pengelompokkan variabel Z hasil standarisasi dari X, yang ditunjukkan pada bagian (2).
k	:	banyaknya komponen utama yang terbentuk dari proses komponen utama non linear mulai 1 hingga m
ε	:	error

Berdasarkan model (2), parameter yang akan ditaksir adalah α dengan menggunakan metode estimasi maksimum likelihood. Diketahui Y_i berdistribusi binomial, sehingga fungsi kepadatan peluangnya adalah sebagai berikut:

$$f(y_i) = (\pi(w_i))^{y_i} (1-\pi(w_i))^{1-y_i}, y_i = 0,1.$$

Selanjutnya fungsi likelihood diperoleh sebagai berikut:

$$L(\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_p)=\prod_{i=1}^{n} f(y_i)$$
$$=\prod_{i=1}^{n}(\pi(w_i))^{y_i}(1-\pi(w_i))^{1-y_i},$$

dan fungsi ln likelihoodnya adalah:

$$\ln L\left(\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_p\right)=\ln\left\{\prod_{i=1}^{n}(\pi(w_i))^{y_i}(1-\pi(w_i))^{1-y_i}\right\}$$
$$=\sum_{i=1}^{n}\left\{y_i(\pi(w_i))-\ln\left(1+e^{\pi(w_i)}\right)\right\}$$

Sehingga turunan pertamanya :

$$\frac{\partial \ln L\left(\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_m\right)}{\partial\alpha_0}=\sum_{i=1}^{n}\left\{y_i-\frac{e^{\pi(w_i)}}{1+e^{\pi(w_i)}}\right\}=\sum_{i=1}^{n}\left\{y_i-\frac{\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}{1+\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}\right\} \quad (3)$$

$$\frac{\partial \ln L\left(\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_m\right)}{\partial\alpha_1}=\sum_{i=1}^{n}\left\{y_i-\frac{w_{1i}e^{\pi(w_i)}}{1+e^{\pi(w_i)}}\right\}=\sum_{i=1}^{n}\left\{y_i-w_{1i}\frac{\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}{1+\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}\right\} \quad (4)$$

$$\frac{\partial \ln L\left(\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_m\right)}{\partial\alpha_2}=\sum_{i=1}^{n}\left\{y_i-\frac{w_{2i}e^{\pi(w_i)}}{1+e^{\pi(w_i)}}\right\}=\sum_{i=1}^{n}\left\{y_i-w_{2i}\frac{\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}{1+\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}\right\} \quad (5)$$

$$\vdots$$

$$\frac{\partial \ln L\left(\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_m\right)}{\partial\alpha_m}=\sum_{i=1}^{n}\left\{y_i-\frac{w_{mi}e^{\pi(w_i)}}{1+e^{\pi(w_i)}}\right\}=\sum_{i=1}^{n}\left\{y_i-w_{mi}\frac{\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}{1+\exp(\alpha_0+\sum_{k=1}^{m}\alpha_k w_k}\right\}. \quad (6)$$

Terlihat dari (3), (4), (5), dan (6), turunan pertama dari fungsi ln likelihood terhadap setiap parameter memberikan penyelesaian yang implisit, sehingga digunakan iterasi Newton Raphson dalam penaksiran parameternya, dan diperoleh sebagai berikut:

$$\begin{bmatrix} \hat{\alpha}_{0(t+1)} \\ \hat{\alpha}_{1(t+1)} \\ \hat{\alpha}_{2(t+1)} \\ \vdots \\ \hat{\alpha}_{m(t+1)} \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_{0(t)} \\ \hat{\alpha}_{1(t)} \\ \hat{\alpha}_{2(t)} \\ \vdots \\ \hat{\alpha}_{m(t)} \end{bmatrix} - \mathbf{H}^{-1}\mathbf{d},$$

dimana:

$\hat{\alpha}_{0(t+1)}$: konstanta pada iterasi ke-t+1

$\hat{\alpha}_{0(t)}$: konstanta pada iterasi ke-t

$\hat{\alpha}_{1(t+1)}$: parameter regresi pada iterasi ke-t+1

$\hat{\alpha}_{1(t)}$: parameter regresi pada iterasi ke-t

d : matriks turunan pertama terhadap parameternya, yaitu

$$\mathbf{d} = \begin{bmatrix} \dfrac{\partial \ln L(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m)}{\partial \alpha_0} \\ \dfrac{\partial \ln L(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m)}{\partial \alpha_1} \\ \dfrac{\partial \ln L(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m)}{\partial \alpha_2} \\ \vdots \\ \dfrac{\partial \ln L(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m)}{\partial \alpha_m} \end{bmatrix}$$

H : matriks turunan kedua terhadap parameter-parameternya.

Setelah diperoleh taksiran parameter regresi logistik dengan metode maksimum likelihood, maka taksiran model regresi logistik biner komponen utama non linear diperoleh sebagai berikut:

$$\hat{\pi}(W) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 W_1 + \hat{\alpha}_2 W_2 + \ldots + \hat{\alpha}_m W_m)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 W_1 + \hat{\alpha}_2 W_2 + \ldots + \hat{\alpha}_m W_m)} \quad (7)$$

Pers. (7) menunjukkan peubah yang digunakan masih peubah W yaitu peubah komponen-komponen utama yang terbentuk dari analisis komponen utama non linear yang dinyatakan oleh beberapa variabel Z yang merupakan hasil standarisasi dari variabel X yang diketahui. Akibatnya, pers (7) dapat ditransformasi ulang ke dalam variabel X berdasarkan persamaan dari analisis komponen utama non linear, sehingga variabel respon Y dapat dinyatakan dengan jelas oleh variabel prediktor X.

5. Kesimpulan

1. Penanganan masalah multikolinearitas pada data respon biner dengan mengandung variabel prediktor yang banyak dan berskala campuran dapat diselesaikan dengan model regresi logistik biner komponen utama non linear.
2. Estimasi parameter model regresi logistik komponen utama linear menggunakan metode maksimum likelihood dengan iterasi Newton Raphson, dan diperoleh:

$$\begin{bmatrix} \hat{\alpha}_{0(t+1)} \\ \hat{\alpha}_{1(t+1)} \\ \hat{\alpha}_{2(t+1)} \\ \vdots \\ \hat{\alpha}_{m(t+1)} \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_{0(t)} \\ \hat{\alpha}_{1(t)} \\ \hat{\alpha}_{2(t)} \\ \vdots \\ \hat{\alpha}_{m(t)} \end{bmatrix} - \mathbf{H}^{-1}\mathbf{d}$$

3. Estimasi model regresi logistik biner komponen utama non linear diperoleh:

$$\hat{\pi}(W) = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}_1 W_1 + \hat{\alpha}_2 W_2 + \ldots + \hat{\alpha}_m W_m)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}_1 W_1 + \hat{\alpha}_2 W_2 + \ldots + \hat{\alpha}_m W_m)}$$

DAFTAR PUSTAKA

[1] Agresti, A., 2000, *Categorical Data Analysis*. John Wiley & Sons. New York.

[2] Aguilera., 2000, " Principal Component Logistic Regression", *Proceding in Computation Statistics*, Physicalverlag,175-180.

[3] Gifi, A., 1990, *Nonlinear Multivariate Analysis*, Chichester, UK, Wiley.

[4] Islamiyati dan Talangko, 2010, "Identifikasi Prestasi Belajar Mahasiswa pada Pendidikan Tinggi dengan Analisis Komponen Utama Non Linear", *Proceding, Seminar Sains FMIPA Universitas Terbuka* Tangerang Banten.

[5] Johnson, R.A. & Winchern, D.W., 2002, *Applied Multivariate Statistical Analysis, 5th edition*, Pearson Education International.

[6] Kroonenberg, P.M., Harch, B.D., Basford K.E., and Cruickshank, A., 1997, "Combined Analysis of Categorical and Numerical Descriptors of Australian Groundnut Accessions Using Nonlinear Principal Componen Analysis", *Journal of Agrucultural, Biological, and Environmental Statistics*, Vol.2, No. 3, 294-312.