Partial Least Squares (PLS) Generalized Linear dalam Regresi Logistik [1]

Retno Subekti

Abstrak

Kasus multikolinieritas seringkali dijumpai dalam regresi yang mengakibatkan salah interpretasi model regresi yang terbentuk. Seperti halnya dalam regresi linear, dalam regresi logistic kasus multikolinearitas juga dapat menjadi masalah, karena adanya korelasi yang cukup tinggi antara variable prediktornya. Sehingga untuk mengatasi masalah seperti ini, akan digambarkan aplikasi prosedur partial least squares terhadap suatu kasus regresi logistic khususnya dalam contoh kasus makalah ini adalah regresi logistic ordinal.

Kata kunci : Partial Least Square generalized linier , multikolinieritas, regresi logistic.

Pendahuluan

Beberapa hal yang perlu diperhatikan ketika kita melakukan analisis regresi linear antara lain adalah asumsi-asumsi seperti normalitas, linearitas dan homoskedastisitas. Apabila diantara variable prediktor/independennya ternyata terdapat korelasi yang cukup tinggi, kita biasanya mengganggap adanya indikasi multikolinieritas yang cukup tinggi, beberapa sumber menyebutkan jika korelasi yang ada melebihi 80% maka kita perlu lebih serius menangani masalah multikolinieritas yang terjadi pada data sehingga tidak menimbulkan salah penafsiran saat menginterpretasikan output yang dihasilkan. Demikian juga pada analisis regresi logistic, asumsi tidak adanya multikolineritas ini perlu diperhatikan. Sehingga beberapa pendekatan untuk mengatasi masalah ini dapat dicoba untuk mendapatkan kesimpulan yang lebih beralasan, misalnya regresi stepwise, PCR dan PLS. Seperti pada [1] sebelumnya penulis mencoba memaparkan bagaimana PLS dalam contoh kasus regresi berganda, kali ini akan kita lihat bagaimana jika masalah yang dihadapi adalah regresi logistic.

Secara ringkasnya algoritma PLS-GLR :

1. Komputasi m komponen PLS t_h ($h = 1, 2, \ldots, m$)
2. GLR dari y pada m komponen PLS yang digunakan
3. Transformasi komponen PLS ke variable aslinya.

[1] Disampaikan dalam Seminar MIPA Nasional yang diselenggarakan oleh FMIPA UNY, Yogyakarta 25 Agustus 2007

Tujuan

Multikolinieritas adalah problem yang sering dijumpai saat melakukan regresi, sehingga perlu dilakukan pendekatan lain agar tidak menghasilkan interpretasi model ataupun koefisien regresi yang tidak tepat dan mungkin saja kesalahan pengambilan keputusan. Karena umumnya kita akan mengambil tindakan membuang variabel yang saling berkorelasi cukup tinggi, padahal kenyataannya variabel tersebut cukup berpengaruh terhadap variabel responnya. Selain karena adaya korelasi yang cukup tinggi antar variabel independennya, multikolinieritas dapat juga disebabkan karena jumlah observasi yang relatif kecil dengan variabel independen yang cukup banyak. Jika pada makalah sebelumnya penulis mencoba mengaplikasikan prosedur PLS pada kasus regresi berganda, kali ini akan dicoba bagaimana jika masalah multikolinieritas terjadi pada kasus regresi logistik.

Regresi Partial Least Square Secara Singkat

Jika terdapat sejumlah p variabel independen dan sebuah variabel dependen/respon, dalam proses PLS kita asumsikan semua variabel sudah dalam bentuk baku/standard.

Model regresi PLS dengan m komponen dirumuskan sebagai :

$$Y = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_h^* x_j\right) + \text{sisa} \qquad (1)$$

Perhitungan komponen pls pertama, $t_1 = Xw_1^*$ didefinisikan sebagai

$$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \text{cov}(y, x_j)^2}} \sum_{j=1}^{p} \text{cov}(y, x_j)^2 x_j \qquad (2)$$

variabel x_j ini dipilih yang berkorelasi tinggi dengan y dan cukup kuat variabilitasnya.

Selanjutnya untuk koefisien regresi a_{1j} dapat digunakan untuk menaksir seberapa penting variabel x_j dalam pembentukan t_1. Regresi sederhana y terhadap x_j dirumuskan:

$$Y = a_{1j} x_j + \text{sisa} \qquad (3)$$

Jika a_{1j} tidak signifikan atau tidak berbeda nyata dengan 0 maka dalam (2) setiap kovariansi yang tidak signifikan dapat diganti dengan 0 dan artinya kita dapat mengabaikan hubungan variabel independennya.

Perhitungan komponen pls kedua, t_2

Komponen pls kedua, t_2 didefinisikan sebagai

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}(y_1, x_{1j})^2}} \sum_{j=1}^{p} \mathrm{cov}(y_1, x_{1j}) x_{1j} \quad (4)$$

dimana sebelumnya dilakukan dua hal yaitu :

1. regresi sederhana y terhadap setiap x_j
2. regresi x_j terhadap t_1

$$Y = c_1 t_1 + y_1 \quad (5)$$

$$X_j = p_{1j} t_1 + x_{1j} \quad (6)$$

komponen pls kedua dapat juga dituliskan sebagai

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}(y_1, x_j | t_1)^2}} \sum_{j=1}^{p} \mathrm{cov}(y_1, x_j | t_1) x_{1j} \quad (7)$$

Karena korelasi parsial antara y dan x_j jika diketahui t_1 didefinisikan sebagai korelasi antara residual y_1 dan x_{1j} maka kovariansi parsial antara y dan x_j diketahui t_1 juga didefinisikan sebagai kovariansi antara residu y_1 dan x_{1j}.

$$\mathrm{Cov}(y, x_j | t_1) = \mathrm{cov}(y_1, x_{1j}) \quad (8)$$

untuk melihat kontribusi x_j dalam pembentukan t_2, dapat diketahui melalui regresi y terhadap t_1 dan x_j.

$$Y = c_{1j} t_1 + a_{2j} x_j + \text{residu} \quad (9)$$

Sedangkan uji koefisien regresi a_{2j} dapat digunakan untuk menaksir seberapa penting variabel x_{1j} dalam pembentukan t_2. Jika tidak signifikan maka hubungan variabel independennya tidaklah penting dalam pembentukan komponen pls kedua tersebut.

Perhitungan komponen PLS berikutnya dan aturan penghentiannya.

Dengan prosedur yang sama seperti mencari t_2 , dicari komponen PLS ke-h, $t_h = Xw_h^*$.

Pencarian komponen baru berhenti jika semua kovariansi parsialnya tidak signifikan.

Persamaan Regresi *Partial Least Square*

Dalam (1) koefisien c_h diestimasi oleh regresi berganda dari y terhadap komponen PLS t_h. Persamaan regresi estimasinya selanjutnya dapat ditulis ke dalam variabel x_j yang asli.

$$\hat{y} = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^* x_j\right) = \sum_{j=1}^{m} \left(\sum_{h=1}^{m} c_h w_{hj}^*\right) x_j = \sum_{j=1}^{p} b_j x_j \quad (10)$$

Regresi logistic

Seperti halnya pada saat kita akan melakukan analisis regresi, kita biasanya melihat bagaimana pola sebaran datanya terlebih dahulu, apakah ada kecenderungan linier, kuadratik atau pola lainnya. Jika variabel dependen/responnya adalah data kuantitatif dan adanya pola linier maka analisis regresi linier dimungkinkan untuk digunakan dalam mengolah data tersebut. Tetapi jika kita menghadapi kasus dimana variable respon adalah data kualitatif (nominal, ordinal, kategorik), misalnya Y mempunyai dua nilai, kita anggap 0 dan 1 (Y dinamakan dikotomus) maka regresi linier biasa bukanlah alat yang tepat untuk mengolah data tersebut melainkan kita perlu mengubahnya menjadi regresi dalam bentuk peluang atau regresi logistik.

$$Y = \begin{cases} 0 \\ 1 \end{cases}$$

$E(Y) = P(Y = 1) = \pi$

Jadi model regresi logistic

$$\pi = P(Y = 1) = \frac{e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k}}$$

$P(Y = 1) = \pi$ maka $P(Y = 0) = 1-\pi = \frac{1}{1+e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k}}$

Transformasi dari π ini yang dinamakan logit transformation, yaitu

$$\text{Odds} = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi}{1-\pi} = e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k}$$

$$\text{Log odds} = \log e^{\alpha+\beta_1 X_1+\cdots+\beta_k X_k} = \alpha + \beta_1 X_1 + \beta X_2 + \cdots \beta_k X_k$$

Sedangkan jika variable respondennya adalah data ordinal (bertingkat) maka regresi logistic yang dipilih adalah regresi logistic ordinal.

Model Regresi logistic Ordinal

$$P(y \leq k) = \frac{e^{\alpha_k+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\alpha_k+\beta_1 X_1+\cdots+\beta_k X_k}}$$

Contoh Kasus Regresi Logistic

Dalam makalah ini digunakan data Bordeaux Wine dalam Bastien, P., Vinzi, VE., Tenenhaus, M., 2004, yang ternyata dengan bantuan software yang sudah banyak beredar seperti SPSS 14 dan Minitab 14 kita dapat mengaplikasikan algoritma di atas untuk

mendapatkan komponen PLS. Bahkan MINITAB 14 atau 15 sudah menyediakan tambahan menu untuk regresi berganda PLS dan regresi multivariate PLS.

Table 1. Korelasi Pearson antara variable independent

	TEMPERATURE	SUNSHINE	HEAT
sunshine	0,712		
	0,000		
heat	0,865	0,646	
	0,000	0,000	
rain	-0,410	-0,473	-0,401
	0,016	0,005	0,019

Terlihat jelas adanya korelasi yang cukup erat antara variabel *sunshine, temperature* dan *heat*, dengan masing-masing angka korelasinya > 0,5. Sehingga mengindikasikan adanya multikolinieritas yang perlu diperhatikan.

Dengan variable independent distandardisasikan lebih dulu, maka model regresi logistic ordinalnya adalah :

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	Upper
Const(1)	-2,66382	0,926633	-2,87	0,004			
Const(2)	2,29406	0,978207	2,35	0,019			
t	3,42677	1,80293	1,90	0,057	30,78	0,90	1054,17
s	1,74618	1,07602	1,62	0,105	5,73	0,70	47,24
h	-0,889079	1,19488	-0,74	0,457	0,41	0,04	4,28
r	-2,36683	1,12922	-2,10	0,036	0,09	0,01	0,86

$$P(y = 1) = \frac{e^{-2,6638+3,4268\,t+1,7462\,s-0,8891\,h-2,3668\,r}}{1+e^{-2,6638+3,4268\,t+1,7462\,s-0,8891\,h-2,3668\,r}}$$

$$P(y \leq 2) = \frac{e^{-2,2941+3,4268\,t+1,7462\,s-0,8891\,h-2,3668\,r}}{1+e^{-2,2941+3,4268\,t+1,7462\,s-0,8891\,h-2,3668\,r}}$$

Dimana hanya variable *temperature* dan *rain* yang signifikan, sedangkan variable *sunshine* dan *heat* tidak signifikan karena p-valuenya > 0,05 padahal variable *sunshine* dan *heat* cukup berperan penting dalam mempengaruhi kualitas *wine*. Ini yang menjadi salah satu akibat dari adanya multikolinieritas. Untuk mengetahui apakah memang kedua variable tersebut berpengaruh signifikan terhadap *wine* dapat dilihat dari table berikut, yaitu hubungan regresi logistik ordinal antara variable *wine* dengan masing-masing variable independent.

Table 2. Koefisien regresi

PREDICTOR	COEF	SE COEF	Z	P
t	3,01169	0,795932	3,78	0,000
s	3,34015	0,886485	3,77	0,000
h	2,14457	0,607721	3,53	0,000
r	-1,79056	0,568878	-3,15	0,002

Jika model regresi logistic ordinal digunakan maka nilai prediksi dibandingkan observasi untuk variable responnya adalah :

observasi \ prediksi	good	Average	poor
Good	8	3	0
Average	2	8	1
Poor	0	1	11

Dari table di atas terlihat bahwa ada 7 prediksi yang tidak sesuai dengan observasi.

Pembentukan PLS regresi logistic ordinal

1) Perhitungan komponen PLS pertama, t_1

Untuk membangun t_1 dapat dilihat dari tabel 2, dimana semua variable independent ternyata signifikan berpengaruh terhadap kualitas *wine*, jadi t_1 dibangun oleh keempat variable tersebut.

$$t_1 = \frac{3{,}0117x_1^* + 3{,}3401x_2^* + 2{,}1446x_3^* - 1{,}7906x_4^*}{\sqrt{3{,}0117^2 + 3{,}3401^2 + 2{,}1446^2 + 1{,}7906^2}}$$

$$= 0{,}5693 \text{ temperature} + 0{,}6314 \text{ sunshine} + 0{,}4054 \text{ heat} - 0{,}3385 \text{ rain}$$

Ordinal Logistic Regression: quality versus komponen PLS1

```
Logistic Regression Table
                                                    Odds     95% CI
Predictor           Coef   SE Coef       Z      P  Ratio  Lower  Upper
Const(1)       -2,26510  0,864387  -2,62  0,009
Const(2)        2,29912  0,848043   2,71  0,007
komponen PLS1   2,68776  0,714920   3,76  0,000  14,70   3,62  59,68
Log-Likelihood = -15,251
Test that all slopes are zero: G = 44,145, DF = 1, P-Value = 0,000
```

2) Perhitungan komponen PLS kedua, t_2

Sebelumnya perlu dilihat dulu apakah masih ada variable independent yang membangun komponen PLS kedua ini. Untuk mengetahuinya kita lihat dari regresi logistic antara kualitas dengan t_1 dan masing-masing variable independent. Berikut hasil output masing-masing koefisien regresi dan nilai p-valuenya.

```
Logistic Regression Table
Predictor       Coef   SE Coef      Z      P
```

t	-0,630714	1,51211	-0,42	0,677
s	0,646112	1,24113	0,52	0,603
h	-1,94076	1,17392	-1,65	0,098
r	-0,979772	0,859642	-1,14	0,254

karena p-value masing-masing variable independent ternyata > 0,05 maka semua variable independennya sudah tidak ada yang signifikan membangun komponen PLS yang kedua. Sehingga tidak ada komponen PLS baru lagi atau komponen PLS yang terbentuk hanya satu.

Kesimpulan

Jadi regresi logistic ordinal PLS nya

$$P(y = 1) = \frac{e^{-2,265+2.688\,t}}{1+e^{-2,265+2,688\,t}}$$

$$P(y \leq 2) = \frac{e^{2,299+2.688\,t}}{1+e^{2,299+2,688\,t}}$$

Jika diubah ke bentuk variabel aslinya maka :

$$2,688\ t_1 = 1,5303\ t + 1,6972\ s + 1,0897\ h - 0,9099\ r$$

$$P(y = 1) = \frac{e^{-2,265+1,5303\,t+1,6972\,s+1,0897\,h\ \ 0,9099\,r}}{1+e^{-2,265+1,5303\,t+1,6972\,s+1,0897\,h\ \ 0,9099\,r}}$$

$$P(y \leq 2) = \frac{e^{2,299+1,5303\,t+1,6972\,s+1,0897\,h\ \ 0,9099\,r}}{1+e^{2,299+1,5303\,t+1,6972\,s+1,0897\,h\ \ 0,9099\,r}}$$

Daftar Pustaka

[1] Bastien, P., Vinzi, VE., Tenenhaus, M., 2004. *Partial Least Square Generalized Linear Regression*. Computational Statistics & Data Analysis 48 (2005) 17-46

[2] Herve Abdi (2003). *Partial Least Square (PLS) Regression*. Encyclopedia of Social Sciences Research Methods

[3] Hosmer,.D.W&Lemeshow,S. (1989). *Applied Logistic Regression*. New York, NY : John Willey & Sons

[4] Myers, R.H. (1996). *Classical and Modern Regression with Applications*. Boston : PWS-KENT Publishing Company

[5] Neter, J., W. Wasserman, Kutner, MH. (1990). *Applied Linear Statistical Models Third Edition*, Richard D. Irwin, Inc., Homewood, Illinois