

PEMBANDINGAN REGRESI KOMPONEN UTAMA DAN ALGORITMA PROYEKSI DALAM MENGATASI MULTIKOLINEARITAS

By Khoirin Nisa

PEMBANDINGAN REGRESI KOMPONEN UTAMA DAN ALGORITMA PROYEKSI DALAM MENGATASI MULTIKOLINEARITAS

Khoirin Nisa

Jurusan Matematika FMIPA UNILA

Email : Nisa_stat@unila.ac.id

ABSTRACT

Principal component regression (PCR) is often used in regression with multicollinearity, but it is not a method which directly maximizes coefficient of determination (R^2) since the first k principal components (PCs) with the largest variances do not always the variables that are most highly correlated with the response variable. To maximize the R^2 we have to select the PCs by a procedure such as the stepwise procedure. Projection algorithm is a method proposed by Filzmoser and Croux (2001) which can be used to overcome multicollinearity and can also maximize the R^2 . In this study we aimed to compare the R^2 resulted by the PCR, Stepwise-PCR and the projection algorithm by Monte Carlo simulation. We generated data from normal distribution with 1900 replications. The result shows that projection algorithm and Stepwise-PCR yield almost the same R^2 and each of the two methods performs better than the PCR.

Keywords : Principal component regression, projection algorithm, multicollinearity.

I. PENDAHULUAN

Salah satu masalah yang sering dihadapi dalam analisis regresi berganda adalah multikolinieritas, yaitu adanya hubungan linier (korelasi) antara dua variabel bebas atau lebih dalam suatu persamaan regresi. Hal ini menyebabkan matriks $X^T X$ menjadi singular sehingga persamaan normalnya tidak lagi mempunyai jawaban yang tunggal, akibatnya dugaan koefisien regresi kuadrat terkecil [$b = (X^T X)^{-1} X^T Y$] memiliki simpangan baku yang besar yang menjadikannya tidak lagi terandalkan.

Berbagai metode untuk mengatasi masalah multikolinearitas telah dikembangkan, salah satunya adalah Algoritma Proyeksi (Filzmoser dan Croux, 2001). Algoritma ini mengatasi multikolinieritas dengan cara membentuk komponen-komponen utama yang saling bebas dari variabel prediktor yang sekaligus juga memiliki korelasi yang tinggi dengan Y .

Tulisan ini membahas metode algoritma proyeksi dalam mengatasi multikolinieritas dan membandingkannya dengan metode regresi komponen utama (RKU) dan regresi komponen utama stepwise (RKU Stepwise) dengan melakukan simulasi Monte Carlo menggunakan program SAS/IML.

Algoritma Proyeksi

12 Misalkan diperoleh suatu data dari hasil pengamatan yang mempunyai p buah variabel bebas (X) dan sebuah variabel tak-bebas (Y) sehingga model regresi linear berganda yang terbentuk sebagai berikut,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (1)$$

Ditulis dalam bentuk matriks,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Konsep dasar dalam algoritma proyeksi adalah bagaimana mencari komponen-komponen utama $z_1, z_2, \dots, z_k$ yang menghasilkan nilai SMC setinggi mungkin dengan syarat bahwa komponen-komponen ini saling bebas dan memiliki ragam 1. Pemaksimuman nilai SMC dilakukan secara sequensial. Pertama dengan memilih $z_1 = \mathbf{X}\mathbf{b}$ yang memaksimumkan kudrat korelasi dengan Y , selanjutnya mencari komponen lain yang juga memaksimumkan kudrat korelasi dengan Y dan tidak berkorelasi dengan komponen-komponen sebelumnya.. Jadi untuk mencari z_1 dilakukan dengan menentukan sebuah vektor $\mathbf{b}$ yang memaksimumkan fungsi berikut,

$$\mathbf{b} \rightarrow |\text{Corr}(\mathbf{Y}, \mathbf{X}\mathbf{b})| \quad (2)$$

dimana

$$\mathbf{X}\mathbf{b} = b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

Korelasi pada persamaan (2) merupakan korelasi biasa antara dua vektor kolom. Karena fungsi (2) invarian terhadap perkalian skalar maka perlu ditambahkan syarat $\text{Var}(\mathbf{X}\mathbf{b}) = 1$. Agar nilai $\text{Corr}(\mathbf{Y}, \mathbf{X}\mathbf{b})$ mencapai nilai maksimum, maka $\mathbf{b}$ harus merupakan dugaan kuadrat terkecil dari persamaan (1), yaitu $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

Fungsi (2) dapat ditulis menjadi

$$\mathbf{b} \rightarrow |\text{Corr}(\mathbf{Y}, \mathbf{z}_1)|. \quad (3)$$

Untuk menghindari masalah multikolineritas maka yang diinginkan bukan maksimum global dari fungsi (3) tetapi nilai maksimum fungsi pada himpunan

$$B_{n,1}(\mathbf{X}) = \left\{ \frac{\mathbf{X}_i}{\|\mathbf{X}_i\|}; i = 1, 2, \dots, n \right\}.$$

Komponen kedua diperoleh dengan menyeleksi semua vektor yang mempunyai korelasi nol dengan z_1 . Misalkan $\mathbf{X}_j$ menotasikan kolom ke- j dari data matriks $\mathbf{X}$. Semua vektor data $\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$ diregresikan terhadap komponen pertama (z_1). Galat yang diperoleh dari masing-masing persamaan regresi dinotasikan dengan $\mathbf{Y}^1, \mathbf{X}_1^1, \mathbf{X}_2^1, \dots, \mathbf{X}_p^1$, selanjutnya variabel-variabel baru ini digunakan untuk mencari komponen z_2 dan semua variabel ini tidak berkorelasi dengan z_1 . Komponen kedua z_2 diperoleh dengan mencari sebuah vektor $\mathbf{b}$ yang menghasilkan nilai tertinggi pada fungsi

$$\mathbf{b} \rightarrow |\text{Corr}(\mathbf{Y}^1, \mathbf{X}^1 \mathbf{b})| \quad (4)$$

Agar nilai $\text{Corr}(\mathbf{Y}^l, \mathbf{X}^l \mathbf{b})$ mencapai nilai maksimum, maka $\mathbf{b}$ juga harus merupakan dugaan kuadrat terkecil. Sehingga diperoleh komponen utama kedua $\mathbf{z}_2 = \mathbf{X}^1 \mathbf{b}$ yaitu

$$\mathbf{z}_2 = \mathbf{X}^1 ((\mathbf{X}^1, \mathbf{X}^1)^{-1} \mathbf{X}^1, \mathbf{Y}^1)$$

Nilai-nilai fungsi (4) dicari pada himpunan

$$B_{n,2}(\mathbf{X}) = \left\{ \frac{\mathbf{X}_i^1}{\|\mathbf{X}_i^1\|}; n = 1, 2, \dots, n \right\}.$$

Untuk mencari komponen lain z_l ($l = 3, 4, \dots, k$), dilakukan dengan cara yang sama seperti pada pembentukan z_2 . . Komponen z_l diperoleh dengan cara sebagai berikut

$$\begin{aligned} z_{l-1} &= b Y^{l-2} + \varepsilon_0 \\ z_{l-1} &= b X_1^{l-2} + \varepsilon_1 \\ z_{l-1} &= b X_2^{l-2} + \varepsilon_2 \\ &\vdots \\ z_{l-1} &= b X_p^{l-2} + \varepsilon_p \end{aligned} \quad (5)$$

Galat dari masing-masing persamaan regresi pada (5) merupakan variabel-variabel baru untuk mencari komponen z_l , yaitu

$$\begin{aligned} \varepsilon_0 &= Y^{l-1} \\ \varepsilon_1 &= X_1^{l-1} \\ \varepsilon_2 &= X_2^{l-1} \\ &\vdots \\ \varepsilon_p &= X_p^{l-1} \end{aligned} \quad (6)$$

Persamaan- persamaan pada (6) dinotasikan dengan $\mathbf{Y}^{l-1}$, $\mathbf{X}_1^{l-1}$, $\mathbf{X}_2^{l-1}$, ..., $\mathbf{X}_p^{l-1}$ dan semua vektor ini tidak berkorelasi dengan semua $\mathbf{z}$ sebelumnya. Komponen pertama z_l diperoleh dengan memberikan sebuah vektor $\mathbf{b}$ yang menghasilkan nilai tertinggi pada fungsi

$$\mathbf{b} \rightarrow |\text{Corr}(\mathbf{Y}^{l-1}, \mathbf{X}^{l-1} \mathbf{b})| \quad (7)$$

Solusi dari persamaan (7) di hitung pada himpunan

$$B_{n,l}(\mathbf{X}) = \left\{ \frac{\mathbf{X}_i^{l-1}}{\|\mathbf{X}_i^{l-1}\|}; 1 \leq i \leq n \right\}.$$

7 II. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah studi pustaka dan simulasi data dengan menggunakan komputer. Tahapan-tahapan yang dilakukan adalah sebagai berikut :

1. Membuat makro program SAS/IML untuk algoritma proyeksi.
2. Simulasi Data. Simulasi dilakukan dengan membangkitkan data sampel berdistribusi normal berukuran $n=100$ dengan banyaknya peubah bebas X dalam beberapa dimensi p , dimana $p=2,3,4,\dots,20$. Selanjutnya untuk memunculkan korelasi antar peubah dibangun matriks transformasi T berukuran $p \times p$ yang elemen-elemennya diperoleh dari distribusi uniform. Vektor-vektor kolom dari matriks T merupakan koefisien kombinasi linier dari fungsi peubah asal yang digunakan untuk menghasilkan peubah-peubah baru yang saling berkorelasi. Peubah respon Y diperoleh dari hasil penjumlahan peubah bebas X dengan galat ε dimana ε dibangun dari distribusi normal baku, yaitu $\varepsilon \sim N(0,1)$. Ulangan dalam simulasi dilakukan sebanyak 1900 kali, yaitu 100 ulangan untuk masing-masing jumlah peubah $p=2,3,4,\dots,20$. Selanjutnya seluruh data pada setiap ulangan dianalisis dengan menggunakan tiga metode, yaitu, metode regresi komponen utama, regresi komponen utama stepwise, dan metode algoritma proyeksi. Pembangkitan dan analisis data dengan ketiga metode dilakukan dengan menggunakan software SAS 8.0.
3. Membandingkan hasil yang diperoleh.

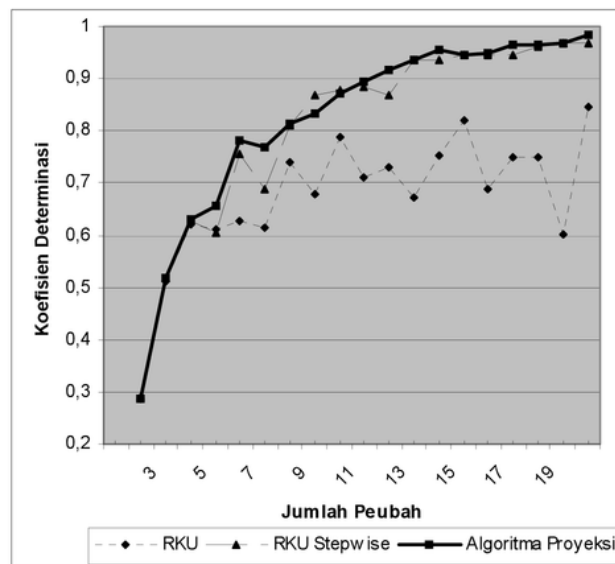
III. HASIL PENELITIAN DAN PEMBAHASAN

Hasil simulasi data dan analisis dengan menggunakan ketiga metode memberikan nilai koefisien determinasi yang disajikan pada Tabel 1. Pada Tabel 1 terlihat bahwa rata-rata koefisien determinasi metode algoritma proyeksi cenderung lebih baik dari pada kedua metode lainnya. Data pada Tabel 1 disajikan dalam bentuk grafik pada Gambar 1.

Tabel 1. Rata-rata koefisien determinasi ketiga metode

Jumlah Peubah	Rata-rata Koefisien Determinasi		
	RKU	RKU Stepwise	Algoritma Proyeksi
2	0.2885	0.2918	0.2886
3	0.5112	0.5185	0.5189
4	0.6211	0.6262	0.6296
5	0.6102	0.6055	0.6578
6	0.6261	0.7557	0.7801
7	0.6154	0.6875	0.7689
8	0.7393	0.8116	0.8129
9	0.6779	0.8693	0.8328
10	0.7875	0.8777	0.8716
11	0.7121	0.8854	0.8961
12	0.7292	0.8685	0.9155
13	0.6726	0.9365	0.9368

14	0.7537	0.9362	0.9557
15	0.8197	0.9459	0.9465
16	0.6877	0.9469	0.9484
17	0.7508	0.9469	0.9658
18	0.7495	0.9617	0.9634
19	0.6025	0.9675	0.9685
20	0.8452	0.9674	0.9835



Gambar 1. Plot koefisien determinasi metode RKU, RKU Stepwise dan Algoritma proyeksi

Dari Gambar 1 terlihat bahwa rata-rata koefisien determinasi yang dihasilkan oleh metode algoritma proyeksi lebih tinggi dibandingkan dengan metode RKU dan metode RKU Stepwise. Plot rata-rata koefisien determinasi terhadap jumlah peubah dengan menggunakan algoritma proyeksi cenderung semakin meningkat dengan semakin bertambahnya jumlah peubah yang dimasukkan ke dalam model. Hal ini berlaku pula untuk metode RKU stepwise. Sedangkan plot rata-rata koefisien determinasi yang dihasilkan oleh metode RKU terlihat tidak stabil.

Secara umum metode algoritma proyeksi dan RKU memiliki prinsip yang sama, yaitu menghilangkan korelasi antar peubah bebas dengan membentuk komponen-komponen utama yang tidak berkorelasi yang merupakan kombinasi linier dari peubah asal. Komponen utama pada RKU dibentuk dengan menggunakan vektor eigen dari matriks ragam koragam atau matriks korelasi, selanjutnya komponen utama yang memiliki keragaman terbesar dipilih sebagai peubah bebas baru dan diregresikan dengan peubah respon. Hal ini dapat mengatasi masalah kolinieritas

dengan baik namun sayangnya komponen utama yang memiliki keragaman terbesar tidak selalu memiliki korelasi tinggi dengan peubah respon, akibatnya tidak ada jaminan bahwa model yang terbentuk akan memiliki koefisien determinasi yang tinggi. RKU Stepwise merupakan metode perbaikan dari RKU, yaitu komponen-komponen utama yang terbentuk diseleksi berdasarkan korelasinya dengan peubah respon, sehingga model yang diperoleh akan memiliki koefisien determinasi yang besar. Algoritma proyeksi identik dengan metode RKU stepwise, hanya bedanya algoritma proyeksi melakukan hal tersebut di atas secara simultan, yaitu membentuk komponen utama yang saling bebas dan sekaligus memiliki korelasi yang tinggi dengan peubah respon dalam satu tahap.

Metode algoritma proyeksi dan RKU menggunakan tiga komponen utama sebagai peubah bebas dalam model regresi pada semua jumlah peubah kecuali untuk $p=2$, untuk dua peubah X maka hanya tersedia dua komponen utama saja sehingga komponen utama yang digunakan dalam model regresi juga hanya dua. Sedangkan metode RKU stepwise menggunakan semua komponen utama yang secara statistik memberikan kontribusi signifikan dalam model, yaitu yang memiliki nilai p -value pada uji t kurang dari 0.05. Dalam simulasi ini, banyaknya komponen utama yang digunakan oleh RKU stepwise rata-rata lebih dari 50% dari peubah asal. Jadi untuk $p \geq 6$ maka RKU stepwise menggunakan komponen utama yang lebih banyak dibandingkan dengan metode RKU dan algoritma proyeksi. Secara teoritis, semakin banyak jumlah peubah bebas dalam model maka akan meningkatkan koefisien determinasi. Namun penelitian ini telah memperlihatkan bahwa dengan jumlah peubah bebas yang lebih sedikit dalam model, metode algoritma proyeksi dapat memberikan hasil yang sebanding bahkan lebih baik dari metode RKU stepwise.

10

IV. KESIMPULAN

Dari hasil penelitian ini maka dapat ditarik kesimpulan sebagai berikut :

Koefisien determinasi yang dihasilkan oleh metode algoritma proyeksi lebih baik dibandingkan dengan metode RKU dan RKU stepwise, dan nilainya semakin meningkat seiring dengan semakin banyaknya jumlah peubah bebas walaupun hanya tiga kompoenen utama yang digunakan dalam model. Ini berarti metode algoritma proyeksi sangat baik dan efektif digunakan dalam analisis regresi dengan multikolinieritas.

DAFTAR PUSTAKA

- [1] Croux, C dan Dehon, C. 2001. Estimator of the Multiple Correlation Coefficient : Local Robustness and Confidence Interval. *Statistical Papers*. <http://www.econ.kuleuven.ac.be/christophe.croux>.
- [2] Croux, C dan Gazen, A.R. 2000. High Breakdown Estimators for Principal Components : the Projection-Pursuit Approach Revisited. *Computational Statistics*. Physica-Verlag, Heidelberg.
- [3] Croux, C., Ruiz-Gazen, A., 1996. A fast algorithm for robust principal components based on projection pursuit. *Compstat: Proceedings in Computational Statistics*, Heidelberg: Physica-Verlag, hal 211-216.
- [4] Davies, A.M.C. 2005. Back to basic : application of Principal Component Analysis. *SpectroscopyEurope*. Vol 17 No 25.
- [5] Filzmoser, P dan Cristophe Croux. 2001. A Projection Algorithm for regression with collinearity, dalam *Clasification, Clustering and Data Analysis*. Springer-Verlag, Berlin.
- [6] Filzmoser, P. 2001. Robust Principal Component Regression dalam S. Aivazian, Yu. Karin dan H. Rieder : *Computer Data Analysis and Modeling Robust and Computer Intensive Methods*. Belarusian State University.
- [7] Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-KENT Publishing Company, Boston

PEMBANDINGAN REGRESI KOMPONEN UTAMA DAN ALGORITMA PROYEKSI DALAM MENGATASI MULTIKOLINEARITAS

ORIGINALITY REPORT

13%

SIMILARITY INDEX

PRIMARY SOURCES

1	Filzmoser, Peter, and Christophe Croux. "A Projection Algorithm for Regression with Collinearity", Studies in Classification Data Analysis and Knowledge Organization, 2002. Crossref	44 words — 2%
2	www.tandfonline.com Internet	27 words — 1%
3	es.scribd.com Internet	27 words — 1%
4	www.emeraldinsight.com Internet	24 words — 1%
5	www.scribd.com Internet	23 words — 1%
6	www.statistik.tuwien.ac.at Internet	21 words — 1%
7	Sudirman S., Sri Kurniati A.. "ANALISIS PERBANDINGAN PENGGUNAAN TIPE PENYALAAN KONTROL JARAK SAMA DAN SUDUT SAMA PADA PENYEARAH TERKENDALI TIGA PHASA", MAKARA of Technology Series, 2010 Crossref	14 words — 1%
8	docobook.com Internet	14 words — 1%

9	wis.kuleuven.be Internet	12 words — 1%
10	eprints.radenfatah.ac.id Internet	12 words — 1%
11	www.econ.kuleuven.ac.be Internet	11 words — 1%
12	kalibanteng.wordpress.com Internet	11 words — 1%
13	www.peipfi-komdasulsel.org Internet	10 words — < 1%
14	Joaquín Pacheco, Silvia Casado, Santiago Porras. "Exact methods for variable selection in principal component analysis: Guide functions and pre-selection", Computational Statistics & Data Analysis, 2013 Crossref	7 words — < 1%
15	Ricardo A. Maronna, R. Douglas Martin, Víctor J. Yohai. "Robust Statistics", Wiley, 2006 Crossref	6 words — < 1%

EXCLUDE QUOTES OFF
EXCLUDE BIBLIOGRAPHY ON

EXCLUDE MATCHES OFF