S-30

SENSITIFITAS INDIKATOR KESELURUHAN MULTIKOLINEARITAS DALAM MODEL REGRESI LINEAR MULTIPEL

Dien Sukardinah

Staf Pengajar Jurusan Statistika FMIPA Universitas Padjadjaran.

Abstrak

Makalah ini menyajikan indikator keseluruhan multikolinearitas, yaitu Bilangan Kondisi (Belsley dkk., 1980), indikator *Red* (Kovacs P., Petres T.,Toth L., 2005), indikator *DEF* (Curto J.D., Pinto J.C., 2007). Bilangan Kondisi mengandung unsur nilai eigen dari matriks korelasi variabel bebas, indikator *Red* memungkinkan kita mengkuantifikasi persentase kolinearitas dari 0% sampai dengan 100%, indikator *DEF* diperoleh berdasarkan pada koefisien jalur, yang memungkinkan kita menilai pengaruh langsung maupun tak langsung dari satu variabel terhadap variabel lainnya. Dengan simulasi replikasi 1000 kali diperoleh proporsi kemunculan /terjadinya multikolinearitas dari indikator-indikator multikolinearitas keseluruhan, kemudian dari grafik fungsi indikator *Red, DEF* dan bilangan kondisi dapat dilihat sensitifitas dari masing-masing indikator.

Kata kunci : Regresi Linear Multipel; indikator multikolinearitas; matriks korelasi; analisis jalur.

1 Pendahuluan

Misalkan diasumsikan model regresi linear multipel dengan m variabel regresor sebagai berikut;

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + \varepsilon_i \; ; \; i : 1,2,\ldots,N \qquad \ldots \qquad (1)$$

atau dalam bentuk matriks sebagai berikut;

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \ldots \qquad (2)$$

Menurut Ragnar Frisch (1934) multikolinearitas terjadi karena adanya hubungan linear sempurna / hampir sempurna di antara variabel-variabel regresor dalam suatu model regresi (Gujarati, 2003). Multikolinearitas timbul ketika terdapat penyimpangan keortogonalan dari himpunan regresor dan hal ini bisa menjadi masalah yang serius

(Grill, 1998). Multikolinearitas dapat dibedakan menjadi multikolineritas sempurna dan multikolinearitas hampir sempurna, sebagaimana yang dijelaskan berikut ini. Multikolinearitas sempurna berarti adanya hubungan linear sempurna antar variabel regresor, $\sum_{i=1}^{m} a_i x_i = 0$ dengan $a_{i;}$ konstanta sembarang yang tidak semuanya nol untuk setiap *i;1,2,...,m.* Hal ini mengakibatkan determinan (**X'X**) = 0, sehingga matriks (**X'X)**$^{-1}$ tidak ada dan taksiran kuadrat terkecil biasa (`ordinary least square' – OLS)

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},\ \hat{Var}(\hat{\beta}) = S^2(\mathbf{X}'\mathbf{X})^{-1},\ S^2 = \frac{\mathbf{e}'\mathbf{e}}{n-m-1},\ \mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta} \quad \ldots \quad (3)$$

tidak dapat ditentukan. Multikolinearitas hampir sempurna, yang berarti bahwa antar variabel regresornya terdapat hubungan linear yang hampir sempurna, $\sum_{i=1}^{m} a_i x_i = \mu_i$, μ_i adalah galat stokastik, yang mengakibatkan det (X'X) ≈ 0 dan matriks (X'X) hampir singular.

Menurut Norman R.D. (1981) kondisi tersebut dikatakan *ill conditioned*. Namun demikian matriks (X'X)$^{-1}$ ada dan taksiran dari β masih dapat dicari, tetapi variansnya semakin besar sejalan dengan bertambahnya tingkat multikolinearitas. Jadi interval konfidensi untuk parameter regresi semakin melebar, dan menghasilkan inferensi yang buruk, sehingga perlu dilakukan pendeteksian ada tidaknya multikolinearitas sedini mungkin. Sudah banyak referensi mengenai indikator multikolinearitas, dari yang klasik (bilangan kondisi) dikemukakan oleh Belsley dkk (1980), yang baru indikator Red (Kovacs dkk, 2005), sampai yang terbaru yaitu indikator DEF yang dikemukakan oleh Curto dkk (2007). Dalam makalah ini penulis ingin membandingkan ketiga indikator multikolinearitas tersebut.
Dengan melakukan simulasi untuk membangkitkan data yang dapat diatur berdasarkan banyaknya variabel regresor (m), banyaknya observasi (n), dan tingkat korelasi antar variabel (noise) untuk menghitung nilai – nilai indikator multikolinearitas secara keseluruhan, serta proporsi kemunculan multikolinearitasnya. Replikasi 1000 kali.

2. Indikator keseluruhan Multikolinearitas

Indikator keseluruhan multikolinearitas, yaitu indikator yang memperhitungkan semua regresor secara serempak, dalam ekononetrika indikator multikoinearitas klasik adalah bilangan kondisi (Belsley dkk, 1980), yang baru adalah indikator *Red* (Kovacs dkk., 2005) dan yang terbaru adalah indikator *DEF* (Curto dkk., 2007)

2.1 Bilangan kondisi (Belsley dkk., 1980)

Bilangan kondisi (κ) didefinisikan sebagai rasio nilai eigen maksimum dengan nilai eigen minimum dari matriks (X'X) sebagai berikut:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \quad . \quad . \quad . \quad (4)$$

Dengan $\lambda_i > 0$, *i = 1,2,...,m* adalah nilai eigen dari matriks (X'X). Pada umumnya jika bilangan kondisi lebih kecil dari 100 maka masalah multikolinearitas dianggap tidak begitu serius.
Multikolinearitas dianggap sedang/moderat sampai kuat, jika $100 \leq \kappa \leq 1000$. Multikolinearitas sangat kuat jika $\kappa > 1000$, dan dalam kondisi ini masalah multikolinearitas dianggap berbahaya (Montgomery & Peck, 1992).

2.2 Indikator Red (Kovacs dkk, 2005)

Indikator Red adalah merupakan indikator multikolinearitas baru yang mengkuantifikasikan tingkat keberlebihan (redundancy) himpunan data (database).
Indikator *Red* didefinisikan sebagai berikut:

$$\text{Red} = \sqrt{\frac{tr\left(\mathbf{R}^2 - \mathbf{I}\right)}{m(m-1)}} = \sqrt{\frac{tr\left[(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X}) - \mathbf{I}\right]}{m(m-1)}} \quad . \quad . \quad . \quad (5)$$

Dalam hal ini *tr*(**R**) adalah terace dari matrik **R,** dan **R** adalah matriksi korelasi dengan $\mathbf{R} = \mathbf{X}'\mathbf{X}$ untuk variabel yang dibakukan. Indikator yang diperoleh dengan cara ini akan digunakan untuk mengkuantifikasikan tingkat keberlebihan (*redundancy*). Indikator *Red* mengukur keberlebihan banyak variabel dalam data. Untuk dua database atau lebih database dengan ukuran berbeda dibandingkan, indikator Red hanya dapat digunakan untuk menentukan seberapa berlebihan masing-masing database tersebut. Dalam kasus tidak berlebihan Red = 0 atau 0 %, untuk kasus berlebihan maksimum Red = 1 atau 100 %.
Indikator Red ini menggambarkan korelasi secara lebih tepat, baik secara kualitas maupun besarnya. Berbagai kasus multikolinearitas yang ekstrim juga dapat ditemukan dengan menggunakan indikator ini. Nilainya akan sangat besar apabila semua elemen dari matriks korelasi sama dengan satu.

2.3 *Direct Effects Factor* (DEF)

Curto.J.D. & Pinto J.C (2007), menemukan indikator multikolinearitas baru yang didasarkan pada koefisien jalur, yaitu Faktor Pengaruh Langsung (*Direct Effects Factor*) yang untuk selanjutnya ditulis dengan *DEF*. Koefisien jalur diekspresikan dalam salah satu dari dua metrik, metrik pertama untuk yang tak terbakukan dengan skala menggunakan skala pengukuran dari variabel semula. Metrik kedua untuk yang terbakukan, sebagaimana biasa analisis jalur merupakan perluasan model regresi, variabel yang dibakukan dengan rata-rata 0 dan varians 1. Misalkan dari model regresi linear multipel yang memenuhi persamaan (1), diperoleh model regresi linear yang terbakukan adalah sebagai berikut:

$$\frac{Y-\mu_Y}{\sqrt{\sigma_{YY}}} = \beta_1 \frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{YY}}}\left(\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}}\right) + \ldots + \beta_m \frac{\sqrt{\sigma_{mm}}}{\sqrt{\sigma_{YY}}}\left(\frac{X_m-\mu_m}{\sqrt{\sigma_{mm}}}\right) + \frac{\sqrt{\sigma_{\in\in}}}{\sqrt{\sigma_{YY}}}\left(\frac{\in}{\sqrt{\sigma_{\in\in}}}\right) \ldots \quad (6)$$

Atau dapat dituliskan sebagai berikut:

$$Y_s = p_{y1}Z_1 + p_{y2}Z_2 + \ldots + p_{ym}Z_m + p_{y\varepsilon}\varepsilon_s \qquad \ldots \qquad (7)$$

Dengan Y variabel respon dengan rata-rata μ_y dan varians σ_{yy}, X_j menyatakan variabel regresor j *(j=1,…,m)* dengan rata-rata μ_j dan varians σ_{jj}, ε adalah gangguan dengan rata-rata 0 dan varians $\sigma_{\varepsilon\varepsilon}$, β_j koefisien variabel regresor ke j yang tak dibakukan, koefisien jalur $p_{yj} = \beta_{yj}\frac{\sqrt{\sigma_{jj}}}{\sqrt{\sigma_{yy}}}$ adalah koefisien regresi yang dibakukan (betas), Y_s dan Z_j merupakan variabel – variabel yang dibakukan.

Taksiran koefisien jalur memungkinkan untuk melihat pengaruh langsung maupun tak langsung dari satu variabel terhadap variabel lainnya.

Berdasarkan model (12) korelasi antara Y_s dan setiap Z_j dapat diuraikan sebagai berikut:

$$p_{yj} = Corr(Y_s, Z_j) = Cov\left(\sum_{i=1}^{m} p_{yi}Zi, Z_j\right) = \sum_{i=1-1}^{m} p_{yi}\rho_{ij} \quad ; \text{untuk } j = 1,2,\ldots,m, \quad \ldots \quad (8)$$

Dalam hal ini ρ_{ij} adalah koefisien korelasi antara variabel regresor i dengan j.

Varians Y_s dapat dipecah kedalam tiga bagian (Johnson & Wichern, 1992), sehingga diperoleh persamaan berikut:

$$Var(Y_s) = 1 = Var\left(\sum_{j=1}^{m} p_{yi}Zi + p_{y\varepsilon}\varepsilon_s\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} p_{yi}\rho_{ij}p_{yj} + p_{y\varepsilon}^2 \quad \ldots \quad (9)$$

$$= \sum_{j=1}^{m} p_{yi}^2 + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} p_{Yi}\rho_{ij}p_{Yj} + p_{y\varepsilon}^2$$

Dengan asumsi gangguan tidak berkorelasi dengan Z_j, dan $\sum_{i=1-1}^{m} p_{yi}^2$ adalah proporsi varians yang diberikan langsung oleh koefisien jalur, $2\sum_{i=1}^{m}\sum_{i=1-1}^{m} p_{yi}\rho_{ij}p_{yj}$ adalah proporsi varians yang disebabkan oleh interkorelasi diantara variabel regresor, dan $p_{y\varepsilon}^2$ adalah proporsi varians yang disebabkan oleh gangguan.

Vektor korelasi di antara variabel respon dan variabel regresor $\rho_{yz} = \left[\rho_{y1}\ \rho_{y2}\ldots\rho_{ym}\right]'$, matriks korelasi variabel regresor yang berorde m x m adalah $\Omega_{zz} = \{\rho_{ij}\}$, dan vektor koefisien jalur $p_{yz} = \left[p_{y1}\ p_{y2}\ldots p_{ym}\right]'$.

Jadi

$$\rho_{yz} = \Omega_{zz} p_y \text{ dan } p_{yz} = \Omega_{zz}^{-1} \rho_{yz} \quad . \ . \ . \quad (10)$$

Sehingga akan diperoleh

$$p_{y\varepsilon}{}^2 = 1 - \rho_{yz}{}^2 \quad . \ . \ . \quad (11)$$

Kuadrat galat koefisien jalur $p_{y\varepsilon}{}^2$ ditaksir oleh 1- R^2 (Dillon & Goldstein,1984) dengan R^2 adalah koefisien determinasi dari regresi Y_s pada semua regresornya, R^2 taksiran dari $\rho_{yz}{}^2$.

Berdasarkan hal tersebut dan sesuai dengan pembakuan variabel, dapat disimpulkan bahwa koefisien determinasi adalah jumlah pengaruh langsung dan tak langsung;

$$\rho_{yz}^2 = \sum_{i=1}^{m} p_{yi}^2 + \sum_{i=1}^{m} \sum_{j=i+1}^{m} p_{yi} \rho_{ij} p_{yj} \quad . \ . \ . \quad (12)$$

Kaitan antara $\rho_{yz}{}^2$ dengan pengaruh langsung / taklangsung dapat digunakan untuk menentukan Faktor Pengaruh Langsung / *Direct Effects Factor* (*DEF*) sebagai ukuran korelasi baru diantara variabel regresor dalam model regresi linear multipel.

Direct Effects Factor (*DEF*) didefinisikan sebagai berikut:

$$DEF = \frac{2\sum_{i=1}^{m} \sum_{j=i+1}^{m} p_{Yi} \rho_{ij} p_{Yj}}{\sum_{i=1}^{m} p_{Yi}^2 + 2\sum_{i=1}^{m} \sum_{j=i+1}^{m} p_{Yi} \rho_{ij} p_{Yj}} \quad . \ . \ . \quad (13)$$

$$= \frac{\sum_{i=1}^{m} \sum_{j=1(j \neq i)}^{m} p_{Yi} \rho_{ij} p_{Yj}}{\sum_{i=1}^{m} p_{Yi}^2 + \sum_{i=1}^{m} \sum_{j=1(i \neq j)}^{m} p_{Yi} \rho_{ij} p_{Yj}}$$

Ukuran ini bervariasi dari 0 sampai 1, yang membandingkan pengaruh langsung variabel regresor pada variabel respon dengan pengaruh tak langsung yang dihasilkan dari interkorelasi diantara variabel-variabel regresor. Jika pengaruh langsung semua variabel regresor kecil dibandingkan pengaruh taklangsung maka Indeks *DEF* mendekati 1 dan variabel regresor terkorelasi secara kuat. Jika proporsi variasi yang disebabkan interkorelasi diantara variabel regresor mendekati nol, maka indeks *DEF* mendekati nol dan korelasi diantara variabel regresor lemah. Jadi masalah multikolinearitas terjadi ketika nilai *DEF* mendekati satu.

3. **Simulasi**

Untuk membandingkan ketiga Indikator multikolinearitas keseluruhan, dilakukan simulasi dengan menggunakan program Statistika, adapun tujuan dan penggunaan simulasi ini adalah membangkitkan data yang dapat diatur berdasarkan, banyaknya variabel regresor (m), banyaknya observasi (n), tingkat korelasi antar variabel regresor (noise), dengan replikasi 1000 kali. Untuk membangkitkan data yang mengandung multikolinearitas penulis menggunakan algoritma Norliza dkk (2006).

Algoritma simulasinya sebagai berikut;

3.1 Bangkitkan data sebanyak *n* observasi x *m* var regresor dengan rumus sebagai berikut:

$x_1 = N(0,1)$

$x_p = N(0,'noise') + x_{1,}$ untuk setiap nilai *p, p=2,3,...,m*

$Y = x_1 + x_2 + \ldots + x_m$ + random

3.2 Jalankan regresi (*OLS*) dengan komputasi tambahan untuk menghitung indikator-indikator multikolinearitas.

a. Ekstrak data ke dalam matriks **X** dan **Y.**

b. Hitung B, koefisien regresi taksiran dengan rumus $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

c. Hitung STD, standard deviasi dari masing-masing variabel dalam data.

d. Hitung BETA, koefisien jalur /Betas koefisien regresi yang sudah distandarisasi dengan rumus BETA[i] = B[i] * STD[i] / STD[y]

e. Hitung *COREL*, korelasi dari variabel regresor,

f. Hitung *Direct Effect*, dari matriks *COREL* dan BETA.

g. Hitung *DEF = Indirect Effect / Total Effect* atau

$$DEF = \frac{2\sum_{i=1}^{m}\sum_{j=i+1}^{m} p_{Yi}\rho_{ij}p_{Yj}}{\sum_{i=1}^{m} p_{Yi}^2 + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} p_{Yi}\rho_{ij}p_{Yj}}$$

h. Hitung *EVAL*, eigen value dari matriks *COREL*, cari juga maximum dan minimumnya.

i. Hitung Bilangan Kondisi/ C*ondition Number* dengan rumus

$$K = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

j. Hitung Red*,* indikator *RED* dengan rumus

$$\mathrm{Red} = \sqrt{\frac{tr\left(\mathrm{R}^2 - \mathrm{I}\right)}{m(m-1)}} = \sqrt{\frac{tr\left[(\mathrm{X}'\mathrm{X})(\mathrm{X}'\mathrm{X}) - \mathrm{I}\right]}{m(m-1)}}$$

3.3 Dengan replikasi 1000 kali untuk melihat perbedaan nilai indikator-indikator multikolinearitas secara keseluruhan, *RED, DEF* dan Bilangan Kondisi (κ), dilakukan simulasi komparasi, definisikan banyaknya variabel *m* dan banyaknya observasi *n* dalam dua buah array, *Cases* dan *Variables*, lalu lakukan perhitungan (1) dan (2) sebanyak *m x n.* Kriteria pengklasifikasiannya sebagai berikut, klasifikasi untuk *RED* $\geq$0,5 , yang berarti ada multikolinearitas diberikan nilai *1*, lainnya *0*, untuk DEF $\geq$ 0,5 yang berati ada multikolinearitas diberikan nilai *1*, lainnya *0*, untuk $\kappa \geq 100$ diberikan 1 dan lainya *0.*

4**. Hasil simulasi**.

Proporsi kemunculan /terjadinya multikolinearitas dari indikator-indikator multikolinearitas keseluruhan dengan n = 30, m = 3, replikasi 1000 kali adalah sebagai berikut,

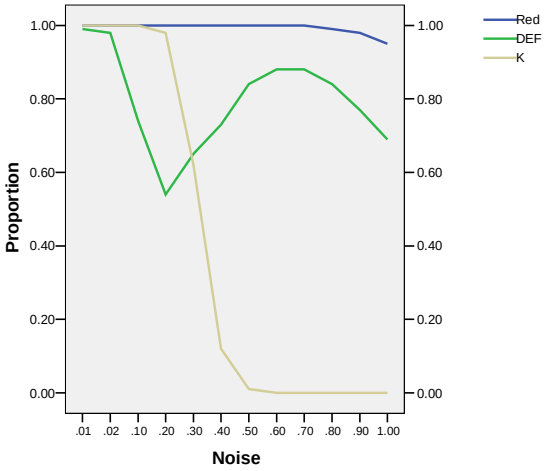
Tabel 4.1 Nilai-nilai indikator multikolinearitas keseluruhan

Noise	0.01	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Red	1	1	1	1	1	1	1	1	1
DEF	0.99	0.98	0.74	0.54	0.65	0.73	0.84	0.88	0.88
K	1	1	1	0.98	0.62	0.12	0.01	0	0

Tabel Lanjutan

Noise	0,8	0,9	1
Red	0,99	0,98	0,95
DEF	0,84	0,77	0,69
K	0	0	0

Dari tabel 4.1 diperoleh grafik fungsi indikator *Red, DEF* dan bilangan kondisi sebagai berikut,



Gambar 4.1 Fungsi Red, DEF,K

Dari grafik dapat disimpulkan bahwa,indikator Red lebih sensitif dari pada indikator lainnya dan makin kecil nilai noise, makin makin tinggi derajat multikolinearitas.

Daftar Pustaka

1. Belsley David A dkk *(1980*), *Regression Diagnostics Identifying Influential Data and Sources of Collinearity,* John Wiley & son, New York.
2. Curto J.D, Pinto J.C, (*2007*), New Multicollinearity Indicators in Linear Regression Models , *Int'l Statistical Rev. 75,1,114-121 , Lisboa, Portugal.*
3. Dillon W.R., (1984), *Multivariate Analysis Methods and Applications*, John Wiley & Sons, Inc., Canada.
4. Greene William H., (*1990)*, *Econometric Analysis,* Prentice Hall, Englewood Cliffs, New Yersey.
5. Gujarati Damodar N.*(2003*), *Basic Econometrics,* Prentice Hall inc. Singapore.
6. Johnson, Wichhern, (*1982)*, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
7. Kovacs P., Petres T.and Toth L.(*2005)*, A New Measure of Multicollinearity in Linear Regression Models, *Int'l Statistical Rev. ,73,3,405-412, printed in Wales by Cambrian Printers , Budape, Hungary.*
8. Law A.M., Kelton W.D., *Simulation Modeling And Analysis*, McGraw-Hill, Inc., New York.
9. Mishra, (2004), An Algorithm To Generate Variates With Desired Intercorrelation Matrix, *Dept. of Economics NEHU, Shillong, India.*
10. Montgomery, D.C, (*1992*), *Introduction To Linear Regression Analysis*, John Wiley & Sons, Inc. New York.
11. Norliza Adnan, dkk (2006), A Comparative Study On Some methods For Handling Multicollinearity Problems*, Department of Mathematic UTM Skundai, Johor, Malaysia.*

12. Rice J., (1995), *Mathematical Statistics and Data Analysis,* An Imprint of Wadsworth Publishing Company, Belmont, California.
13. Sembiring R. K., *(2003), Analisis Regresi* , Penerbit ITB Bandung.