Peranan dari Pemilihan Level sebagai Referensi pada Variabel Bebas Bertipe Kategori terhadap Derajat Multikolinieritas dalam Model Regresi Linier

[1]Seny Mustikawati, [2]Anneke Iswani A., [3]Abdul Kudus
[1,2,3]Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Jl. Tamansari No. 1 Bandung 40116
e-mail : [1]senymustika@ymail.com, [2]annekeiswani11@gmail.com, [3]Akudus69@yahoo.com

Abstrak. Penggunaan variabel bebas bertipe kategori dalam model regresi linier cenderung akan menimbulkan masalah multikolinieritas. Multikolinieritas juga merupakan salah satu faktor yang menyebabkan galat baku menjadi besar sehingga menyebabkan selang kepercayaan untuk parameter juga cenderung akan lebih lebar. Statistik yang bisa digunakan untuk mendeteksi adanya indikasi multikolinieritas salah satunya adalah bilangan kondisi. Nilai bilangan kondisi bergantung pada pemilihan level kategori yang dijadikan sebagai referensi dari variabel bebas bertipe kategori. Variabel bebas tersebut dalam praktiknya diwakili oleh variabel *dummy*. Pengubahan level kategori sebagai referensi ini kesemuanya ada 96 skenario. Hasil perhitungan untuk skenario 1 diperoleh nilai eigen 6.344, 2.039, 1.309, 1.013, 1.012, 0.845, 0.669, 0.386, 0.142, 0.085, 0.073, 0.039, 0.029, dan 0.016. Sehingga dari perhitungan diperoleh nilai bilangan kondisi sebesar 20.128, artinya bilangan kondisi ini berada diantara nilai 10-30 yang berarti adanya masalah multikolinieritas yang medium. Adapun nilai bilangan kondisi terbesar yaitu pada skenario ke-78 dengan nilai bilangan kondisi sebesar 51.242 yang nilai tersebut berada >30, hal ini menandakan bahwa masalah multikolinieritas dianggap serius. Sementara nilai bilangan kondisi yang terkecil yaitu pada skenario ke-50 adalah sebesar 17.891 dan nilai tersebut berada diantara nilai 10-30, hal ini menunjukan adanya masalah multikolinieritas yang medium. Berdasarkan hasil analisis menunjukan bahwa untuk derajat multikolinieritas dalam penelitian ini adalah medium/sedang dengan nilai R-Square yang diperoleh tetap yaitu sebesar 0.89.

Kata Kunci : Variabel Bebas Bertipe Kategori, Multikolinieritas, Bilangan Kondisi, Model Regresi Linier.

A. Pendahuluan

Menghadapi kasus variabel yang tidak bersifat kontinu, sebagai gantinya peneliti mempergunakan variabel yang bersifat kategori. Jenis variabel kategori tersebut seringkali menunjukkan keberadaan klasifikasi (kategori) tertentu, dan variabel yang bertipe kategori dapat berperan sebagai variabel bebas (X) yang digunakan dalam persamaan regresi linier. Variabel bebas bertipe kategori tersebut bisa diwakili dengan penggunaan variabel *dummy*. Ketika level kategori pada variabel *dummy* ini dimasukkan ke dalam model analisis regresi linier, maka variabel *dummy* tersebut akan mirip atau mencerminkan nilai yang sama dengan kolom intersepnya sehingga penggunaan variabel *dummy* di dalam model tersebut cenderung akan menimbulkan masalah multikolinieritas. Menurut Sembiring (2003), jika variabel bebas dengan penggunaan variabel dummy terlalu banyak dimasukkan ke dalam model, maka akan menimbulkan adanya masalah multikolinieritas. Hal ini, Wissmann dkk. (2007) mengajukan pengujian untuk melihat efek dari pemilihan variabel bertipe kategori dalam mengurangi derajat multikolinieritas dengan menggunakan statistik bilangan kondisi atau nilai *eigen* X'X di dalam model regresi linier, dimana pengujian ini memiliki kelebihan bahwa multikolinearitas yang dengan variabel kategori dapat dikurangi dengan memilih kategori yang benar. Pada umumnya, bilangan kondisi merupakan indikator untuk masalah multikolinieritas secara keseluruhan sedangkan untuk mengetahui secara individu diperiksa dengan menggunakan *indeks kondisi*.

Menurut Wissmann dkk (2007) jika bilangan kondisi (κ) ini lebih kecil dari 10, maka tidak ada masalah multikolinieritas. Sedangkan, jika (κ) antara 10 sampai 30 menunjukan adanya masalah multikolinieritas yang medium, akan tetapi jika nilai (κ) lebih besar daripada 30, maka masalah multikolinieritas dianggap serius. Sementara, untuk *indeks kondisi* yang lebih besar daripada 30 dan proporsi varians yang lebih besar dari 0,5, maka kita harus mewaspadai adanya indikasi masalah multikolinieritas (Montgomery & Peck, 1992).

Penerapan pengujian ini dapat dilihat dalam kasus yang diambil dari skripsi Suci Rachmadini (2010) dengan judul "Membandingkan Model Regresi Pengeluaran Rumah Tangga Penduduk Komplek Dan Non Komplek Di Kelurahan Kujang Sari Dengan *Dummy Variable*". Tujuan penelitian Suci (2010) adalah menentukan model regresi terbaik untuk model regresi pengeluaran rumah tangga di Kelurahan Kujang Sari Kecamatan Bandung Kidul dengan memperhatikan pendapatan rumah tangga, jumlah anggota rumah tangga, dan umur kepala rumah tangga, jenis kelamin kepala rumah tangga, pekerjaan kepala rumah tangga, kepala rumah tangga yang berada di komplek dan non komplek dengan tingkat pendidikan kepala rumah tangga.

Tujuan dari pembuatan makalah ini adalah untuk mengetahui prosedur pemeriksaan adanya indikasi masalah multikolinieritas dengan menggunakan statistik bilangan kondisi. Dan untuk mengetahui pengaruh dari level kategori yang dijadikan sebagai referensi dari variabel bebas bertipe kategori terhadap derajat multikolinieritas.

B. Landasan Teori

1. Analisis Regresi Linier Berganda

Bentuk umum model regresi linier berganda dengan p variabel bebas adalah :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_P X_{Pi} + \varepsilon_i \quad ; \text{i} = 1, 2, \ldots \text{N} \qquad (2.1)$$

Dan model regresi linier berganda di atas dapat ditaksir berdasarkan sebuah sampel acak yang berukuran n dengan model sebagai berikut :

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_p X_{pi} \quad ; \text{i} = 1, 2, \ldots \text{n} \qquad (2.2)$$

Asumsi-asumsi pada regresi linier diantaranya adalah galat mengikuti distribusi normal, tidak terjadi heteroskedastis(terjadi homoskedastis), tidak ada otokorelasi diantara keasalahan pengganggu, dan tidak terjadi multikolinieritas.

Multikolinearitas pertama kali ditemukan oleh Frisch (1934) yang berarti adanya hubungan liniear yang "sempurna" atau pasti diantara beberapa atau semua variabel bebas dari model regresi berganda. Menurut Sumodiningrat (1994:282-283), masalah multikolinieritas bisa timbul karena: adanya sifat-sifat yang terkandung dalam kebanyakan variabel-variabel ekonomi yang berubah bersama-sama sepanjang waktu dan variabel-variabel tersebut dipengaruhi oleh faktor-faktor yang sama. Dan penggunaan Lag, sehingga terbentuk model terdistribusi lag (*distributed lag*). Multikolinearitas diperkirakan akan muncul dalam kebanyakan hubungan-hubungan ekonomi. Lebih sering muncul dalam data deret waktu dan bisa pula muncul dalam data *cross sectional*.

Prosedur penanggulangan efek multikolinieritas yang sering terjadi sangat tergantung sekali pada kondisi penelitian, misalnya prosedur penggunaan informasi apriori sangat tergantung dari ada atau tidaknya dasar teori (literatur) yang sangat kuat untuk mendukung hubungan matematis antara variabel bebas yang saling berkolinear, prosedur mengeluarkan variabel bebas walaupun

seringkali membuat banyak peneliti keberatan karena prosedur ini akan mengurangi obyek penelitian yang diangkat, prosedur lainya seperti menghubungkan data *cross sectional* dan *time series*, prosedur *first difference* dan penambahan data baru untuk memberikan efek penanggulangan yang kecil pada masalah multikolinearitas.

Gejala multikolinieritas dapat di diagnosis salah satunya dengan bilangan kondisi dimana **X** adalah matriks desain regresi berorde (*n x p*), dimana λ_j, (j= 1,2...p) adalah nilai eigen dari matriks **X'X**. Sehingga, *bilangan kondisi* dari matriks **X** adalah :

$$\kappa(X) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \quad \ldots\ldots\ldots$$

dengan :

λ_{max} = nilai eigen terbesar

λ_{min} = nilai eigen terkecil

Dan *Indeks Kondisi* dari matriks **X'X** adalah :

$$\eta_j = \frac{\mu_{max}}{\mu_j} \;\; ; \; \mu_j, \text{ (j=1, 2 ... p)}$$

karena $\mu = \sqrt{\lambda}$, maka $\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$

dimana λ_j, (j=1, 2 ... p)

Nilai tebesar untuk $\eta_j$adalah bilangan kondisi dari matriks X. Dan untuk menentukan nilai eigen dapat menyelesaikan persamaan:

$$[\lambda I - A]X = 0$$

Persamaan (2.5) terpenuhi jika dan hanya jika:

$$\det [\lambda I - A] = 0$$

2. Variabel *Dummy*

Variabel dalam persamaan regresi yang sifatnya kategori biasanya menunjukkan ada tidaknya suatu "quality" atau "atribute', misalnya laki-laki atau perempuan, sarjana atau bukan, dan seterusnya. Salah satu metode untuk mengkuantitatifkan atribut yang bersifat kualitatif tersebut adalah dengan cara membentuk variabel yang sifatnya *artificial* (*dummy*) ke dalam model persamaan regresi dengan mengambil nilai 1 (satu) yang menunjukan adanya atribut atau 0 (nol) yang menunjukan tidak ada atribut.

Maka, model persamaan regresinya dengan variabel dummy adalah :

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + \varepsilon_i \qquad (2.7)$$

Dengan mengamati adanya indikasi masalah multikolinieritas yang kuat, jika hanya 4 persen dari pengamatan dalam kategori referensi, ini menunjukkan bahwa masalah multikolinearitas dapat dikurangi dengan memilih kategori referensi yang berbeda dari variabel dummy. Seperti ketika kategori referensi berubah, maka 96 persen dari pengamatan terletak pada kategori referensi. Hal ini jelas bahwa dengan mengubah kategori referensi, kesalahan standar koefisien regresi juga tidak berubah.

Oleh karena itu, pengkodean variabel dummy dan pilihan kategori referensi mempengaruhi stabilitas numerik desain matriks serta varians dari intersep. Ini adalah fungsi dari proporsi "nol'. Hal ini, derajat multikolinearitas meningkat karena kategori referensi yang dipilih. Nilai bilangan kondisi bergantung pada pemilihan level kategori yang dijadikan sebagai referensi dari

variabel bebas bertipe kategori.

C. Metode

Banyaknya data pengamatan yaitu berjumlah 138 rumah tangga. Dalam hal ini, yang menjadi variabel respon adalah pengeluaran rumah tangga (Y), dan variabel bebas yaitu pendapatan rumah tangga (X_1), jumlah anggota rumah tangga (X_2), umur kepala rumah tangga (X_3), jenis kelamin kepala rumah tangga (X_4) dengan kategori 1=laki-laki dan 2=perempuan, pekerjaan kepala rumah tangga (X_5) dengan kategori 1=pegawai pabrik, 2=wiraswasta, 3=buruh, 4=PNS, tempat tinggal (X_6) dengan kategori 1=non-komplek dan 2=komplek, dan pendidikan terakhir kepala rumah tangga (X_7) dengan kategori 1=tamat SD, 2=tamat SLTP, 3=tamat SMA, 4=D3, 5=S1 dan 6=S2.

Langkah-langkah dalam pemilihan level kategori pada variabel bebas bertipe kategori dan menghitung bilangan kondisi dapat dilakukan dengan beberapa langkah. Langkah yang digunakan adalah:

1. Variabel bebas yang berbentuk kategori, diubah menjadi variabel dummy, dengan menetapkan level kategori pertama sebagai referensi.
2. Menentukan taksiran model regresi linier berganda dengan metode kuadrat terkecil.
   a. Pemeriksaan Multikolinieritas
   b. Menghitung matriks A = (**X'X**) dan mencari vektor eigen dengan menghitung determinan dari $[\lambda I - A] = 0$.
   c. Menghitung Nilai Eigen dari setiap fungsi persamaan.
   d. Menghitung Bilangan Kondisi dan Proporsi Varians.
3. Mengulangi langkah 2 dan langkah 3 dengan terlebih dahulu mengubah referensi menjadi kategori selanjutnya (ke-2, ke-3, dan seterusnya).

Sebelum membandingkan dan memberikan kesimpulan, terlebih dahulu ditentukan prosedur pengubahan level yang dijadikan sebagai referensi. Diketahui dalam penelitian ini, ada 4 variabel bebas kategori yang terdiri dari:

1. Jenis Kelamin terdiri dari 2 level kategori. Dengan kategori pertama jenis kelamin laki-laki, dan kedua jenis kelamin perempuan.
2. Pekerjaan terdiri dari 4 level kategori. Dengan kategori pertama pegawai pabrik, kategori kedua wiraswasta, kategori ketiga buruh dan kategori keempat adalah PNS.
3. Tempat Tinggal terdiri dari 2 level kategori. Dengan kategori pertama adalah non-komplek dan kategori kedua adalah komplek
4. Pendidikan Terakhir terdiri dari 6 level kategori. Dengan kategori pertama tamat SD, kategori kedua tamat SMP, kategori ketiga tamat SMA, kategori keempat Diploma, kategori kelima S1, dan kategori keenam adalah S2.

Dengan demikian, maka akan terdapat 2 x 4 x 2 x 6 = 96 skenario penentuan referensi dari variabel-variabel bebas bertipe kategori tersebut sebagaimana dijelaskan pada bagian berikut:

Skenario 1 dengan penyusunan variabel dummy:

Variabel Jenis Kelamin diubah menjadi 1 variabel dummy dengan menetapkan kategori laki-laki sebagai referensi.

JK	D1
1	0
2	1

Variabel Pekerjaan diubah menjadi 3 variabel dummy dengan menetapkan kategori Pegawai Pabrik sebagai referensi.

P	D2	D3	D4
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Variabel Tempat Tinggal diubah menjadi 1 variabel dummy dengan menetapkan kategori non-komplek sebagai referensi.

TT	D5
1	0
2	1

Variabel Pendidikan Terakhir diubah menjadi 5 variabel dummy dengan menetapkan kategori Tamat SD sebagai referensi.

PT	D6	D7	D8	D9	D10
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1

Dan terdapat 96 skenario pengubahan level kategori sebagai referensi sebagai berikut:

Skenario	**Jenis Kelamin**	**Pekerjaan**	**Tempat Tinggal**	**Pendidikan Terakhir**
2	Laki-Laki	Pegawai Pabrik	Non-Komplek	Tamat SMP
3	Laki-Laki	Pegawai Pabrik	Non-Komplek	Tamat SMA
4	Laki-Laki	Pegawai Pabrik	Non-Komplek	Diploma3
5	Laki-Laki	Pegawai Pabrik	Non-Komplek	S1
6	Laki-Laki	Pegawai Pabrik	Non-Komplek	S2
7	Laki-Laki	Pegawai Pabrik	Komplek	Tamat SD
8	Laki-Laki	Pegawai Pabrik	Komplek	Tamat SMP
9	Laki-Laki	Pegawai Pabrik	Komplek	Tamat SMA
10	Laki-Laki	Pegawai Pabrik	Komplek	Diploma3
⋮	⋮	⋮	⋮	⋮
50	Perempuan	Pegawai Pabrik	Non-Komplek	Tamat SMP
⋮	⋮	⋮	⋮	⋮
78	Perempuan	Buruh	Non-Komplek	S2
⋮	⋮	⋮	⋮	⋮
96	Perempuan	PNS	Komplek	S2

5. Bandingkan hasil pemeriksaan multikolinieritas untuk setiap kategori yang dijadikan referensi.

D. Kesimpulan

Berdasarkan hasil penelitian dapat diambil kesimpulan sebagai berikut :

1. Hasil perhitungan untuk skenario pertama diperoleh nilai eigen 2.039, 1.309, 1.013, 1.012, 0.845, 0.669, 0.386, 0.142, 0.085, 0.073, 0.039, 0.029, dan 0.016. Sehingga dari perhitungan diperoleh nilai bilangan kondisi sebesar 20.128, artinya bilangan kondisi ini berada diantara nilai 10-30 yang berarti adanya masalah multikolinieritas yang medium. Dan perubahan pada level kategori sebagai referensi ini kesemuanya ada 96. Adapun nilai bilangan kondisi terbesar yaitu pada skenario ke 78 dengan nilai bilangan kondisi sebesar 51.242 yang nilai tersebut berada >30, hal ini menandakan bahwa masalah multikolinieritas dianggap serius. Sementara nilai bilangan kondisi yang terkecil yaitu pada skenario ke 50 adalah sebesar 17.891 dan nilai tersebut berada di antara nilai 10-30, hal ini menunjukan adanya masalah multikolinieritas yang medium. Dengan nilai R-Square yang diperoleh adalah tetap yaitu sebesar 0.890, artinya kontribusi dari tiga variabel bebas dan sepuluh variabel dummy terhadap variabel terikat sebesar 89% atau 89% variabel pengeluaran rumah tangga dapat dijelaskan oleh pendapatan rumah tangga, jumlah anggota rumah tangga, umur kepala rumah tangga, jenis kelamin, pekerjaan, tempat tinggal dan pendidikan terakhir. Sedangkan sisanya yaitu sebesar 11% dijelaskan oleh faktor lain diluar model.
2. Tujuan penelitian ini yaitu ingin mengetahui prosedur pemeriksaan adanya indikasi masalah multikolinieritas dengan menggunakan statistik bilangan kondisi, diketahui bahwa masalah multikolinieritas dalam kasus ini nilai kondisinya mengalami perubahan apabila referensinya berubah.
3. Untuk pengaruh dari level kategori yang dijadikan sebagai referensi dari variabel dummy pada kasus ini tentunya mempunyai peranan bahwa dengan adanya perubahan level kategori sebagai referensi menimbulkan peningkatan dan penurunan dalam derajat multikolinieritas.

Daftar Pustaka

Montgomery, D. C., and Peck, E.A. 1992. *Introduction to Linear Regression Analysis. Second Edition.* New York: John Wiley & Sons.

Rachmadini, Suci. 2010. Membandingkan Model Regresi Pengeluaran Rumah Tangga Penduduk Komplek Dan Non Komplek Kelurahan Kujang Sari Dengan *Dummy Variable*. Bandung: Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam(MIPA), Universitas Islam Bandung.

Sembiring, R.K. 2003. Analisis Regresi. Edisi Kedua. ITB, Bandung.

Sumodiningrat, G. 1994. "*Ekonometrik Pengantar* ". BPFE Yogyakarta. Badan Penerbit Ekonomi Indonesia, Jakarta.

Wissmann, M, Toutenburg, H, and Shalabh. 2007. *Role of Categorical Variables in Multicollinearity in the Linear Regression Model*. http://www.stat.uni-muenchen.de.