 http://dx.doi.org/10.29313/.v7i1.25565

Mendeteksi dan Mengatasi Multikolinieritas pada Data Penelitian Diabetes Melitus Wanita Suku Indian Tahun 2018

Mohamad Ridwan*, Siti Sunendiari

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*mohamadridwan2218@gmail.com, diarisunen22@gmail.com

Abstract. Fulfillment of the multicollinearity assumption in logistic regression is very important. If the assumption of multicollinearity is not met, the regression model is no longer efficient because the standard error value of the regression coefficient becomes very large (overestimate). This indicates that the parameters in the model cannot be estimated and the output in the form of a path diagram cannot be displayed. It is also possible that if the parameters are successfully estimated and the path diagram output is successfully displayed, the results are biased. In this study, we will try to detect and overcome the multicollinearity problem by looking at the correlation value and the VIF value and overcoming it by eliminating the independent variable that has a VIF value > 10. The data used in this study is secondary data from the Indian Women's Diabetes Research 2018. The dependent variable has two categories, namely not suffering from and suffering from. From the research it can be concluded that there are two variables detected that multicollinearity occurs.

Keywords: Correlations, Logistic Regression, Multicollinierity, VIF .

Abstrak. Pemenuhan asumsi multikolinieritas dalam regresi logistik sangat lah penting. Jika asumsi multikolinieritas tidak terpenuhi, model regresi tidak lagi efisien karena nilai standar error koefisien regresi menjadi sangat besar *(overestimate)*. Hal ini mernunjukkan bahwa parameter dalam model tidak dapat diestimasi dan output dalam bentuk diagram jalur tidak dapat ditampilkan. Bisa juga jika parameter berhasil diestimasi dan output diagram jalur berhasil ditampilkan, tetapi hasilnya menjadi bias. Dalam penelitian ini akan dicoba mendeteksi dan mengatasi masalah multikolinieritas dengan cara melihat nilai korelasi dan nilai VIF dan mengatasinya dengan cara menghilangkan variabel independen yang memiliki nilai VIF > 10. Data yang digunakan pada penelitian ini adalah data sekunder Penelitian Diabetes Wanita Indian Tahun 2018. Variabel dependen memiliki dua kategori, yaitu tidak mengidap dan mengidap. Dari penelitian dapat disimpulkan bahwa terdapat dua variabel yang terdeteksi terjadi multikolinieritas.

Kata Kunci: Korelasi, Multikolinieritas, Regresi Logistik, VIF.

1. Pendahuluan

Untuk menganalisis data penelitian diabetes diperlukan suatu metoda analisis yang tepat, dalam penelitian ini akan digunakan regresi logistik. Yang harus diperhatikan dalam regresi logistik adalah multikolinieritas. Mendeteksi multikolinieritas sangat penting, jika dibiarkan akan menyebabkan estimasi yang tidak stabil dan nilai varians yang tidak akurat

sehingga mempengaruhi interval kepercayaan dan uji hipotesis.

Regresi logistik adalah bagian dari analisis regresi yang dapat digunakan jika variabel dependen (Y) merupakan variabel dikotomi. Variabel dikotomi biasanya hanya terdiri atas dua nilai, yang mewakili kemunculan atau tidak adanya suatu kejadian yang biasanya diberi angka 0 atau 1. Sedangkan multikolinearitas adalah suatu kondisi adanya hubungan linier atau korelasi yang tinggi diantara masing-masing variabel bebas dalam sebuah model regresi. Oleh sebab itu dalam regresi logistik harus mendeteksi multikolinieritas terlebih dahulu.

Terdapat beberapa cara mendeteksi multikolinieritas, dalam penelitian ini akan menggunakan nilai dari korelasi dan VIF. Korelasi digunakan untuk mengukur eratnya hubungan antara variabel bebas dengan variabel tak bebas. Korelasi yang digunakan adalah Korelasi Pearson. Cara mendeteksi multikolinieritasnya apabila korelasi kuat atau diatas 0,8 maka ditemukan masalah multikolinieritas. Nilai VIF atau *Variance Inflation Factor* adalah kenaikan varians dari dugaan parameter antar variabel X. Jika nilai VIF lebih dari 10, maka taksiran parameter kurang baik.

Regresi logistik biner adalah bentuk regresi yang digunakan untuk memodelkan hubungan antara variabel dependen dan variabel independen. Ketika variabel adalah sebuah data dengan ukuran biner/dikotomi (misal: ya atau tidak, sukses atau gagal, puas atau tidak puas, bagus atau rusak, mati atau hidup) maka regresi logistik tersebut menggunakan regresi logisitik biner.

Regresi logistik biner mensyaratkan variabel dependen (Y) memiliki 2 kategori Dalam skripsi ini akan mendeteksi dan mengatasi masalah multikolineritas pada data sekunder yang diperoleh dari *National Institute Of Diabetes and Disgeftive and Kidney Diseases.* Data yang berjumlah 88 yang meneliti tentang diabetes mellitus pada Wanita Indian tahun 2018, akan dideteksi multikolinieritasnya dengan cara melihat nilai korelasi dan VIF. Melalui cara tersebut maka masalah multikolinieritas dapat teratasi.

2. Landasan Teori

Regresi Logistik

Regresi logistik adalah suatu metode yang dapat digunakan untuk mencari hubungan antara variabel dependen (Y) yang bersifat dikotomi (skala nominal/ordinal dengan dua kategori) dengan satu atau lebih variabel independen berskala kategori atau kontinu. Model regresi logistik terdiri dari regresi logistik dengan respon biner, ordinal, dan multinomial. (Hosmer dan lemeshow, 2000). Regresi logistik biner adalah suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel dependen (Y) yang bersifat biner (dikotomi) dengan variabel independen (X) yang bersifat kategorik atau kontinu (Hosmer dan lemeshow, 2000). Variabel respon terdiri dari dua kategori seperti sukses dan gagal dengan notasi $y = 1$ jika sukses dan $y = 0$ jika gagal.

Model dari regresi logistik adalah:

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})} \qquad (2.1)$$

Persamaan di atas perlu dilakukan transformasi logit untuk memperoleh fungsi yang linear. Maka didapat model linier sebagai berikut:

$$Logit\ (p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}; -\infty \leq Logit(p_i) \leq \infty \qquad (2.2)$$

Pendugaan parameter dalam regresi logistik dilakukan dengan cara metode maksimum *likelihood*. Metode tersebut menduga koefisien β dengan cara memaksimumkan fungsi *likelihood* dan mensyaratkan bahwa data harus mengikuti suatu distribusi tertentu. Pada regresi logistik biner, setiap pengamatan mengikuti distribusi Bernoulli sehingga dapat ditentukan fungsi *likelihood*nya.

Jika x_i dan y_i adalah pasangan variabel independen dan dependen pada pengamatan ke-i dan diasumsikan bahwa setiap pasangan pengamatan saling bebas dengan pasangan pengamatan lainnya, $i = 1,2, \ldots, n$ maka fungsi peluang untuk setiap pasangan adalah sebagai berikut :

$$f(x_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \quad ; y_i = 0,1 \qquad (2.3)$$

Setiap pasangan pengamatan diasumsikan bebas sehingga fungsi *likelihood*nya merupakan gabungan dari fungsi distribusi masing – masing pasangan yaitu sebagai berikut :

$$l(\beta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad (2.4)$$

Fungsi likelihood dimaksimumkan dalam bentuk ln $l(\beta)$ dan dinyatakan dengan $L(\beta)$.

$$l(\beta) = \ln[p_1^{y_1} \dots p_n^{y_n}(1-p_1)^{1-y_1} \dots (1-p_n)^{1-y_n}] \quad (2.5)$$

Berikut adalah turunan pertama yang disajikan dalam bentuk matriks.

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{in} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{bmatrix} . \begin{bmatrix} y_1 - p_1 \\ y_2 - p_2 \\ \vdots \\ y_k - p_k \end{bmatrix} = X^T(Y - P_l) \quad (2.6)$$

Berikut adalah turunan kedua yang disajikan dalam bentuk matriks.

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{bmatrix} . \begin{bmatrix} p_i(1-p_i) & 0 & \dots & 0 \\ 0 & p_2(1-p_1) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n(1-p_i) \end{bmatrix} . \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{21} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} = \mathbf{X^T V} \quad (2.7)$$

Metode newton raphson digunakan untuk menyelesaikan persamaan aljabar yang tidak linear. Dalam regresi logistik biner diperoleh persamaan yang belum linear, maka dari itu diperlukan metode iterasi newton raphson agar mendapatkan persamaan yang linear.

Berikut langkah-langkah metode iterasi newton raphson:

1. Dipilih taksiran awal untuk $\boldsymbol{\beta}$ misal $\hat{\beta} = 0$
2. Dihitung $\boldsymbol{X^T(Y - p_i)}$ dan $\boldsymbol{X^T V X}$ selanjutnya dihitung invers dari $\boldsymbol{X^T V X}$.
3. Pada setiap $i + 1$ dihitung taksiran baru yaitu $\widehat{\beta_{i+1}} = \widehat{\beta_i} + \{x^T V X\}^{-1}\{X^T(Y - P_i)\}$.
4. Iterasi berakhir jika diperoleh $\widehat{\beta_{i+1}} \cong \widehat{\beta_i}$.

Multikolinieritas

Dalam model persamaan struktural, asumsi secara empiris yang tidak boleh dilanggar adalah multikolinearitas, karena multikolinearitas dapat menyebabkan nilai *Confidence Interval* sangat lebar, sehingga akan menjadi sangat sulit untuk menolak hipotesis nol pada sebuah penelitian. Jika dalam penelitian tersebut terdapat multikolinearitas dapat memberikan efek yang fatal yaitu model menjadi *non identified* yang berarti parameter dalam model tidak dapat diestimasi dan output dalam bentuk diagram jalur tidak dapat ditampilkan atau jika parameter berhasil diestimasi dan *output* diagram jalur berhasil ditampilkan, tetapi hasilnya dapat bias (Wijanto, 2008). Ada beberapa cara untuk mendeteksi multikolinearitas, diantaranya:

1. Nilai korelasi (korelasi antar variabel bebas)

Uji multikolinearitas bertujuan untuk menguji adanya korelasi antar variabel independen pada model. Asumsi multikolinearitas mengharuskan tidak adanya korelasi yang sempurna atau besar diantara variabel-variabel independen. Analisis koefisien korelasi bertujuan untuk mempelajari apakah ada hubungan antara dua variabel. Koefisien korelasi antar variabel independen haruslah lemah (dibawah 0.8). Jika korelasi kuat, terjadilah masalah multikolinearitas. Koefisien korelasi pearson dirumuskan sebagai berikut:

$$r_{ab} = \frac{n\sum ab - n\sum a\sum b}{\sqrt{[n\sum a^2 - (\sum a)^2][n\sum b^2 - (\sum b)^2]}} \quad (2.8)$$

Untuk pengujian multikolinieritas adalah pengujian antara variabel X dengan variabel X lainnya, dimana a adalah inisial untuk variabel X dan b adalah inisial variabel X lainnya pada model, sedangkan *n* adalah banyaknya sampel yang diguna kan (Santoso, 2012).

2. Variance Inflation Factor (VIF) dan Toleransi

Metode untuk menguji adanya multikolinearitas dapat dilihat pada nilai toleransi atau *variance inflation factor* (VIF). Kedua ukuran ini menunjukkan setiap variabel independen

manakah yang dijelaskan oleh variabel independen lainnya. Toleransi mengukur variabilitas variabel independen yang terpilih yang tidak dijelaskan oleh variabel independen lainnya. Nilai toleransi yang rendah sama dengan nilai VIF yang tinggi. Nilai VIF dapat diperoleh dengan rumus berikut:

$$VIF = \frac{1}{Toleransi} \tag{2.9}$$

Dengan rumus toleransi sebagai berikut:

$$Toleransi = 1 - R^2 \tag{2.10}$$

Batas nilai toleransi adalah 0,10 dan nilai VIF adalah 10. Jika VIF > 10 dan nilai toleransi < 0.10, maka tejadi multikolinearitas tinggi antar variabel independen dengan variabel independen lainnya. Jika VIF < 10 dan nilai toleransi > 0.10, maka dapat diartikan tidak terdapat multikolinearitas pada penelitian tersebut. Regresi yang baik memiliki VIF disekitar angka 1 dan mempunyai angka Toleransi mendekati 1 (Santoso, 2012).

Salah satu cara yang bisa dilakukan untuk menanggulangi jika terjadi multikolinearitas adalah dengan menghapus variabel independen yang memiliki nilai VIF > 10. Proses tersebut dapat dilakukan berulang-ulang sampai tidak terdapat multikolinieritas pada variabel independen.

Diabetes

Pengertian dari diabetes lainnya menurut *American Diabetes Association* (ADA) adalah suatu kelompok penyakit metabolic dengan karakteristik hiperglikemia yang terjadi karena kelebihan sekresi insulin, gangguan kerja insulin atau keduanya, yang menimbulkan berbagai komplikasi kronik pada mata, ginjal, saraf dan pembuluh darah. (Hastuti, 2008).

3. Hasil Penelitian dan Pembahasan

Data

Data yang digunakan penulis dalam penelitian ini adalah data sekunder yang diperoleh dari *National Institute Of Diabetes and Disgeftive and Kidney Diseases* pada tahun 2018. Sampel dari penelitian ini adalah wanita berasal dari Suku Indian yang berumur minimal 21 tahun. Variabel independen (X) dalam penelitian ini adalah sebagai berikut:

X1 = Kehamilan
X2 = Kandungan glukosa.
X3 = Tekanan darah.
X4 = Ketebalan kulit.
X5 = Kadar insulin.
X6 = IMT.
X7 = Silsilah diabetes.
X8 = Umur.

Variabel dependent (Y) yaitu terkena atau tidaknya diabetes, Variabel Y merupakan variabel yang kategoreik dengan keterangan 0 = “Tidak Mengidap Diabetes”, dan 1 = “Mengidap Diabetes”.

Tabel 1. Data penelitian diabetes wanita suku indian tahun 2018

Sampel	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	8	164	86	20	123	25.1	0.078	21	0
2	9	165	86	14	148	24.8	0.143	21	0
3	9	165	88	44	187	24.7	0.206	21	0
4	1	92	60	22	139	25	0.253	22	0

5	12	193	100	26	120	32.4	0.433	22	0
.									
88	12	193	100	31	188	30.8	0.493	22	0

Mendeteksi multikolinieritas

Mendeteksi multikolinieritas melalui nilai korelasi adalah dengan cara menghitung korelasi dari setiap pasang variabel independen hingga membentuk matriks korelasi 8 x 8. Setelah didapat nilai korelasinya, apabila terdapat korelasi yang lebih dari 0,8 maka dapat disimpulkan terdapat multikolinieritas pada pasangan variabel tersebut. Melalui pengujian melalui *software* SPSS didapat matriks korelasi sebagai berikut:

$$r = \begin{pmatrix} 1 & 0{,}908 & 0{,}99 & -0{,}03 & 0{,}072 & 0{,}102 & -0{,}128 & -0{,}016 \\ 0{,}908 & 1 & 0{,}899 & -0{,}08 & -0{,}047 & -0{,}031 & -0{,}142 & -0{,}128 \\ 0{,}99 & 0{,}899 & 1 & -0{,}016 & 0{,}082 & 0{,}116 & -0{,}114 & -0{,}011 \\ -0{,}03 & -0{,}08 & -0{,}016 & 1 & -0{,}016 & 0{,}453 & 0{,}081 & 0{,}122 \\ 0{,}072 & -0{,}047 & 0{,}082 & -0{,}016 & 1 & 0{,}241 & 0{,}103 & 0{,}001 \\ 0{,}102 & -0{,}031 & 0{,}116 & 0{,}453 & 0{,}241 & 1 & 0{,}147 & -0{,}017 \\ -0{,}128 & -0{,}142 & -0{,}114 & 0{,}081 & 0{,}103 & 0{,}147 & 1 & 0{,}206 \\ -0{,}016 & -0{,}128 & -0{,}011 & 0{,}122 & 0{,}001 & 0{,}017 & 0{,}206 & 1 \end{pmatrix}$$

Dari matriks korelasi diatas terdapat 1 pasang korelasi yang memiliki nilai lebih dari 0,8 yaitu pada pasangan variabel korelasi kehamilan (X_1), kandungan glukosa (X_2), dan tekanan darah (X_3) terdeteksi multikolinieritas.

Mendeteksi multikolinieritas melalui nilai VIF adalah dengan cara mencari nilai dari R^2 dilanjutkan dengan mencari nilai toleransi, dan didapat nilai VIF. Berikut adalah hasil perhitungan nilai VIF.

Tabel 2. Hasil perhitungan VIF

Variabel	Toleransi	VIF
Kehamilan	0,018	55,402
Kandungan Glukosa	0,145	6,889
Tekanan Darah	0,020	50,903
Ketebalan Kulit	0,751	1,331
Kadar Insulin	0.911	1,098
IMT	0.637	1,571
Silsilah Diabetes	0,903	1,108
Umur	0,850	1,176

Dari tabel diatas terdapat 2 variabel yang terdeteksi multikolinieritas yaitu pada variabel kehamilan (X_1) dengan nilai VIF sebesar 55,402, dan variabel tekanan darah (X_3) dengan nilai VIF sebesar 50,903.

Mengatasi Multikolinieritas

Dikarenakan pada pendeteksian multikolinieritas melalui dua metode yaitu nilai korelasi dan nilai VIF terdeteksi adanya multikolinieritas, maka langkah selanjutnya adalah mengatasi multikolinieritas dengan cara menghapus variabel yang terdeteksi terdapat multikolinieritas.

Selanjutnya akan dilakukan mengatasi multikolinieritas dengan cara menghapus variabel yang terdeteksi multikolinieritas. Setelah dilakukan penghapusan variabel yang bermasalah, didapatkan matriks korelasi sebagai berikut:

$$r = \begin{pmatrix} 1 & -0{,}016 & 0{,}453 & 0{,}081 & 0{,}122 \\ -0{,}016 & 1 & 0{,}241 & 0{,}103 & 0{,}001 \\ 0{,}453 & 0{,}241 & 1 & 0{,}147 & -0{,}017 \\ 0{,}081 & 0{,}103 & 0{,}147 & 1 & 0{,}206 \\ 0{,}122 & 0{,}001 & -0{,}017 & 0{,}206 & 1 \end{pmatrix}$$

Dari matriks korelasi diatas tidak terdapat gejala multikolinieritas setelah dilakukan penghapusan 3 variabel yang bermasalah.

Dikarenakan pada pengujian multikolinieritas melalui nilai VIF didapat 2 variabel yang terdeteksi multikolinieritas yaitu variabel kehamilan (X_1), dan tekanan darah (X_3). Selanjutnya akan dilakukan mengatasi multikolinieritas dengan cara menghapus variabel yang terdeteksi multikolinieritas. Setelah dilakukan penghapusan variabel yang bermasalah, didapatkan nilai VIF sebagai berikut:

Tabel 3. Tabel Pengujian VIF Setelah Penghapusan 2 Variabel

Variabel	Toleransi	VIF
Kandungan Glukosa	0.963	1,038
Ketebalan Kulit	0,760	1,317
Kadar Insulin	0,914	1,094
IMT	0,718	1,393
Silsilah Diabetes	0,917	1,091
Umur	0,927	1,079

Dari tabel diatas tidak terdapat gejala multikolinieritas setelah dilakukan penghapusan 2 variabel yang mempunyai nilai VIF > 10 yaitu variabel kehamilan (X_1), dan variabel tekanan darah (X_3).

4. Kesimpulan

Hasil analisis pada data penelitian dapat disimpulkan bahwa saat pertama kali dideteksi multikolinieritasnya melalu nilai VIF, didapat 2 variabel yang terdeteksi multikolinieritas yaitu pada variabel kehamilan (X_1) dan variabel tekanan darah (X_3). Selanjutnya ketika dideteksi multikolinieritasnya melalui nilai korelasi, didapat 3 variabel yang terdeteksi multikolinieritas yaitu pada variabel kehamilan (X_1), variabel kandungan glukosa (X_2) dan variabel tekanan darah (X_3). Berdasarkan data penelitian, maka variabel kehamilan, kandungan glukosa dan tekanan darah tidak lagi dilibatkan dalam model karena terdeteksi adanya multikolinieritas pada data diabetes militus Wanita Suku Indian tahun 2018. Dengan menghilangkan ketiga variabel diatas, maka masalah multikolinieritas teratasi. Variabel yang dapat dilibatkan untuk mengetahui penderita diabetes adalah ketebalan kulit, kadar insulin, IMT, silsilah diabetes dan umur.

5. Saran

Bila ingin mendeteksi ada atau tidaknya multikolinieritas, Penulis menganjurkan memakai metode korelasi dan nilai VIF dikarenakan lebih mudah digunakan.

Daftar Pustaka

[1] Hastuti, T. (2008). Faktor-Faktor Resiko Ulkus Diabetika Pada Penderita Diabetes Melitus. Semarang: Universitas Dipenogoro.

[2] Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic regression.* United States of American: Sons Inc.

[3] Irawan, D. (2010). *Prevalendsi Dan Faktor Risiko Kejadian Diabetes Melitus Tipe 2 Di Daerah Urban Indonesia.* Jakarta: Fakultas kesehatan masyarakat, Universitas Indonesia.

[4] Kaban, S. (2007). *Diabetes Melitus Tipe 2 Di kota Sibolga Tahun 2005.* Sibolga: Majalah Kedokteran Nusantara.

[5] Kemenkes. (2010). Diabetes Melitus Dapat Dicegah. Diakses pada 9 Desember 2020. http://www.depkes.go.id/index.php?vw=2&id=2383

[6] Santoso, Singgih. (2012). *Panduan Lengkap SPSS Versi 20*. Jakarta: PT Elex Media Komputindo

[7] Senaviratna, SAMR. & Cooray, TMJA. (2019). *Diagnosing Multicollinearity Of Logistic Regression Model.* Srilanka: The Open University Of Srilanka.

[8] Soegondo, S. (2009). *Buku Ajar Penyakit Dalam: Asidosis Lakat, Jilid III, Edisi 4.* Jakarta: FK UI.

[9] Wijanto, S.H. (2008). *Structural Equation Moedeling dengan Lisrel 8.8.* Yogyakarta: Graha Ilmu.