Estimasi, Vol. 1, No. 1, Januari, 2020, Hal. 40-47
DOI: http://dx.doi.org/10.20956/ejsa.v1i1.9262

Kemampuan Estimator Spline Linier dalam Analisis Komponen Utama

Samsul Arifin[1*], **Anna Islamiyati**[2], **Raupong**[3]
[1,2,3]Departemen Statistika, Fakultas MIPA,
Universitas Hasanuddin, Makassar, 90245, Indonesia

*** Corresponding author, email: zula4717@gmail.com**

Abstract

In the formation of a regression model there is a possibility of a relationship between one predictor variable with other predictor variables known as multicollinearity. In the parametric approach, multicollinearity can be overcome by the principal component analysis method. Principal component analysis (PCA) is a multivariate analysis that transforms the originating variables that are correlated into new variables that are not correlated by reducing a number of these variables so that they have smaller dimensions but can account for most of the diversity of the original variables. In some research data that do not form parametric patterns also allows the occurrence of multicollinearity on the predictor variables. This study examines the ability of spline estimators in the analysis of the main components. The data contained multicollinearity and was applied to diabetes mellitus data by taking cholesterol type factors as predictors. Based on the estimation results, one main component is obtained to explain the diversity of variables in diabetes data with the best linear spline model at one knot point.

***Keywords**: Principal Component Analysis, Diabetes Data, Linear, Multicollinearity, Spline.*

Abstrak

Pada pembentukan model regresi terdapat kemungkinan adanya hubungan antara variabel prediktor satu dengan variabel prediktor lainnya yang dikenal dengan istilah multikolinearitas. Pada pendekatan parametrik, multikolinieritas dapat diatasi dengan metode analisis komponen utama. Analisis komponen utama (AKU) adalah analisis multivariat yang mentransformasi variabel asal yang saling berkorelasi menjadi variabel baru yang tidak berkorelasi. AKU bekerja dengan cara mereduksi sejumlah variabel sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya. Pada beberapa data penelitian yang tidak membentuk pola parametrik juga memungkinkan terjadinya multikolinearitas pada variabel prediktornya. Penelitian ini mengkaji tentang kemampuan estimator spline pada analisis komponen utama. Data mengandung multikolinieritas dan diaplikasikan pada data diabetes melitus dengan mengambil faktor jenis kolesterol sebagai variabel prediktornya. Berdasarkan hasil estimasi, diperoleh satu komponen utama untuk menjelaskan keragaman variabel pada data diabetes dengan model spline linier terbaik pada satu titik knot.

Kata Kunci: Analisis Komponen Utama, Data Diabetes, Linier, Multikolinearitas, Spline.

1. Pendahuluan

Analisis regresi merupakan salah satu metode statistika yang digunakan untuk menyelidiki pola hubungan antara variabel respon dan variabel prediktor. Bentuk pola hubungan fungsional antara variabel respon dan variabel prediktor dapat diperkirakan dengan membuat *scatter plot* yang memuat informasi tentang kedua hubungan tersebut. Jika pola datanya diketahui dan mengikuti suatu pola tertentu seperti linier, kuadratik, dan kubik maka dapat digunakan pendekatan regresi parameterik. Akan tetapi, jika pola hubungan keduanya tidak dapat diketahui bentuknya atau tidak tersedia informasi

terkait pola datanya maka digunakan pendekatan regresi nonparametrik. Hal tersebut dikarenakan pendekatan regresi nonparametrik memiliki fleksibilitas yang tinggi dalam membentuk kurva regresi [1].

Pada pembentukan model regresi terdapat kemungkinan adanya hubungan antara variabel prediktor satu dengan variabel prediktor lainnya yang dikenal dengan istilah multikolinearitas. Multikolinieritas menyebabkan variabel prediktor yang seharusnya berpengaruh signifikan terhadap variabel respon akan dinyatakan sebaliknya (tidak nyata secara statistik) sehingga mengakibatkan tidak akuratnya pada peramalan [2]. Pada pendekatan parametrik, multikolinieritas dapat diatasi dengan beberapa metode diantaranya regresi *ridge* dan analisis komponen utama. Analisis komponen utama (AKU) adalah analisis multivariat yang mentransformasi variabel asal yang saling berkorelasi menjadi variabel baru yang tidak berkorelasi dengan cara mereduksi sejumlah variabel tersebut sehingga mempunyai dimensi yang lebih kecil namun dapat menerangkan sebagian besar keragaman variabel aslinya [3].

Beberapa penelitian yang telah menggunakan AKU diantaranya Islamiyati (2014) menggunakan AKU pada model regresi logistik [4], Melany (2017) menerapkan *robust* AKU untuk data yang mengandung *outlier* [5] dan Tharwat (2017) menerapkan AKU pada machine learning [6]. Akan tetapi, permasalahan multikolinearitas tidak hanya terjadi pada pendekatan regresi parametrik. Pada beberapa data penelitian yang tidak membentuk pola parametrik juga memungkinkan terjadinya multikolinearitas pada variabel prediktornya. Hoffman, dkk (2009) menggunakan AKU pada model regresi nonparametrik dengan estimator fungsi kernel [7]. Li dan Hsing (2010) menerapkan AKU pada analisis data longitudinal dengan regresi nonparametrik [8].

Pendekatan regresi nonparametrik telah banyak dikembangkan seperti spline, kernel, polinomial lokal, dan fourier. Diantara pendekatan regresi nonparametrik yang banyak dikembangkan hingga saat ini adalah spline. Spline pada hakekatnya adalah generalisasi dari fungsi polinomial, dimana optimasinya masih mengadopsi konsep dalam pendekatan regresi parametrik. Salah satu keunggulan spline adalah dapat mengatasi pola data yang menunjukkan adanya perubahan perilaku pada sub-sub interval tertentu dengan bantuan titik-titik knot, serta kurva yang dihasilkan relatif mulus [9]. Malik (2014) mengestimasi model regresi nonparametrik multivariabel dengan pendekatan spline pada data longitudinal [10]. Akan tetapi, penelitian tersebut belum mempertimbangkan multikolinearitas yang dapat terjadi pada kasus multiprediktor.

Berdarkan uraian tersebut, penulis mengkaji tentang kemampuan estimator spline dalam analisis komponen utama pada data yang mengandung multikolinieritas. Selanjutnya, metode diaplikasikan pada data diabetes melitus dengan mengambil faktor kolesterol sebagai variabel prediktornya. Penelitian ini mengacu pada penelitian diabetes sebelumnya yang menggunakan pendekatan regresi nonparametrik dengan melibatkan faktor gula darah diet dan lama perawatan [11], faktor glukosa 2 jam setelah

makan [12], dan faktor berat badan [13]. Selanjutnya, hasil penelitian ini menghasilkan satu komponen utama dalam menjelaskan keragaman variansi data diabetes berdasarkan faktor kolesterol. Komponen utama yang dianalisis dengan spline linier diperoleh model optimal pada satu titik knot. Hal ini menunjukkan bahwa terdapat dua pola perubahan gula darah yang berbeda dalam komponen utama yang terbentuk. Perubahan pola yang terbentuk melalui spline AKU diharapkan dapat memberikan kontribusi mendasar dalam permasalahan glukosa diabetes.

2. Material dan Metode

Penelitian ini merupakan penelitian terapan menggunakan metode regresi nonparametrik spline pada data yang mengandung multikolinearitas. Data yang digunakan dalam penelitian ini merupakan data primer, yaitu data penderita penyakit diabetes yang diperoleh dari Rumah Sakit Pendidikan Universitas Hasanuddin tahun 2014-2018. Variabel yang digunakan terdiri dari dua variabel yaitu glukosa darah (mg/dL) sebagai variabel respon dan variabel prediktor yang terdiri dari kolesterol HDL (mg/dL), kolesterol LDL (mg/dL), dan kolesterol total (mg/dL).

Pendekatan regresi nonparametrik merupakan pendekatan regresi dimana bentuk pola datanya tidak diketahui dan hanya diasumsikan mulus dalam arti termuat dalam suatu ruang fungsi tertentu sehingga mempunyai sifat fleksibilitas yang tinggi. Model regresi nonparametrik seperti pada persamaan (1) berikut ini:

$$y_i = f(x_i) + \varepsilon_i; i = 1,2,\dots,n \tag{1}$$

dengan y_i adalah variabel respon sedangkan $f(x_i)$ merupakan kurva regresi dengan x_i sebagai variabel prediktor dan ε_i adalah sisaan yang diasumsikan independen berdistribusi normal dengan rataan nol dan variansi σ^2.

Variabel prediktor pada persamaan (1) diasumsikan terjadi multikolinearitas. Multikolinearitas adalah suatu kondisi dimana terjadi korelasi yang kuat antar variabel prediktor yang satu dengan yang lainnya. Salah satu cara untuk mengetahui adanya multikolinearitas dengan melihat nilai korelasi pada variabel prediktor berdasarkan matriks korelasi. Multikolinearitas ini dapat diatasi dengan metode AKU yang melibatkan nilai eigen dan vektor eigen. Dari proses AKU diperoleh komponen-komponen utama yang memuat variabel-variabel prediktor. Selanjutnya, komponen utama yang terbentuk dimodelkan dengan pendekatan regresi nonparametrik. Model AKU pada regresi nonparametrik seperti pada persamaan (2) berikut ini:

$$y_i = g(w_i) + \varepsilon_i; i = 1,2,\dots,n \tag{2}$$

dengan w_i adalah komponen utama yang terbentuk dari AKU. Fungsi spline linear pada setiap komponen utama yang terbentuk seperti pada persamaan (3) berikut ini:

$$g(w_i) = \beta_0 + \beta_1 w_i + \sum_{h=1}^{r} \beta_{(1+h)}(w_i - k_h)_+ \tag{3}$$

Persamaan diatas dapat dinyatakan dalam bentuk matriks berikut ini:

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

atau

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & w_1 & (w_1 - k_1)_+ & \dots & (w_1 - k_r)_+ \\ 1 & w_2 & (w_2 - k_1)_+ & \dots & (w_2 - k_r)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & w_n & (w_n - k_1)_+ & \cdots & (w_n - k_r)_+ \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{(1+1)} \\ \vdots \\ \beta_{(1+r)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Untuk memperoleh penaksir $\boldsymbol{\beta}$ dilakukan optimasi *least square* yaitu dengan menyelesaikan persamaan berikut ini:

$$\min(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \min[(\boldsymbol{y} - \mathbf{W}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{W}\boldsymbol{\beta})]$$

Memisalkan fungsi $\boldsymbol{Q}(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{W}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{W}\boldsymbol{\beta})$, kemudian fungsi tersebut diturunkan terhadap $\boldsymbol{\beta}$ dan hasil turunannya disamadengankan nol maka diperoleh hasil estimasi parameter $\boldsymbol{\beta}$ seperti pada persamaan (4) berikut ini:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{y} \qquad (4)$$

3. Hasil dan Diskusi

Langkah awal yang dilakukan sebelum memodelkan glukosa darah adalah melakukan statistik deskriptif. Statistik deskriptif ini bertujuan untuk memberikan informasi awal mengenai variabel yang digunakan dalam penelitian. Statistik deskriptif yang digunakan seperti rata-rata, standar deviasi, minimum, dan maksimum seperti pada Tabel 1.

Tabel 1. Statistik deskriptif data diabetes melitus tahun 2014-2018

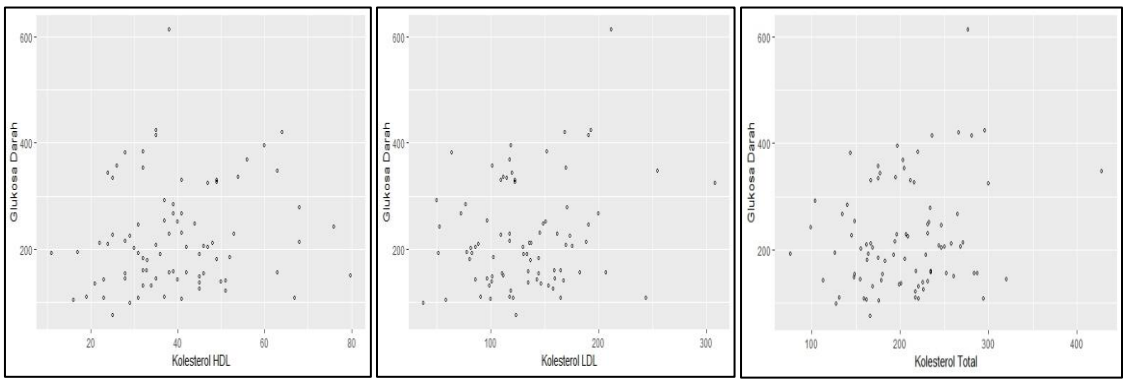
Variabel	Rata-rata	Standar deviasi	Minimum	Maksimum
$\boldsymbol{y}$	224,10	99,80	76	613
$\boldsymbol{x_1}$	39,55	13,70	11	79,8
$\boldsymbol{x_2}$	134,78	48,29	38	308
$\boldsymbol{x_3}$	205,38	56,72	76	428

Sumber: Data diolah 2020

Untuk melihat pola hubungan yang terbentuk antara glukosa darah (y) dengan kolesterol HDL (x_1), kolesterol LDL (x_2), dan kolesterol total (x_3), divisualisasikan seperti pada Gambar 1. Hasil plot menunjukkan pola hubungan antara variabel kolesterol HDL, kolesterol LDL dan kolesterol total terhadap glukosa darah, tidak membentuk suatu pola tertentu sehingga estimasi model menggunakan pendekatan regresi nonparametrik.

Selanjutnya, dilakukan uji multikolinearitas untuk mengetahui adanya korelasi antar variabel prediktor. Untuk mengetahui adanya multikolinearitas pada variabel prediktor digunakan matriks korelasi seperti pada Tabel 2. Matriks korelasi

menunjukkan bahwa terdapat korelasi yang kuat antara variabel x_2 dengan variabel x_3 yaitu 0,856. Hal ini menunjukkan terjadi multikolinearitas pada data sehingga metode AKU digunakan dalam mereduksi variabel. Jumlah komponen utama yang terbentuk digunakan proporsi variansi kumulatif minimal 80%. Nilai proporsi variansi kumulatif seperti pada Tabel 3 dan diperoleh satu komponen utama yang mampu menjelaskan keragaman variansi variabel sebesar 90,5%.



Gambar 1. Scatter plot antara glukosa darah dengan variabel prediktor pada data diabetes melitus tahun 2014-2018.

Tabel 2. Matriks korelasi pada data diabetes melitus tahun 2014-2018

Korelasi	x_1	x_2	x_3
x_1	1,000	0,285	0,430
x_2	0,285	1,000	0,856
x_3	0,430	0,856	1,000

Sumber: Data diolah, 2020

Tabel 3. Proporsi variansi dan proporsi variansi kumulatif

	w_1	w_2	w_3
Proporsi variansi	0,905	0,071	0,024
Proporsi variansi kumulatif	0,905	0,976	1,000

Sumber: Data diolah, 2020

3.1 Pemilihan Titik Knot

Dalam pendekatan regresi nonparametrik spline, dikenal adanya titik knot yaitu titik perpaduan bersama dimana terjadi perubahan pola perilaku kurva pada interval yang berbeda. Letak titik knot dan banyaknya knot merupakan hal yang sangat penting. Metode GCV digunakan untuk menentukan titik knot optimum. Pemilihan titik knot optimum didasarkan nilai GCV minimum. Nilai GCV dari pemodelan dengan menggunakan satu titik knot dan dua titik knot seperti pada Tabel 4 dan Tabel 5. Model terbaik adalah model yang memberikan nilai GCV minimun.

Berdasarkan Tabel 6 terlihat bahwa GCV minimum terdapat pada model spline linier satu titik knot dengan nilai GCV sebesar 9.845,45. Titik knot yang terpilih adalah 273,35 pada komponen utama. Hasil dengan satu titik knot menunjukkan adanya dua pola perubahan glukosa pasien diabetes pada komponen utama yang terbentuk.

Tabel 4. Nilai GCV optimum dengan satu titik knot

No	Titik Knot	Nilai GCV
	k_1	
1	191,09	9.903,87
2	195,16	9.914,18
3	203,81	9.936,06
4	225,04	9.939,35
5	233,91	9.918,74
6	264,86	9.859,27
7	**273,35**	**9.845,45**
8	287,82	9.865,82

Sumber: Data diolah, 2020

Tabel 5. Nilai GCV optimum dengan dua titik knot

No	Titik Knot		Nilai GCV
	k_1	k_2	
1	195,16	244,72	10.118,69
2	201,39	247,94	10.116,72
3	209,42	264,86	10.099,22
4	217,67	268,03	10.092,88
5	226,14	270,22	10.077,56
6	**231,42**	**273,35**	**10.076,60**
7	233,91	275,53	10.076,89
8	235,15	277,67	10.077,55

Sumber: Data diolah, 2020

Tabel 6. Perbandingan nilai GCV minimum

Model	**Nilai GCV**
Model 1 titik knot	**9.845,45**
Model 2 titik knot	10.076,60

Sumber: Data diolah, 2020

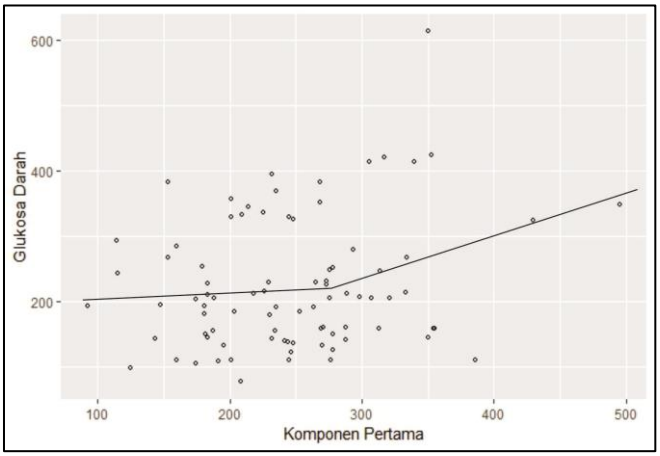
3.2 Pemodelan Glukosa Darah dengan Estimator Spline Linear

Dari pemilihan titik knot diperoleh nilai GCV minimum dengan menggunakan satu titik knot yang berada pada titik $k_1 = 273{,}35$ dengan nilai GCV sebesar 9.845,45.

Adapun estimasi parameter dan estimasi kurva model spline linear dengan satu titik knot adalah sebagai berikut:

$$\hat{y} = 195{,}13 - 0{,}08w + 0{,}58(w - 273{,}35)_+$$

Berdasarkan hasil estimasi model yang bersesuaian dengan Gambar 2 menunjukkan terdapat dua pola perubahan kadar glukosa pasien diabetes pada komponen utama yang terbentuk. Kedua pola menunjukkan kecenderungan naik, namun pola kenaikan antara pola pertama dan kedua yang berbeda. Pola pertama memiliki kencederungan naik perlahan sedangkan pola kedua meningkat tajam.



Gambar 2. Estimasi kurva model spline linear

4. Kesimpulan

Berdasarkan hasil yang diperoleh model terbaik yang digunakan untuk memodelkan glukosa darah adalah model spline linear dengan satu titik knot. Model yang terbentuk seperti berikut ini:

$$\hat{y} = 195{,}13 - 0{,}08w + 0{,}58(w - 273{,}35)_+$$

Berdasarkan hasil estimasi model, variabel prediktor yang termuat dalam komponen utama meningkatkan glukosa darah pada pasien diabetes melitus dengan dua pola peningkatan yang berbeda.

Daftar Pustaka

[1] Islamiyati, A. Spline Polynomial Truncated dalam Regresi Nonparametrik. *Jurnal Matematika, Statistika & Komputasi*, 14 (1) : 54-60, 2017.

[2] Daoud, J.I. Multicollinearity and Regression Analysis. *J. Phys. Conf. Ser*. 949 01200, 2017.

[3] Johnson, R.A. & Wichern, D. W. 2002. *Applied Multivariate Statistical Analysis*. Pentice Hall Inc, New Jersey.

[4] Islamiyati, A. Estimasi Parameter Model Regresi Logistik Biner Komponen Utama Non Linear dengan Maksimum Likelihood. *Jurnal Matematika, Statistika dan Komputasi*, 11 (2) : 122-128, 2015.

[5] Melany, M. Penerapan *Robust Principal Component Analysis* untuk data yang mengandung *Outlier*. *Skripsi*. Universitas Hasanuddin, Makassar. 2017.

[6] Tharwat, A. *Principal Component Analysis-a Tutorial*. ResearcGate. 2017.

[7] Hoffman, H., Schaal, S. & Vijayakumar, S. Local Dimensionality Reduction for Non-Parametric Regression. *Neural Process Lett*, 29 : 109-131, 2009.

[8] Li, B.Y & Hsing, T. Uniform Convergence Rates For Nonparametric Regression and Principal Component Analysis in Longitudinal Data. *The Annals of Statistics*, 38 (6) : 3321-3351, 2010.

[9] Islamiyati, A. Regresi Spline Polynomial Truncated Biprediktor untuk Identifikasi Perubahan Jumlah Trombosit Pasien Demam Berdarah Dengue. *Al khwarizmi*, 7 (2) : 97-110, 2019.

[10] Islamiyati, A., Fatmawati & Chamidah, N. Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with Penalized Spline Regression. *Journal of Physics: Conf. Series.* 979 pp 012093, 2018.

[11] Islamiyati, A., Fatmawati & Chamidah, I.N. Penalized Spline Estimator with Multi Smoothing Parameters in Biresponse Multipredictor Regression Model for Longitudinal Data. *Songklanakarin Journal of Science and Technology*, In Press SJST-2018-0423.R2, 2019.

[12] Islamiyati, A., Fatmawati & Chamidah, I.N. Changes in Blood Glucosa 2 Hours After Meals in Type 2 Diabetes Patients based on Length of Treatment at Hasanuddin University Hospital, Indonesia. *Rawal Medical Journal*, 45 (1) : 31-34, 2020.

[13] Islamiyati, A., Raupong & Anisa. Use of Penalized Spline Linear to Identify Change in Pattern of Blood Sugar based on the Weight of Diabetes Patients. *Int. J. Acad. Appl. Res*., 3 Issue 12 : 75-78, 2019.