ESTIMASI: Journal of Statistics and Its Application
Vol. 5, No. 2, Juli, 2024, Hal. 266-281
DOI: https://doi.org/10.20956/ejsa.v5i2.32283

Perbandingan Analisis Komponen Utama Robust *Minimum* Covarian Determinant dengan Least Trimmed Square pada Data Produk Domestik Regional Bruto

Wa Ode Sitti Amni[1], Andi Kresna Jaya[2], Nirwan[3]
[123]Departemen Statistika, Fakultas MIPA, Universitas Hasanuddin, Makassar, 90245, Indonesia
*** Corresponding author, email: waodesittiamni@gmail.com**

Abstract

Regression analysis is a method to examine the relationship between variables and determine their influence. However, the problem of multicollinearity often arises in linear regression analysis and can cause interpretation problems. To handle multicollinearity, Principal Component Analysis (PCA) is used. However, this method has a weakness when the data contains outliers. Therefore, it was developed into robust PCA using the Minimum Covariance Determinant (MCD) method and the Least Trimmed Square (LTS) estimation method. This study uses Gross Regional Domestic Product data in Indonesia in 2020, which has problems with multicollinearity and outliers. This data is modeled using two robust PCA methods, namely MCD and LTS. The robust PCA model with MCD has an adjusted $R2$ value of 87.87% and an MSE value of 0.0700. However, in the robust PCA regression model with LTS, the adjusted $R2$ value is 98.93% and the MSE value is 0.0325. Thus, the effective method in handling multicollinearity and outliers is the LTS method because it shows better results.

Keywords: *Multicollinearity, Outlier, Principal Component Analysis, Minimum Covariance Determinant, Least Trimmed Square.*

Abstrak

Analisis regresi merupakan metode untuk melihat hubungan antara variabel dan menentukan pengaruhnya. Namun, masalah multikolinearitas sering muncul dalam analisis regresi linear dan dapat menyebabkan masalah interpretasi. Untuk menangani multikolinearitas, digunakan Analisis Komponen Utama (AKU). Namun, metode ini memiliki kelemahan ketika data mengandung outlier. Oleh karena itu, dikembangkan menjadi AKU robust dengan menggunakan metode Minimum Covariance Determinant (MCD) dan metode estimasi Least Trimmed Square (LTS). Penelitian ini menggunakan data Produk Domestik Regional Bruto di Indonesia tahun 2020, yang mengalami masalah multikolinearitas dan outlier. Data ini dimodelkan menggunakan dua metode AKU robust, yaitu MCD dan LTS. Model AKU robust dengan MCD dengan nilai adjusted $R2$ sebesar 87,87% dan nilai MSE sebesar 0,0700. Namun pada Model regresi AKU robust dengan LTS nilai adjusted $R2$ sebesar 98,93% dan nilai MSE sebesar 0,0325. Dengan demikian metode yang efektif dalam menangani multikolinearitas dan outlier yaitu metode LTS karena menunjukkan hasil yang lebih baik.

Kata Kunci: Multikolinearitas, Outlier, Analisis Komponen Utama, Minimum Covariance Determinant, Least Trimmed Square.

1. Pendahuluan

Analisis regresi merupakan teknik statistik yang memungkinkan kita untuk mengeksplorasi dan memahami hubungan antara dua atau lebih variabel, serta memprediksi pengaruh variabel tersebut terhadap satu sama lain [1]. Jika hubungan yang

diteliti melibatkan satu variabel respon dan satu variabel prediktor disebut analisis regresi linear sederhana. Sebaliknya, analisis regresi linear berganda melibatkan satu variabel respon dan lebih dari satu variabel prediktor [2]. Dalam metode regresi linear berganda, terdapat beberapa asumsi yang harus dipenuhi agar model yang dihasilkan dapat dianggap baik. Beberapa asumsi tersebut yaitu asumsi kenormalan, homoskedastisitas, tidak adanya autokorelasi, dan tidak adanya multikolinearitas [3].

Dalam analisis regresi linear berganda, sering kali muncul masalah multikolinearitas antara variabel prediktor. Multikolinearitas terjadi ketika terdapat korelasi yang kuat antara variabel prediktor [4]. Metode statistika yang dapat digunakan untuk mengatasi pengaruh multikolinearitas, seperti mengumpulkan data tambahan, perbandingan dan evaluasi estimator, spesifikasi model dengan menghapus suatu variabel yang berkorelasi, metode regresi ridge dan metode analisis komponen utama.

Analisis Komponen Utama (AKU) adalah teknik statistik yang digunakan untuk mengatasi multikolinearitas dalam regresi linear berganda. Metode ini mereduksi data berdimensi besar dengan mentransformasikan variabel yang berkorelasi menjadi variabel baru yang bebas dari korelasi, yang disebut sebagai komponen utama [5]. Pembentukan metode AKU melibatkan dua langkah: pertama, komponen utama dibentuk berdasarkan matriks varian kovarian atau matriks korelasi variabel prediktor dengan memanfaatkan vektor eigen; kedua, komponen utama yang terpilih diregresikan dengan variabel respon menggunakan regresi komponen utama (RKU) dengan metode ordinary least square (OLS) [6].

AKU klasik memiliki kekurangan dalam menghadapi data yang mengandung outlier, karena baik vektor rata-rata maupun matriks kovarian atau korelasi sampel sangat rentan terhadap outlier. Hal ini juga berlaku untuk metode OLS dalam RKU, yang dapat menghasilkan perkiraan parameter regresi yang kurang akurat dan efisien jika data mengandung outlier [7]. Oleh karena itu akan dikembangkan menjadi regresi robust yang menerapkan metode robust pada kedua tahap tersebut. Regresi robust merupakan metode regresi yang digunakan pada saat distribusi dari error yang tidak normal dalam model. Metode robust yang digunakan dalam penelitian ini adalah metode minimum covariance determinant (MCD) pada saat melakukan AKU dan metode estimasi least trimmed square (LTS) pada saat melakukan RKU.

Berdasarkan dari kedua penelitian dengan metode yang berbeda penulis ingin mengetahui cara pengaplikasian analisis komponen utama dengan MCD dan analisis komponen utama dengan LTS serta mengetahui metode manakah yang terbaik digunakan dalam mengatasi multikolinearitas dan outlier yang terjadi pada data produk domestik regional bruto di Indonesia tahun 2020.

2. Material dan Metode

2.1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang bersumber dari Badan Pusat Statistika, Kementrian Kesehatan Repoblik Indonesia, Kementrian Pekerjaan Umum dan Perumahan Rakyat tahun 2020. Variabel respon yang digunakan adalah produk domestik regional bruto (PDRB) setiap provinsi di Indonesia. Sedangkan, variabel prediktor yang digunakan pada penelitian ini terdiri dari Panjang jalan, distribusi listrik, infrastruktur kesehatan, infrastruktur pendidikan, infrastruktur pariwisata, infrastruktur perumahan dan fasilitas industri.

2.2. Analisis Regresi

Regresi linear berganda adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu variabel respon dan dua atau lebih variabel prediktor. Analisis ini bertujuan untuk menentukan arah hubungan antara variabel- variabel tersebut, masing-masing variabel prediktor bisa berhubungan positif atau negatif, serta untuk memprediksi nilai variabel respon yang mengalami kenaikan atau penurunan berdasarkan perubahan pada variabel prediktor [8]. Dengan bertambahnya variabel prediktor maka bentuk umum dari persamaan regresi linear berganda yang mencakup dua atau lebih variabel prediktor dapat ditulis sebagai berikut [9]:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon \quad (1)$$

Berdasarkan Persamaan (1) maka bentuk matriks dari model regresi linear berganda dapat dituliskan sebagai berikut:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Atau dapat ditulis dengan $\boldsymbol{Y} = \boldsymbol{X}\mathrm{Q} + \mathrm{s}$, dengan: $\boldsymbol{Y}$ adalah vektor pengamatan berukuran $n\times 1$, $\boldsymbol{X}$ adalah matriks variabel prediktor ukuran $n \times (p+1)$, β adalah vektor parameteryang akan ditaksir berukuran $(p+1) \times 1$ dan s adalah vektor random *error* berukuran $n\times 1$.

2.3. Estimasi Parameter Model

Menurut [10] estimasi parameter dapat diperoleh dengan menggunakan metode *ordinary least squares* (OLS) yaitu dengan meminimumkan jumlah kuadrat *error* sesuaipersamaan (2) berikut:

$$S(Q) = sTs = YTY - 2QTXTY + QTXTXQ \quad (2)$$

Untuk mendapatkan estimator OLS (Q) yang meminimumkan $S(Q)$ disyaratkan bahwa:

$$\left.\frac{\partial S(Q)}{\partial Q}\right|_{\beta=\hat{\beta}} = 0$$

Maka:

$$\left.\frac{\partial S(\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = 0$$

$$-2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Perbedaan unit satuan pada model regresi yang tidak distandarkan dapat menyebabkan koefisien regresi tidak bisa dibandingkan. Oleh karena itu, perlu dilakukan standarisasi menggunakan rumus sebagai berikut:

$$Y_i^* = \frac{Y_i - \bar{Y}}{S_y} \text{ dan } X_i^* = \frac{X_{ji} - \bar{X}_J}{S_j} \quad (4)$$

dengan $S_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$ dan $S_j = \sqrt{\frac{\sum_{i=1}^{n}(X_{ji}-\bar{X}_j)}{n-1}}$ Sehingga diperoleh model regresi standar sebagai berikut:

$$Y_1^* = \beta_{1i}^* X_{1i}^* + \beta_{2i}^* X_{2i}^* + \cdots + \beta_{ki}^* X_{ki}^* + \varepsilon_i^*$$

2.4. Uji Ukuran Kebaikan Model Regresi

Metode yang digunakan untuk menentukan ukuran kebaikan model dilakukan dengan menghitung nilai mean square error (MSE). Perhitungan MSE dilakukan menggunakan persamaan berikut: [10]

$$MSE = \frac{JKE}{n-p-1} \quad (5)$$

dengan JKE adalah jumlah kuadrat residual, n adalah jumlah sampel, dan p adalah banyaknya variabel prediktor. Jika nilai MSE semakin kecil hingga mendekati nol maka dapat dikatakan bahwa model regresi semakin baik. Selain metode tersebut, untuk menentukan ukuran kebaikan model juga dapat menggunakan koefisien determinasi

adjusted R^2. Kecocokan model lebih baik jika nilai *adjusted* R^2 semakin mendekati satu. Adapun rumus untuk mendapatkan nilai *adjusted* R^2 sebagai berikut:

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n-p-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 / (n-1)} \quad (6)$$

dengan $0 \leq adjusted\ R2 \leq 1$, apabila $adjusted\ R^2 = 0$ artinya tidak ada hubungan di antara X dan Y atau secara model regresi yang terbentuk tidak tepat untuk medefinisikan Y dan apabila nilai *adjusted* $R^2 = 1$ artinya garis regresi yang terbentuk dapat mendefinisikan Y secara sempurna.

2.5. Multikolinearitas

Multikolinearitas terjadi ketika terdapat keterkaitan linear yang signifikan antara variabel prediktor dalam model regresi berganda. Hubungan linear yang kuat antara variabel prediktor dapat menyebabkan estimasi parameter menjadi tidak stabil dan sulit untuk diinterpretasikan dengan baik [11].

Menurut [10] cara mendeteksi adanya multikolinearitas pada model regresi dapat dilakukan dengan menggunakan nilai *variance inflation factory* (VIF). Berikut persamaan yang digunakan dalam menghitung nilai VIF:

$$VIF = \frac{1}{\left(1 - R_j^2\right)} = \frac{1}{Tolerance} \quad (7)$$

$$R^2_{adj} = 1 - \frac{JKE/(n-k-1)}{JKR/(n-1)}$$

Keterangan:
JKE = Jumlah Kuadrat Error
JKR = Jumlah Kuadrat Regresi

Dengan R^2 adalah koefisien determinasi ke-j, di mana $j = 1, 2, 3, \cdots, p$. Batas nilai yang digunakan untuk VIF adalah 10 dan batas nilai dari *tolerance* adalah 0.1. Apabila nilai yang diperoleh $VIF \geq 10$ dan $tolerance < 0.1$ maka terjadi multikolinearitas yang kuat diatara variabel prediktor dan sebaliknya.

2.6. Outlier

Outlier adalah data yang tidak mengikuti pola yang umumnya diikuti oleh model atau data yang berada di luar model yang dibentuk, serta tidak berada dalam daerah selang kepercayaan yang telah ditentukan [12]. Adanya *outlier* pada variabel prediktor dapat dideteksi dengan menghitung jarak mahalanobis. Untuk mengukur jarak mahalanobis digunakan vektor rata-rata dan matriks kovarian. Sebuah pengamatan X_i dideteksi sebagai *outlier* jika jarak mahalanobisnya:

$$d_{MD}^2 = (X_i - \bar{X})^T S^{-1} > X_{(p;2)}^2 \quad (8)$$

dengan $\bar{X}$ dan S adalah vector rata-rata dan matriks kovarian dari data.

Selanjutnya, untuk mendeteksi adanya outlier pada variabel respon dapat dilakukan dengan melihat *error* dari model regresi. Salah satu metode yang dapatdigunakan adalah metode DFFITS (*Difference in Fit Standardized*). Perhitungan DFFITS dapat ditulis sebagai berikut [10]:

$$DFFITS_i = t_i \left(\frac{h_i}{1-h_i}\right)^{1/2} \text{ dengan } t_i = \varepsilon_i \sqrt{\frac{n-p-1}{JKE(1-h_i)-\varepsilon_i^2}} \quad (9)$$

Sebuah data bisa dikatakan *outlier* apabila $|DFFITS| > 2\sqrt{\frac{k}{n}}$ dengan $k = p + 1$ adalah banyaknya variabel data dan n adalah banyak jumlah observasi.

2.7. Matriks

Vektor mean sampel dari matriks varians-kovarians sampel dapat diperoleh dengan cara sebagai berikut. Nilai varian-kovarian dapat diperoleh dengan mengikuti persamaan sebagai berikut [13]:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

Matriks rata-rata peubah prediktor:

$$\bar{\boldsymbol{X}} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \cdots & \bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & \cdots & \bar{x}_p \end{bmatrix}$$

Selisih antara matriks peubah prediktor dan matriks rata-rata peubah prediktor:

$$\boldsymbol{X} - \bar{\boldsymbol{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & x_{13} - \bar{x}_3 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & x_{23} - \bar{x}_3 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & x_{n3} - \bar{x}_3 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

Persamaan matriks varian-kovarian adalah perkalian antara matriks $(\boldsymbol{X} - \bar{\boldsymbol{X}})^T$ dengan matriks $(\boldsymbol{X} - \bar{\boldsymbol{X}})$:

$$\boldsymbol{S} = \frac{1}{n-1}[(\boldsymbol{X} - \bar{\boldsymbol{X}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}})] \tag{10}$$

Matriks koefisien korelasi adalah matriks simetri yang dapat dinyatakan dalam persamaan berikut ini:

$$\boldsymbol{R} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{21}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{12} & 1 & \cdots & \rho_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \tag{11}$$

2.8. Nilai Eigen dan Vektor Eigen

Misalkan matriks S adalah matriks berukuran $p \times p$, terdapat suatu skalar λ vektor tak nol v sehingga memenuhi persamaan berikut [13]:

$$Sa = \lambda a \tag{12}$$

Skalar λ disebut nilai eigen dari $\boldsymbol{S}$ dan a disebut vektor eigen dari $\boldsymbol{S}$ yang bersesuaian dengan λ. Untuk memperoleh nilai eigen pada persamaan (12) dapat dituliskan menjadi:

$$Sa = \lambda a; \text{ dengan } v \neq 0$$

$$Sa - \lambda a = 0$$

$$Sa - \lambda I = 0$$

$$(S - \lambda I)a = 0$$

Agar λ menjadi nilai eigen, maka harus terdapat solusi tak nol dari persamaan (13). Persamaan tersebut akan memiliki penyelesaian tak nol jika dan hanya jika:

$$|S - \lambda| = 0 \tag{13}$$

2.9. Analisis Komponen Utama

Analisis komponen utama (AKU) adalah suatu teknik analisis statistik yang digunakan untuk mentransformasi sekumpulan variabel yang saling berkorelasi menjadi satu set variabel baru yang bebas dari korelasi [14]. Komponen utama merupakan suatu kombinasi linear berdasarkan pada skala pengukuran variabel acak $X_1, X_2, \cdots, X_p$ yang sama dan memiliki struktur matriks varian kovarian $\boldsymbol{S}$ yang kemudian akan dihasilkan nilai eigen sebanyak $p(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p)$. Maka komponen utama yang merupakan kombinasi linear dari $\boldsymbol{X}^T$ variabel asal dan e sebagai vektor eigen didefinisikan sebagai berikut:

$$\begin{aligned} w_1 &= \boldsymbol{a}_j^T \boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ w_2 &= \boldsymbol{a}_j^T \boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ w_p &= \boldsymbol{a}_j^T \boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{14}$$

Dengan matriks sebagai berikut:

$$\boldsymbol{a}_j^t \boldsymbol{X} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

dengan w_1 adalah komponen pertama yang memenuhi maksimum nilai $\boldsymbol{a}_1^T S \boldsymbol{a}_1 = \lambda_1, w_2$ adalah komponen kedua yang memenuhi sisa keragaman selain komponen pertama dengan memaksimumkan nilai $\boldsymbol{a}_2^T S \boldsymbol{a}_2 = \lambda_2$, dan w_p adalah komponen ke p yang memenuhi sisa keragaman selain dari $w_1, w_2, \cdots, w_p$ dengan memaksimumkan nilai $\boldsymbol{a}_j^T S \boldsymbol{a} = \lambda$. Syarat untuk membentuk komponen utama yang merupakan kombinasi linear dari variabel *X* agar mempunyai variansi maksimum adalah dengan memilih vektor eigen yaitu $\boldsymbol{a}^T = (a_1, a_2, \cdots, a_p)$ sedemikian hingga $var(w) = \boldsymbol{a}_j^T S \boldsymbol{a}$ maksimum dan $\boldsymbol{a}_j^T \boldsymbol{a}_J = 1$ [15].

Kriteria pemilihan komponen utama menggunakan persentase variansi kumulatif terhadap total variansi. Persentase variansi kumulatif yang dijelaskan komponen utama ke-j dapat dihitung menggunakan persamaan berikut ini [16]:

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\% = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \times 100\%; \text{untuk } j = 1,2,\cdots,p \tag{15}$$

Komponen utama dapat ditentukan dengan melihat proporsi kumulatif varians dan mampu menerangkan total variansi data sekitar 70% sampai 80%.

2.10. RegresiKomponenUtama

Metode regresi komponen utama (RKU) adalah suatu pendekatan analisis yang menggabungkan analisis regresi dengan AKU. Dalam metode ini, analisis regresi digunakan untuk menentukan keberadaan hubungan antara variabel respon dan variabel predictor. Metode RKU difungsikan untuk meminimumkan masalah multikolinearitas yang terjadi di dalam model regresi linear berganda dengan cara tidak menghilangkan variabel prediktor yang bersinggungan dengan kolinieritas. Menurut [17] metode RKU dilakukan setelah melakukan analisis menggunakan metode AKU untuk mendapatkan nilai-nilai komponen utama. Kemudian komponen utama yang dipilih akan diregresikan dalam model regresi bersama variabel respon menggunakan metode OLS.

2.11. Minimum Covarian Determinant

Metode *minimum covarian determinant* (MCD) merupakan penaksir *robust* untuk rata-rata dan matriks kovarian dengan mencari sebagian data yang mempunyai kovarian minimum yang digunakan untuk mengidentifikasi *outlier*, menentukan jarak dan residu *robust* yang akan digunakan dalam pembobotan data dan penentuan parameter regresi [18]. Metode MCD bersifat resisten terhadap keberadaan *outlier* di dalam data pengamatan, sehingga sangat berguna untuk mendeteksi *outlier*. Tujuan MCD untuk mendapatkan suatu subsampel berukuran h dari keseluruhan pengamatan n, yang matriks varian kovariannya memiliki determinan terkecil diantara semua kombinasi kemungkinan data, dengan menggunakan persamaan sebagai berikut:

$$h = \frac{n + p + 1}{2} \tag{16}$$

Metode estimasi MCD mudah dihitung dan ditemukan apabila jumlah n pengamatan kecil. Apabila jumlah n pengamatan besar maka akan banyak kombinasi subsampel dari h yang akan dicari dan ditemukan. Oleh karena itu, metode MCD menggunakan persamaan sebagai berikut:

$$\bar{X}_{MCD} = \frac{1}{h}\sum_{i \in h}^{h} [X_i - \bar{X}_{MCD}]^T [X_i - \bar{X}_{MCD}] \tag{17}$$

dan

$$S_{MCD} = \frac{1}{h}\sum_{i \in h}^{h} [X_i - \bar{X}_{MCD}]^T [X_i - \bar{X}_{MCD}] \tag{18}$$

2.12. Least Trimmed Square

Least trimmed squares (LTS) adalah metode estimasi regresi yang digunakan untuk mengatasi dampak data *outlier* pada model regresi. LTS mengambil pendekatan yang berbeda dengan metode OLS dengan mempertimbangkan hanya sebagian kecil data yang paling cocok dengan model dan mengabaikan data yang dianggap sebagai *outlier* [19]. LTS adalah salah satu metode pendugaan parameter pada regresi *robust* yang sangat tahan terhadap keberadaan *outlier* dan memiliki *high breakdown point*. Metode ini memilih subhimpunan data dengan ukuran s, penentuan ukuran s menggunakan rumus $\left(\frac{(3h+w+1)}{4} \leq s \leq h\right)$, di mana s merupakan jumlah pengamatan yang dianggap bukan sebagai *outlier*. Kemudian, metode ini meminimalkan jumlah kuadrat *error* hanya dari subhimpunan data tersebut. Dengan demikian, metode LTS memberikan bobot yang lebih rendah pada data yang dianggap sebagai *outlier* [20].

Pendekatan ini memungkinkan metode LTS untuk menjadi lebih tahan terhadap keberadaan outlier. Dengan hanya mempertimbangkan sebagian kecil data yang dianggap sebagai data yang tidak outlier, metode LTS dapat memberikan estimasi parameter yang lebih konsisten dan robust terhadap pengaruh outlier. Metode LTS menduga koefisien regresi dengan menggunakan metode OLS terhadap subhimpunan data berukuran s dengan menggunakan persamaan sebagai berikut:

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{h}\varepsilon_i^2\right) = \arg\min_{\beta}\left(\sum_{i=1}^{h}\left(Y_i - \hat{Y}_i\right)^2\right), \frac{(3h+w+1)}{4} \leq S \leq h \qquad (19)$$

3. Hasil dan Diskusi

3.1. Uji Multikolinearitas

Salah satu cara yang umum digunakan untuk menguji keberadaan multikolinearitas dalam suatu model regresi adalah dengan menggunakan nilai VIF. Berdasarkan hasil perhitungan menggunakan Persamaan (7) maka diperoleh hasil sebagai berikut:

Table 1. Uji Multikolinearitas

Variabel Bebas	VIF
Panjang Jalan (X_1)	5,1540
Distribusi Listrik (X_2)	62,2107
Infrastruktur Kesehatan (X_3)	22,5643
Infrastruktur Pendidikan (X_4)	46,4837
Infrastruktur Pariwisata (X_5)	108,5055
Infrastruktur Perumahan (X_6)	8,7754
Fasilitas Industri (X_7)	75,4356

Berdasarkan Tabel 1 dapat dilihat bahwa lebih banyak nilai VIF yang melebihi ambang batas 10. Hal ini terjadi pada variabel distribusi listrik, infrastruktur kesehatan, infrastruktur pendidikan, infrastruktur pariwisata, serta fasilitas industri. Oleh karena itu, dapat disimpulkan bahwa pada data PDRB Indonesia tahun 2020, terdapat permasalahan yang signifikan terkait multikolinearitas antara variabel prediktor.

3.2. Uji Outlier

Untuk mengidentifikasi outlier yang muncul dalam variabel prediktor dapat dilakukan dengan menghitung jarak mahalanobis. Berdasarkan perhitungan bahwa terdapat beberapa outlier pada variabel prediktor. Outlier terdapat pada provinsi DKI Jakarta, Jawa Tengah dan Jawa timur. Oleh karena itu, dapat disimpulkan bahwa pada data PDRB di Indonesia tahun 2020 terdapat masalah outlier pada variabel prediktor.

Selanjutnya untuk mendeteksi outlier pada variabel respon dapat dilakukan dengan menggunakan nilai DFFITS. Berdasarkan perhitungan bahwa terdapat beberapa nilai DFFITS yang melewati batas outlier yang telah ditetapkan yaitu 0,9074. Pencilan-pencilan ini terdeteksi di provinsi Riau, Jawa Timur, dan Kalimantan Timur. Oleh

karena itu, dapat disimpulkan bahwa pada data PDRB di Indonesia tahun 2020 terdapat masalah outlier pada variabel prediktor dan variabel respon.

3.3. Analisis Komponen Utama dengan MCD

Tujuan MCD adalah untuk mendapatkan suatu subsampel berukuran h dari keseluruhan pengamatan n, yang matriks varian kovariansnya memiliki determinan terkecil diantara semua kombinasi kemungkinan data, berdasarkan perhitungan menggunakan Persamaan (10) maka diperoleh nilai $h = 31$. Adapun matriks korelasi yang diperoleh berdasarkan MCD sebagai berikut:

$$\boldsymbol{R} = \begin{bmatrix} 1{,}0000 & 0{,}5048 & 0{,}7886 & 0{,}8057 & 0{,}5157 & 0{,}2209 & 0{,}7846 \\ 0{,}5048 & 1{,}0000 & 0{,}7815 & 0{,}7971 & 0{,}9356 & 0{,}6999 & 0{,}6314 \\ 0{,}7886 & 0{,}7815 & 1{,}0000 & 0{,}9259 & 0{,}7832 & 0{,}4852 & 0{,}8154 \\ 0{,}8057 & 0{,}7971 & 0{,}9259 & 1{,}0000 & 0{,}8435 & 0{,}3980 & 0{,}9165 \\ 0{,}5157 & 0{,}9356 & 0{,}7832 & 0{,}8435 & 1{,}0000 & 0{,}4626 & 0{,}7257 \\ 0{,}2209 & 0{,}6999 & 0{,}4852 & 0{,}3980 & 0{,}4626 & 1{,}0000 & 0{,}1883 \\ 0{,}7846 & 0{,}6314 & 0{,}8154 & 0{,}9165 & 0{,}7257 & 0{,}1883 & 1{,}0000 \end{bmatrix}$$

Matriks korelasi yang terbentuk akan digunakan untuk menghitung nilai eigen dan vektor eigen. Untuk menghitung vektor eigen menggunakan Persamaan (12) dan untuk menghitung nilai eigen menggunakan Persamaan (13) maka diperoleh hasil perhitungan berdasarkan Tabel 2 sebagai berikut:

Table 2. Nilai Eigen dan Vektor Eigen

λ	1,7834	0,4191	0,2361	0,1651	0,1323	0,0631	0,0004
a_1	0,0001	0,4281	0,4050	0,4259	0,2803	0,5648	0,2716

a_2	-0,0010	0,2975	-0,6275	-0,4774	0,2212	0,4897	-0,0315
a_3	-0,0007	0,2378	-0,3820	0,2421	0,3976	-0,5432	0,5346
a_4	-0,0001	-0,6509	0,0731	-0,2386	0,0765	0,2646	0,6619
a_5	-0,0003	0,2234	0,5344	-0,6441	0,4204	-0,2696	-0,0123
a_6	0,0004	-0,4450	-0,0728	0,2455	0,7292	0,0581	-0,4486
a_7	-1,0000	-0,0006	0,0008	0,0007	-0,0003	0,0001	-0,0005

Terlihat bahwa komponen utama ke-1 dan ke-2 telah mewakili variansi dari data, sehingga dalam penelitian ini akan dipilih komponen utama ke-1 dan ke-2 untuk digunakan dalam analisis selanjutnya. Proporsi kumulatif varian yang dapat dijelaskan oleh dua komponen utama terpilih adalah sebagai berikut:

$$\frac{1{,}7834 \ + \ 0{,}4191}{1{,}7834 \ + \ 0{,}4191 \ + \ 0{,}2361 \ + \cdots + \ 0{,}0004} \times 100\% = 0{,}9698$$

RKU dibentuk berdasarkan komponen utama terpilih dari proses AKU-MCD sebelumnya. Pemodelan untuk RKU yaitu dengan membentuk model regresi linier berganda menggunakan metode OLS, dengan komponen utama w_1 dan w_2 sebagai variabel prediktor dan y sebagai variabel respon. komponen utama dengan MCD yang dikombinasikan dengan analisis regresi OLS menghasilkan persamaan model sebagai berikut:

$$\hat{Y} = -0{,}2412 + 0{,}3158w_1 + 0{,}0164w_1$$

Jika dikembalikan ke data yang distandarisasi maka model yang didapatkan sebagai berikut:

$$\hat{Y} = -0{,}2412 \ + \ 0{,}00002Z_1 + \ 0{,}1401Z_2 + \ 0{,}1175Z_3 + \ 0{,}1266Z_4 + \ 0{,}0921Z_5$$

$$+ 0{,}1864Z_6 + \ 0{,}0852Z_7$$

Selanjutnya diperoleh persamaan menggunakan data awal didapatkan model sebagai berikut:

$$\hat{Y} = 67090 + \ 0{,}0111X_1 + 4{,}1005X_2 + 121{,}8251X_3 + 0{,}3834X_4 + 0{,}0015X_5$$

$$+ 1{,}1551X_6 + 0{,}2071X_7$$

3.4. Analisis Komponen Utama dengan LTS

Tahap awal dari AKU melibatkan penentuan matriks korelasi. Matriks korelasi digunakan ketika terdapat unit data yang serupa, dan berperan penting dalam perolehan nilai eigen dan vektor eigen. Adapun matriks korelasi yang diperoleh berdasarkan Persamaan (11) sebagai berikut:

$$\boldsymbol{R} = \begin{bmatrix} 1{,}0000 & 0{,}3726 & 0{,}0632 & 0{,}0077 & 0{,}1584 & -0{,}0892 & -0{,}0511 \\ 0{,}3726 & 1{,}0000 & 0{,}8552 & 0{,}8665 & 0{,}8792 & 0{,}7409 & 0{,}7702 \\ 0{,}0632 & 0{,}8552 & 1{,}0000 & 0{,}9735 & 0{,}8805 & 0{,}6921 & 0{,}8837 \\ 0{,}0077 & 0{,}8665 & 0{,}9735 & 1{,}0000 & 0{,}8873 & 0{,}7269 & 0{,}9030 \\ 0{,}1584 & 0{,}8792 & 0{,}8805 & 0{,}8873 & 1{,}0000 & 0{,}5455 & 0{,}9611 \\ -0{,}0892 & 0{,}7409 & 0{,}6921 & 0{,}7269 & 0{,}5455 & 1{,}0000 & 0{,}5318 \\ -0{,}0511 & 0{,}7702 & 0{,}8837 & 0{,}9030 & 0{,}9611 & 0{,}5318 & 1{,}0000 \end{bmatrix}$$

Matriks korelasi yang terbentuk selanjutnya digunakan untuk menghitung nilai eigen dan vektor eigen. Untuk menghitung vektor eigen menggunakan Persamaan (12) dan untuk menghitung nilai eigen menggunakan Persamaan (13) maka diperoleh hasil perhitungan berdasarkan Tabel 3 sebagai berikut:

Table 3. Nilai Eigen dan Vektor Eigen

λ	2,2522	1,0666	0,7673	0,3737	0,1859	0,1512	0,0636
a_1	0,0493	0,4152	0,4277	0,4325	0,4196	0,3340	0,4093
a_2	-0,9276	-0,2570	0,0478	0,0996	-0,0722	0,1983	0,1294
a_3	0,0748	0,2054	-0,0533	-0,0259	-0,3444	0,8007	-0,4347
a_4	0,1028	-0,2563	0,6501	0,4357	-0,4425	-0,2182	-0,2604
a_5	-0,3187	0,6711	-0,1371	0,2367	0,0370	-0,3848	-0,4731
a_6	0,1294	-0,0883	-0,5974	0,6798	-0,2910	0,0130	0,2676
a_7	-0,0511	0,4422	0,1167	-0,3071	-0,6467	-0,1103	0,5134

Terlihat bahwa komponen utama ke-1 dan ke-2 telah mewakili variansi dari data, sehingga dalam penelitian ini dipilih komponen utama ke-1 dan ke-2 untuk digunakan dalam analisis selanjutnya. Proporsi kumulatif varian yang dapat dijelaskan oleh dua komponen utama terpilih adalah sebagai berikut:

$$\frac{2{,}2522 + 1{,}0666}{2{,}2522 + 1{,}0666 + 0{,}7673 + \cdots + 0{,}0636} \times 100\% = 88{,}71$$

Pemodelan untuk RKU menggunakan metode LTS dengan komponen utama w_1 dan w_2 sebagai variabel prediktor dan *Y* sebagai variabel respon. Metode LTS meminimalkan jumlah kuadrat *error* dari subsampel data. Metode LTS menggunakan subsampel s terbaik dihitung menggunakan rumus $\frac{(3n+w+1)}{4} \leq s \leq h$. Sehingga diperoleh sejumlah 26 sampel yang dianggap memiliki galat paling minimum, adapun galat yang paling besar akan diasumsikan sebagai data *outlier*.

Persamaan yang paling baik yang diperoleh menggunakan metode LTS adalah sebagai berikut:

$$\hat{Y} = -0{,}09762 + 0{,}3540w_1 - 0{,}4713w_2$$

Jika dikembalikan ke data yang distandarisasi maka model yang didapatkan sebagai berikut:

$$\hat{Y} = -0{,}0976 + 0{,}4547Z_1 + 0{,}2681Z_2 + 0{,}1288Z_3 + 0{,}1062Z_4 + 0{,}1825Z_5 + 0{,}0248Z_6 + 0{,}0839Z_7$$

Selanjutnya diperoleh persamaan menggunakan data awal didapatkan model sebagai berikut:

$$\hat{Y} = -10337 + 0{,}1600X_1 + 9{,}9348X_2 + 184{,}2580X_3 + 0{,}4413X_4 + 0{,}0028X_5 + 0{,}2297X_6 + 0{,}1841X_7$$

3.5. Uji kebaikan Model

Uji kebaikan model ditentukan berdasarkan nilai MSE dan a$dusted\ \boldsymbol{R^2}$. Model yang lebih baik akan menghasilkan nilai MSE yang lebih kecil dan nilai a$dusted\ \boldsymbol{R^2}$ yang lebih besar. Untuk menghitung nilai MSE menggunakan Persamaan (5) dan untuk menghitung nilai nilai a$dusted\ \boldsymbol{R^2}$ menggunakan Persamaan (6) maka diperoleh hasil perhitungan berdasarkan Tabel 4 sebagai berikut:

Table 4. Uji Kebaikan Model

Metode	Nilai a$dusted\ R^2$	Nilai MSE
AKU-MCD	87,87%	0,0700
AKU-LTS	98,93%	0,0325

Berdasarkan Tabel 4 dapat disimpulkan bahwa dari kedua metode yangdigunakan, metode AKU-LTS merupakan yang paling baik dalam memodelkan data, ditunjukkan oleh nilai $Adusted\ R^2$ sebesar 98,93% dan nilai MSE sebesar 0,0325. Hal ini mengindikasikan bahwa model AKU-LST memberikan informasi yang sangat baik dan memiliki tingkat kesalahan (MSE) yang paling rendah jika dibandingkan dengan metode AKU-MCD. Adapun AKU-MCD memiliki nilai $Adusted\ R^2$ sebesar 87.87% dan nilai MSE sebesar 0,0700.

4. Kesimpulan

Berdasarkan hasil analisis data dan pembahasan pada penelitian ini dapat ditarik kesimpulan bahwa analisis komponen utama *robust least trimmed square* paling efektif untuk menangani multikolinearitas dan outlier pada pemodelan Produk Domestik Regional Bruto di Indonesia Tahun 2020. Efektivitas penerapan metode analisis komponen utama *robust least trimmed square* dalam pemodelan Produk Domestik Regional Bruto di Indonesia Tahun 2020 ditunjukan dengan nilai *Adjusted* R^2 diperoleh sebesar 98,93% dan nilai MSE sebesar 0,0325. Nilai tersebut merupakan nilai *Adjusted*

R^2 yang paling besar dan MSE yang paling kecil dibandingkan dengan metode analisis komponen utama robust minimum covarian determinant dengan nilai *Adjusted* R^2 sebesar 87,87% dan nilai MSE sebesar 0,0700.

Daftar Pustaka

[1] Subandriyo, B. *Buku Ajar Analisis Kolerasi dan Regresi*. Diklat Statistisi Tingkat Ahli BPS Angkatan XXI, 31, 2020.

[2] Padilah, T. N., & Adam, R. I. Analisis Regresi Linier Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang. *FIBONACCI: Jurnal Pendidikan Matematika dan Matematika*, 5(2), 117, 2019, doi: 10.24853/fbc.5.2.117-128.

[3] Shodiqin, A. Perbandingan Metode Regresi Robust yakni Metode Least Trimmed Squares (LTS) dengan metode Estimator-MM (Estimasi-MM). 2018.

[4] Larasati, S. D. A., Nisa, K., & Setiawan, E. Analisis Regresi Komponen Utama Robust dengan Metode Minimum Covariance Determinant – Least Trimmed Square (MCD-LTS). *Jurnal Siger Matematika*, 1(1), pp. 1–9, 2020, doi: 10.23960/jsm.v1i1.2472.

[5] Pendi. *Analisis Regresi dengan Metode Komponen Utama dalam Mengatasi Masalah Multikolinearitas Pendi Intisari*. 2021.

[6] Fabiana. Pemodelan Indeks Pembangunan Manusia di Jawa Tengah dengan Regresi Komponen Utama Robust. 8(1999), pp. 253–271, 2019.

[7] Filzmoser, P. Robust principal component and factor analysis in the geostatistical treatment of environmental data. *Environmetrics,* 10(4), pp. 363–375, 1999, doi: 10.1002/(SICI)1099-095X(199907/08)10:4<363::AID- ENV362>3.0.CO;2-0.

[8] Adiguno, S., Syahra, Y., & Yetri, M. Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, 1(4), p. 275, 2022, doi: 10.53513/jursi.v1i4.5331.

[9] Ningsih, S., & Dukalang, H. H. Penerapan Metode Suksesif Interval pada Analsis Regresi Linier Berganda. *Jambura Journal of Mathematics*, 1(1), pp. 43–53, 2019, doi: 10.34312/jjom.v1i1.1742.

[10] Montgomery, E. A., & Peck, D. C. Introduction To Linier Regression Analysis. *statistich*, p. 872, 2012.

[11] Maubanu, E. ANALISIS KOMPONEN UTAMA UNTUK MENGATASI MULTIKOLINEARITAS PADA FAKTOR-FAKTOR YANG MEMPENGARUHI, 3(1), pp. 21–30, 2018.

[12] Neter, M. H., Wasserman, J., & Kutner, W. Applied Linier Regression Model. p. 561, 1988.

[13] Kosasih, R. Pengenalan Wajah Menggunakan PCA dengan Memperhatikan Jumlah Data Latih dan Vektor Eigen. *Jurnal Informatika Universitas Pamulang*, 6(1), 1, 2021, doi: 10.32493/informatika.v6i1.7261.

[14] Sriningsih, M., Hatidja, D., & Prang, J. D. Penanganan Multikolinearitas Dengan Menggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras Di Provinsi Sulut. *Jurnal Ilmiah Sains*, 18(1), p. 18, 2018, doi: 10.35799/jis.18.1.2018.19396.

[15] Hair, A. R. E. Multivariate Data Analysis. 2014.

[16] Putri, F. K. Pemodelan Persentase Angka Kematian Bayi di Kalimantan Barat dengan Metode Geographically Weighted Regression Principal Component Analysis (GWRPCA). 2021.

[17] Siburian, J. N. J. O., Rahmawati, R., & Hoyyi, A. Regresi Komponen Utama Robust S-Estimator Untuk Analisis Pengaruh Jumlah Pengangguran Di Jawa Tengah. *Jurnal Gaussian*, 8(4), 439–450, 2019, doi: 10.14710/j.gauss.v8i4.26724.

[18] Astuti, L. Analisis Angka Kematian Bayi (AKB) Di Kalimantan Barat Dengan Robust Principal Component Analysis (ROBPCA). 10(1), 61–70, 2021.

[19] Andriany, D., & Susanti, Y. Estimasi Parameter Regresi Robust dengan Metode Estimasi Least Trimmed Squares (LTS) pada Kematian Ibu di Indonesia, 2021.

[20] Setyowati, E., Akbarita, R., & Robby, R. Perbandingan Regresi Robust Metode Least Trimmed Square (Lts) dan Metode Estimasi-S pada Produksi Padi di Kabupaten Blitar. *Jurnal Matematika UNAND*, 10(3), 329–341, 2021.