Dipresentasikan pada SN-SMIAP 2019: Bandar Lampung 24 Oktober 2019

ANALISIS REGRESI KOMPONEN UTAMA *ROBUST* DENGAN METODE *MINIMUM COVARIANCE DETERMINANT – LEAST TRIMMED SQUARE* (MCD-LTS)

Khoirin Nisa(1)*, Siska Diah Ayu Larasati(1) , Eri Setiawan(1)

Jurusan Matematika, FMIPA,
Universitas Lampung, Bandar Lampung, 35145
Jl. Prof. Sumatri Brojonegoro No. 1 Bandar Lampung
**email korespondensi: khoirin.nisa@fmipa.unila.ac.id*

Diterima (Tgl Bulan Tahun), Direvisi (Tgl Bulan Tahun) (Times New Roman, 10pt)

Abstract. *Principal Component Regression (PCR) is a method used to overcome multicollinearity problems by reducing the dimensions of independent variables to obtain new simpler variables without losing most of the information contained in the variables. If the data analyzed contain outliers, a robust method on PCR is required. In this paper we use a robust method which is a combination of Robust Principal Component Analysis using the Minimum Covariance Determinant (MCD) method and Robust Regression Analysis using Least Trimmed Square (LTS) method. The purpose of this study is to examine the robust PCR analysis using the MCD-LTS method and to know the robustness of the method by looking at its sensitivity to outliers. For this purpose we compared the MCD-LTS PCR to the classic PCR based on the bias and Mean Square Error (MSE) values on several different sample sizes and percentages of outliers. The results of this study indicate that robust PCR using MCD-LTS is effective and efficient in overcoming the problem of multicollinearity and outliers in regression analysis.*

Keywords: *multicollinearity, outliers, principal component regression, robust*

Abstrak. Regresi Komponen Utama (RKU) merupakan metode yang digunakan untuk mengatasi masalah multikolinearitas dengan mereduksi dimensi variabel bebas sehingga diperoleh variabel baru yang lebih sederhana tanpa kehilangan sebagian besar informasi yang terkandung pada variabel bebasnya. Apabila pada data pengamatan terindikasi adanya pencilan, maka digunakan metode *robust* pada RKU. Dalam paper ini kami menggunakan metode *robust* yang merupakan kombinasi antara Analisis Komponen Utama *Robust* menggunakan metode *Minimum Covariance Determinant* (MCD) dengan Analisis Regresi *Robust* menggunakan metode *Least Trimmed Square* (LTS). Tujuan dari penelitian ini adalah mengkaji analisis RKU *robust* dengan metode MCD-LTS serta mengetahui ketegaran RKU *robust* dengan melihat kepekaannya terhadap pencilan kemudian dibandingkan dengan RKU klasik berdasarkan bias dan *Mean Square Error* (MSE) pada beberapa ukuran sampel dan persentase pencilan yang berbeda. Hasil dari penelitian ini menunjukkan bahwa RKU *robust* menggunakan MCD-LTS efektif dan efisien dalam mengatasi masalah multikolinearitas dan pencilan pada analisis regresi.

Kata kunci: multikolinearitas, pencilan, regresi komponen utama, *robust*

PENDAHULUAN

Analisis regresi merupakan salah satu metode analisis data yang digunakan untuk menyelidiki hubungan antara variabel respon dengan satu atau beberapa variabel bebas. Apabila hubungan yang diselidiki terdiri dari satu variabel respon dan satu variabel bebas maka disebut dengan analisis regresi linear sederhana. Sedangkan analisis regresi linear berganda terdiri dari satu variabel respon dan lebih dari satu variabel bebas.

Bentuk persamaan regresi dalam bentuk peubah asal *X* dapat ditulis sebagai berikut:

$$Y = \beta_0 + \beta_1 X_1 + \ ... + \beta_k X_k + \varepsilon \qquad (1)$$

dengan *Y* merupakan peubah tak bebas, X_i peubah bebas ke-*i (i = 1, 2, ..., k),* β_i adalah parameter-parameter regresi, dan ε merupakan galat.

Permasalahan yang sering dihadapi dalam analisis regresi linear berganda yaitu adanya multikolinearitas, yang mana terjadi ketika adanya korelasi yang kuat antara variabel bebas. Hal ini dapat menyebabkan $\mathbf{X^T X}$ memiliki kondisi buruk (*ill condition*) atau hampir singular yang pada akhirnya akan menyebabkan nilai penduga ragam bagi parameter regresi menjadi lebih besar [1].

Salah satu metode statistik yang digunakan untuk mengatasi masalah multikolinearitas adalah regresi komponen utama. Menurut Notiragayu & Nisa [2], pada RKU klasik terdapat dua tahap yaitu komponen utama dibentuk menggunakan vektor eigen dari matriks kovarian sampel (**S**) klasik dan diregresikan terhadap Y dengan metode kuadrat terkecil.

Peubah baru sebagai komponen utama (Q) adalah hasil transformasi dari peubah asal (X) yang modelnya dalam bentuk matriks yaitu $\boldsymbol{Q} = \boldsymbol{AX}$, dan komponen utama ke-*j* ditulis:

$$\begin{aligned} Q_j &= a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{kj}X_k \\ &= \boldsymbol{a_j^T X} \end{aligned} \qquad (2)$$

dimana vektor pembobot $\boldsymbol{a_j^T}$ diperoleh dengan memaksimumkan keragaman komponen utama ke-*j*, yaitu $\boldsymbol{S_{Q_j}^2} = \boldsymbol{a_j^T}\ \boldsymbol{S}\ \boldsymbol{a_j}$ dengan kendala $\boldsymbol{a_j^T a_j = 1}$ serta $\boldsymbol{a_h^T a_j = 0},$ untuk $h \neq j$. Vektor pembobot $\boldsymbol{a_j^T}$ diperoleh dari matriks kovarian $\boldsymbol{\Sigma}$ yang diduga dengan matriks $\boldsymbol{S}$, yaitu $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{\bar{x}})(\boldsymbol{x_i} - \boldsymbol{\bar{x}})^T$. Vektor $\boldsymbol{a_j^T}$ yang memenuhi kendala diatas adalah vektor eigen dari matriks kovarian $\boldsymbol{\Sigma}$. Model regresi komponen utama dapat ditulis sebagai berikut:

$$Y = \ \beta_0 + \ \beta_1 Q_1 + \ ... + \beta_j Q_j + \ \varepsilon \qquad (3)$$

dengan $j \leq k$.

Namun analisis seperti ini sangat sensitif terhadap pencilan *(outliers)* dan akan menghasilkan dugaan parameter yang bias akibat terpengaruh oleh data pencilan. Penelitian yang berkaitan dengan RKU *robust* telah dilakukan oleh para peneliti, diantaranya dapat dilihat pada [3-5].

Salah satu cara yang dapat dilakukan untuk mengatasi masalah pencilan dan multikolinieritas yaitu dengan meng-gunakan matriks kovarian *robust* pada analisis komponen utama dan metode *robust* pada tahap analisis regresi. Salah satu metode *robust* bagi matriks kovarian sampel (**S**) yaitu metode MCD dan kemudian hasil komponen-komponen utama *robust* yang terbentuk akan diregresikan dengan peubah respon menggunakan metode LTS.

Penduga MCD adalah pasangan $(\boldsymbol{\bar{X}}, \boldsymbol{S})$, dimana $\boldsymbol{\bar{X}}$ adalah vektor rata-rata dan $\boldsymbol{S}$ adalah matriks kovariansi yang meminimumkan nilai determinan $\boldsymbol{S}$ pada subsampel yang berisikan tepat sebanyak *h* anggota dari *n* pengamatan, dimana nilai standar dari *h*=[(n+k+1)/2]. Pada populasi dengan jumlah pengamatan yang kecil, penduga MCD dapat dengan cepat dihitung dan ditemukan, tetapi jika jumlah pengamatan besar, maka akan banyak sekali kombinasi subsampel dari *H* yang harus ditemukan dan penghitungan pun akan cukup memakan waktu. Dalam mengatasi keterbatasan ini, maka digunakan suatu algoritma baru untuk metode MCD yang dinamakan dengan metode *fast*-MCD [6].

Menurut Dayanti dkk. [7], langkah-langkah penduga MCD dalam menduga nilai koefisien regresi dengan menggunakan *fast*-MCD dimulai dengan mengambil subsampel pertama dari data pengamatan secara acak, misalkan subsampel tersebut H_1 dengan jumlah elemen sebanyak h. Selanjutnya menghitung vektor rata-rata $\bar{\boldsymbol{X}}_{MCD}$ dan matriks kovarians $\boldsymbol{S}_{MCD}$ dari H_1 dengan memisalkan $\bar{\boldsymbol{X}}_1$ dan $\boldsymbol{S}_1$ menggunakan persamaan berikut

$$\bar{\boldsymbol{X}}_{MCD} = \frac{1}{h}\sum_{i\epsilon\, H} x_i \qquad (4)$$

$$\boldsymbol{S}_{MCD} = \frac{1}{h}\sum_{i\epsilon\, H}[x_i - \bar{\boldsymbol{X}}_{MCD}][x_i - \bar{\boldsymbol{X}}_{MCD}]^T \qquad (5)$$

Kemudian menghitung determinan $\boldsymbol{S}_1$. Jika $det(\boldsymbol{S}_1) = 0$, maka berhenti. Namun apabila $det(\boldsymbol{S}_1) \neq 0$, maka dilanjutkan dengan menghitung jarak *robust* (*robust distance* = RD) untuk setiap pengamatan yang kemudian diurutkan dari yang terkecil hingga terbesar menggunakan persamaan berikut

$$RD_i = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{X}}_{MCD})^T \boldsymbol{S}_{MCD}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{X}}_{MCD})} \qquad (6)$$

Pada kasus subsampel selanjutnya, yaitu H_2 akan diambil sebanyak h pengamatan dengan jarak RD terkecil. Demikian seterusnya hingga mencapai konvergen $(S_{i+1}) = (S_1)$. Selanjutnya memilih himpunan H yang memiliki determinan $\boldsymbol{S}_{MCD}$ terkecil, serta mencari $\bar{\boldsymbol{X}}_{MCD}$ dan $\boldsymbol{S}_{MCD}$ dari himpunan H terpilih.

Menurut Nisa [8], metode LTS menduga koefisien regresi dari data yang mengandung pencilan dengan meminimumkan jumlah kuadrat galat terhadap sebaran data yang sudah terpangkas *(trimmed)* yang sering juga disebut dengan istilah "sebaran terwinsorkan" *(winsorized distribution).*

Metode LTS menduga koefisien regresi dengan menggunakan metode *Ordinary Least Square* (OLS) terhadap subhimpunan data berukuran h, yaitu:

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta}\left(\sum_{i=1}^{h} e_i^2\right)\\ &= \arg\min_{\beta}\left(\sum_{i=1}^{h}(y_i - \hat{y}_i)^2\right)\end{aligned} \qquad (7)$$

dengan $\frac{(3n+k+1)}{4} \leq h \leq n.$

Solusi $\hat{\beta}$ pada **Persamaan 7** dapat diperoleh dengan menggunakan turunan/differensial seperti pada penyelesaian pendugaan metode OLS, hanya pada LTS persamaan tersebut dihitung pada subhimpunan H terbaik dilakukan dengan menggunakan algoritma resampling dari seluruh kemungkinan subhimpunan yang dapat dibentuk yaitu sebanyak $\binom{n}{h}$. Subhimpunan H yang diperoleh merupakan sebaran data yang sudah terpangkas [9].

Berdasarkan hal tersebut, maka pada penelitian ini akan mengkaji analisis Regresi Komponen Utama (RKU) *Robust* dengan metode *Minimum Covariance Determinant- Least Trimmed Square* (MCD-LTS) serta mengetahui ketegaran metode tersebut terhadap pencilan.

METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data simulasi. Pada data simulasi akan dibangkitkan data menggunakan *software* SAS 9.4 dengan variabel bebas sebanyak k=6 dan galat berdistribusi normal $N(0,1)$ sehingga Y merupakan kombinasi linear dari k variabel bebas ditambah galat. Masing- masing data dibangkitan dengan ukuran sampel (n = 20, 60, 100, 200) serta 5 jenis persentase pencilan (5%, 10% 15%, 20%, 25%) dan diulang sebanyak 1000 kali. Kemudian pencilan dibangkitkan dari distribusi normal dengan mean 50 dan simpangan baku 0,05 yaitu $\varepsilon^{*} \sim N(50,(0.05)^2)$.

Untuk mendapatkan data kolinearitas pada setiap himpunan data, X_{ik} dibangkitkan menggunakan simulasi Monte Carlo berdasarkan McDonald & Galarneau [10] dengan persamaan sebagai berikut:

$$X_{ij} = (1-\rho^2)^{\frac{1}{2}} x_{ij} + \rho x_{i(k+1)} \qquad (8)$$

dimana :
$i = 1, 2, \dots, n$
$j = 1, 2, \dots, k$

dengan $x_{i1}, x_{i2}, \dots, x_{i(k+1)}$ merupakan data yang dibangkitkan berdistribusi normal N(0,1) dan ρ ditentukan sehingga korelasi antarvariabel bebas diberikan oleh ρ^2. Dua himpunan dari variabel yang saling berkorelasi dalam penelitian ini dibuat berdasarkan nilai $\rho = 0.99$.

Adapun tahapan yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Melakukan simulasi data.
2. Melakukan regresi komponen utama klasik dengan langkah sebagai berikut:
   a. Menghitung matriks kovarian dari variabel asal (X).
   b. Menghitung nilai eigen λ_i dan vektor eigen a dari matriks kovarian.
   c. Menghitung skor komponen utama Q.
   d. Memilih komponen utama yang memiliki nilai eigen lebih dari 1.
   e. Menghitung nilai duga koefisien regresi komponen utama berdasarkan komponen yang terpilih dengan metode OLS, simpan sebagai $\hat{\beta}^{(0)}$.
3. Mencatat nilai duga β pada regresi komponen utama klasik.
4. Membangkitkan sebuah matriks *noise* dari distribusi $N(0,(0.01)^2)$.
5. Membangkitkan matriks pencilan dari distribusi $N(50,(0.05)^2)$ sehingga diperoleh matriks kontaminasi. Elemen dari matriks kontaminasi adalah nol kecuali untuk beberapa elemen yang dijadikan pencilan.
6. Menambahkan matriks *noise* dan matriks kontaminasi pada data simulasi. Sehingga diperoleh matriks untuk data yang telah terkontaminasi pencilan (XYk).
7. Melakukan regresi komponen utama klasik seperti langkah 2 pada data yang telah terkontaminasi pencilan.
8. Mencatat nilai duga β pada regresi komponen utama klasik di langkah 7.
9. Melakukan regresi komponen utama *robust* dengan langkah sebagai berikut:
10. Menghitung matriks kovarian MCD dari data yang terkontaminasi pencilan.
11. Menghitung nilai eigen λ_i dan vektor eigen *a* dari matriks kovarian MCD.
    a. Menghitung skor komponen utama Q.
    b. Memilih komponen utama yang memiliki nilai eigen lebih dari 1.
    c. Menghitung nilai duga koefisien regresi komponen utama berdasarkan komponen yang terpilih dengan *Least Trimmed Square* (LTS).
12. Mencatat nilai duga β pada regresi komponen utama *robust*.
13. Ulangi langkah 4 sampai 10 sebanyak 1000 kali untuk seluruh jumlah data.
14. Menghitung nilai bias dan *Mean Square Error* (MSE) dari RKU klasik dan RKU *robust* dengan menggunakan rumus sebagai berikut:

$$\text{Bias}\left(\hat{\beta}_j\right) = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{\beta}_{ij}^{(s)} - \hat{\beta}_j^{(0)}\right|$$

$$\text{MSE}\left(\hat{\beta}_j\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\beta}_{ij}^{(s)} - \hat{\beta}_j^{(0)}\right)^2$$

dimana,
j = 0, 1, ..., *k*
m = 1000

15. Menganalisa dan membandingkan hasil perhitungan RKU klasik dengan RKU *robust* berdasarkan nilai rata-rata bias dan rata-rata MSE dari seluruh dugaan koefisien regresi yang dihasilkan.

HASIL DAN PEMBAHASAN

Analisis Regresi Komponen Utama *Robust*

RKU klasik tanpa pencilan dan dengan pencilan pada penelitian ini tetap digunakan. RKU klasik membentuk komponen utama menggunakan vektor eigen dari matriks kovarian klasik yang kemudian diregresikan menggunakan metode OLS. Metode RKU klasik tanpa pencilan menjadi acuan sebagai nilai koefisien RKU sebenarnya, sedangkan RKU klasik dengan pencilan menjadi acuan untuk mengetahui tingkat ketegaran RKU *robust.*

$Q_1, Q_2, \dots, Q_j$ disebut komponen utama dari X. Jika matriks kovarian dari variabel asal X_k dilambangkan dengan $\boldsymbol{\Sigma}$, maka diperoleh varian komponen utama yaitu:

$$\begin{aligned} var\,(Q_j) &= E[(Q_j - E(Q_j))(Q_j - E(Q_j))^T] \\ &= E[(\boldsymbol{a}_j^T\boldsymbol{X} - E(\boldsymbol{a}_j^T\boldsymbol{X}))(\boldsymbol{a}_j^T\boldsymbol{X} - E(\boldsymbol{a}_j^T\boldsymbol{X}))^T] \\ &= E[(\boldsymbol{a}_j^T\boldsymbol{X} - \boldsymbol{a}_j^T\,E(\boldsymbol{X}))(\boldsymbol{a}_j^T\boldsymbol{X} - \boldsymbol{a}_j^T\,E(\boldsymbol{X}))^T] \\ &= E[(\boldsymbol{a}_j^T\boldsymbol{X} - \boldsymbol{a}_j^T\boldsymbol{\mu})(\boldsymbol{a}_j^T\boldsymbol{X} - \boldsymbol{a}_j^T\boldsymbol{\mu})^T] \\ &= E[(\boldsymbol{a}_j^T\boldsymbol{X} - \boldsymbol{a}_j^T\boldsymbol{\mu})(\boldsymbol{X}^T\boldsymbol{a}_j - \boldsymbol{\mu}^T\boldsymbol{a}_j)] \\ &= \boldsymbol{a}_j^T E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T]\boldsymbol{a}_j \\ &= \boldsymbol{a}_j^T\boldsymbol{\Sigma}\,\boldsymbol{a}_j \end{aligned} \quad (9)$$

Komponen utama ke-*j* yang merupakan kombinasi linear terbobot variabel asal bertujuan untuk memaksimalkan $var\,(Q_j)$ dan tidak berkorelasi dengan komponen utama yang lain. Oleh karena itu, Q_j harus memenuhi batasan $\boldsymbol{a}_j^T\boldsymbol{a}_j = 1$, $\boldsymbol{a}_i^T\boldsymbol{a}_j = 0\ ;\ i < j$ [11].

Untuk mencari komponen utama pertama, perlu dipertimbangkan $\boldsymbol{a}_1^T\boldsymbol{X}$ sebagaimana mencari $\boldsymbol{a}_1$ yang dapat memaksimumkan $var\,(Q_1) = \boldsymbol{a}_1^T\boldsymbol{\Sigma}\boldsymbol{a}_1$ dengan kendala $\boldsymbol{a}_1^T\boldsymbol{a}_1 = 1$. Pendekatan standarnya adalah dengan teknik pengganda Lagrange atau dalam hal ini akan memaksimumkan fungsi $max\,f(\boldsymbol{a}_1) = \boldsymbol{a}_1^T\boldsymbol{\Sigma}\boldsymbol{a}_1$ dengan kendala $\boldsymbol{a}_1^T\boldsymbol{a}_1 = 1$.
$max\,f(\boldsymbol{a}_1) = \boldsymbol{a}_1^T\boldsymbol{\Sigma}\boldsymbol{a}_1 - \lambda(\boldsymbol{a}_1^T\boldsymbol{a}_1 - 1)$, dimana λ adalah pengganda lagrange. Maka dapat dikaji sebagai berikut:

$$\begin{aligned} \frac{\partial f(\boldsymbol{a}_1)}{\partial \boldsymbol{a}_1} &= 2\boldsymbol{\Sigma}\boldsymbol{a}_1 - 2\lambda\boldsymbol{a}_1 + 0 = 0 \\ \boldsymbol{\Sigma}\boldsymbol{a}_1 - \lambda\,\boldsymbol{a}_1 &= 0 \\ \boldsymbol{a}_1^T\boldsymbol{\Sigma}\boldsymbol{a}_1 - \boldsymbol{a}_1^T\lambda\,\boldsymbol{a}_1 &= 0 \\ var\,(Q_1) - \boldsymbol{a}_1^T\lambda\,\boldsymbol{a}_1 &= 0 \\ var\,(Q_1) &= \boldsymbol{a}_1^T\lambda\,\boldsymbol{a}_1 \\ var\,(Q_1) &= \lambda\,\boldsymbol{a}_1^T\boldsymbol{a}_1 \\ var\,(Q_1) &= \lambda\,.\,1 \\ var\,(Q_1) &= \lambda \end{aligned}$$

Sehingga diperoleh $var\,(Q_1)$ merupakan nilai eigen dari matriks kovarian $\boldsymbol{\Sigma}$.

Pada persamaan $\boldsymbol{\Sigma}\boldsymbol{a}_1 - \lambda\,\boldsymbol{a}_1 = 0$ juga dapat disederhanakan menjadi:
$\boldsymbol{\Sigma}\boldsymbol{a}_1 - \lambda\,\boldsymbol{a}_1 = 0$
$(\boldsymbol{\Sigma} - \lambda\boldsymbol{I})\boldsymbol{a}_1 = 0$
Maka diperoleh $\boldsymbol{\Sigma} - \lambda\boldsymbol{I} = 0$ atau $|\boldsymbol{\Sigma} - \lambda\boldsymbol{I}| = 0$ dan juga $\boldsymbol{a}_1 = 0$ sehingga $\boldsymbol{a}_1$ merupakan vektor eigen dari matriks kovarian $\boldsymbol{\Sigma}$ yang berpadanan dengan nilai eigen. Berdasarkan hal tersebut, telah diperoleh $var\,(Q_1)$ yang merupakan nilai eigen dari matriks kovarian $\boldsymbol{\Sigma}$, maka $\boldsymbol{a}_1$ adalah vektor eigen dari $\boldsymbol{\Sigma}$ yang berpadanan dengan nilai eigen terbesar pertama.

Seperti yang telah dinyatakan diatas, dapat ditunjukkan bahwa komponen utama kedua (Q_2), ketiga (Q_3), keempat (Q_4),... (Q_j), maka vektor eigen $\boldsymbol{a}_{1j}, \boldsymbol{a}_{2j}, \dots, \boldsymbol{a}_{kj}$ adalah vektor eigen dari matriks kovarian $\boldsymbol{\Sigma}$ yang berpadanan dengan nilai eigen terbesar

$\lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_p$ dan selanjutnya $var(\boldsymbol{a^T X}) = \lambda_p$, untuk p = 1, 2,.., k.

Pada analisis komponen utama (AKU) *robust*, untuk menduga matriks kovarian yang *robust* maka dengan menggunakan metode MCD. Pada prinsipnya metode MCD adalah mencari himpunan bagian yang anggotanya sebanyak h elemen dari matriks **X** dengan h merupakan bilangan bulat terkecil dari $\frac{n+k+1}{2}$. Misalkan himpunan bagian itu adalah H_1 maka terdapat sebanyak C_h^n kombinasi yang harus ditemukan untuk mendapatkan dugaan vektor rataan dan matriks kovarian. Apabila n yang digunakan kecil, pendugaan MCD mudah dan relatif lebih cepat untuk ditemukan. Tetapi, apabila n besar. Maka akan banyak kombinasi subhimpunan yang harus diperoleh untuk mendapatkan pendugaan MCD tersebut. Untuk mengatasi keterbatasan ini, maka digunakan pendekatan FAST-MCD yang dikembangkan oleh Rousseeuw dan Van Driessen [6].

Misalkan terdapat $X_k = [x_1, x_2, \ldots, x_k]$ merupakan himpunan data sejumlah n pengamatan dari k peubah. Misalkan $H_1 \subset \{1,2,\ldots,n\}$ dengan $|H_1|$ = h, maka hitung vektor rata-rata seperti pada **Persamaan 4**.

$$\overline{\boldsymbol{X}}_{MCD1} = \frac{1}{h}\sum_{i\epsilon\, H} \boldsymbol{x_i}$$

dan juga hitung matriks kovarian MCD seperti pada **Persamaan 5**:

$$\boldsymbol{S}_{MCD1} = \frac{1}{h}\sum_{i\epsilon\, H}[\boldsymbol{x_i} - \overline{\boldsymbol{X}}_{MCD}][\boldsymbol{x_i} - \overline{\boldsymbol{X}}_{MCD}]^T$$

Jika $\det(\mathbf{S_1}) \neq 0$, maka definisikan jarak *robust* seperti pada **Persamaan 6**:

$$RD_i = \sqrt{(\boldsymbol{x}_i - \overline{\boldsymbol{X}}_{MCD1})^T \boldsymbol{S}_{MCD1}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{X}}_{MCD1})}$$

dengan i=1,2, ..., n. Kemudian mengurutkan nilai RD mulai dari yang terkecil hingga terbesar. Selanjutnya ambil H_2 sejumlah h pengamatan dengan jarak terkecil kemudian hitung $\overline{\boldsymbol{X}}_{MCD2}$ dan $\boldsymbol{S}_{MCD2}$ dari himpunan H_2.

Penjelasan MCD mensyaratkan $\det(\boldsymbol{S_{MCD1}}) \neq 0$, karena jika $\det(\boldsymbol{S_{MCD1}}) = 0$ maka nilai objektif minimum untuk mendapatkan determinan terkecil telah ditemukan. Proses pada metode MCD akan berhenti jika ditemukan himpunan bagian yang konvergen $det(\boldsymbol{S_{i+1}})$ = $det(\boldsymbol{S_i})$. Sehingga memilih subsampel H yang memiliki nilai determinan matriks kovarians terkecil. Kemudian dari subsampel yang terpilih akan dicari nilai $\overline{\boldsymbol{X}}_{MCD}$ dan $\boldsymbol{S}_{MCD}$. Berdasarkan hal tersebut, maka untuk AKU *robust* pada penelitian ini akan diperoleh vektor eigen dari matriks kovarian MCD yang berpadanan dengan nilai eigen terbesar pertama dari matriks kovarian MCD.

Adapun pada penelitian ini hanya menggunakan satu komponen utama. Sehingga langkah selanjutnya akan diregresikan antara komponen utama tersebut dengan variabel tak bebas Y yang dimaksudkan untuk menduga nilai koefisien regresi $\hat{\beta}_0$ dan $\hat{\beta}_1$.

Solusi $\hat{\beta}$ dengan metode LTS dapat diperoleh dengan menggunakan turunan/differensial seperti pada penyelesaian pendugaan metode OLS, hanya pada LTS persamaan tersebut dihitung pada subhimpunan H terbaik dilakukan dengan menggunakan algoritma resampling dari seluruh kemungkinan subhimpunan yang dapat dibentuk yaitu sebanyak $\binom{n}{h}$ dimana $h = \frac{3n+k+1}{4}$. Subhimpunan H yang diperoleh merupakan sebaran data yang sudah terpangkas.

Akan ditunjukkan sebagai berikut:

$$J = \sum_{i=1}^{h} \varepsilon_i^2 = \sum_{i=1}^{h}(y_i - \beta_0 - \beta_1 Q_i)^2 \quad (10)$$

Untuk mendapatkan nilai β_0 dan β_1 pada **Persamaan 10** maka J diturunkan terhadap β_0 dan β_1 kemudian menyamakannya dengan nol. Jika J diturunkan terhadap β_0 maka diperoleh:

$$\frac{\partial J}{\partial \beta_0} = -2\sum_{i=1}^{h}(y_i - \beta_0 - \beta_1 Q_i) = 0$$

$$\sum_{i=1}^{h} y_i - h\beta_0 - \beta_1 \sum_{i=1}^{h} Q_i = 0$$

$$\beta_0 = \frac{\sum_{i=1}^{h} y_i - \beta_1 \sum_{i=1}^{h} Q_i}{h}$$

dengan $\overline{y} = \frac{\sum_{i=1}^{h} y_i}{h}$ dan $\overline{Q} = \frac{\sum_{i=1}^{h} Q_i}{h}$
maka diperoleh:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{Q} \qquad (11)$$

Selanjutnya J diturunkan terhadap β_1 dan samakan dengan nol,

$$\frac{\partial J}{\partial \beta_1} = -2\sum_{i=1}^{h}(y_i - \beta_0 - \beta_1 Q_i)Q_i = 0$$

$$\sum_{i=1}^{h} y_i Q_i - \beta_0 \sum_{i=1}^{h} Q_i - \beta_1 \sum_{i=1}^{h} {Q_i}^2 = 0$$

$$\sum_{i=1}^{h} y_i Q_i - (\overline{y} - \beta_1 \overline{Q}) \sum_{i=1}^{h} Q_i - \beta_1 \sum_{i=1}^{h} {Q_i}^2 = 0$$

$$\sum_{i=1}^{h} y_i Q_i - \overline{y} \sum_{i=1}^{h} Q_i + \beta_1 \overline{Q} \sum_{i=1}^{h} Q_i - \beta_1 \sum_{i=1}^{h} {Q_i}^2 = 0$$

$$\sum_{i=1}^{h} y_i Q_i - \overline{y} \sum_{i=1}^{h} Q_i = \beta_1 \left[\sum_{i=1}^{h} {Q_i}^2 - \overline{Q} \sum_{i=1}^{h} Q_i\right]$$

$$\beta_1 = \frac{\sum_{i=1}^{h} y_i Q_i - \frac{1}{h}\sum_{i=1}^{h} y_i \sum_{i=1}^{h} Q_i}{\sum_{i=1}^{h} {Q_i}^2 - \frac{1}{h}\sum_{i=1}^{h} Q_i \sum_{i=1}^{h} Q_i}$$

$$\beta_1 = \frac{\sum_{i=1}^{h} y_i Q_i - \frac{1}{h}\sum_{i=1}^{h} y_i \sum_{i=1}^{h} Q_i}{\sum_{i=1}^{h} {Q_i}^2 - \frac{1}{h}\sum_{i=1}^{h} (Q_i)^2}$$

Sehingga dapat disederhanakan menjadi:

$$\hat{\beta}_1 = \frac{h\sum_{i=1}^{h} y_i Q_i - \sum_{i=1}^{h} y_i \sum_{i=1}^{h} Q_i}{h\sum_{i=1}^{h} {Q_i}^2 - \sum_{i=1}^{h} (Q_i)^2} \qquad (12)$$

Jadi, $\hat{\beta}_0$ dan $\hat{\beta}_1$ pada **Persamaan 11** dan **Persamaan 12** merupakan penduga bagi nilai koefisien RKU *robust* yaitu menggunakan metode MCD-LTS.

Hasil Simulasi

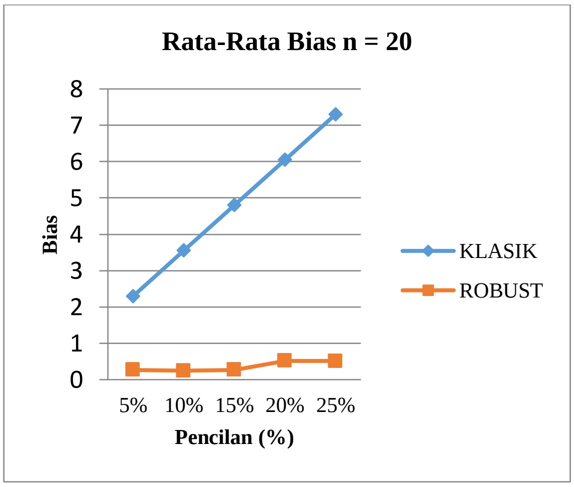
Untuk mempermudah dalam interpretasi data, maka penulis menggunakan nilai rata-rata bias dan rata-rata MSE yang akan disajikan pada tabel berikut.

Tabel 1. Rata-Rata Bias dan MSE Pendugaan Koefisien Regresi Komponen Utama

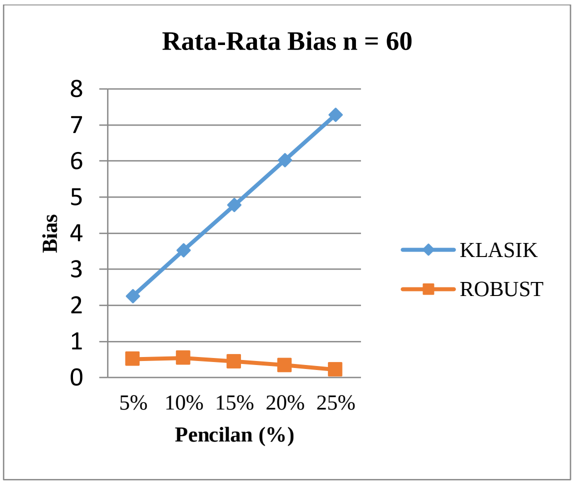
n	Pencilan (%)	Rata-Rata Bias		Rata-Rata MSE	
		RKU Klasik	RKU *Robust*	RKU Klasik	RKU *Robust*
20	5	2,2985	0,2673	5,3236	0,2301
	10	3,557	0,2459	14,7344	0,148
	15	4,8047	0,2688	30,3489	0,1727
	20	6,0513	0,5145	52,2118	0,4211
	25	7,3025	0,5111	80,3393	0,4713
60	5	2,2633	0,5075	5,1785	0,4866
	10	3,5315	0,5388	14,6277	0,4946
	15	4,7819	0,4392	30,2556	0,3221
	20	6,0328	0,3416	52,1328	0,1981
	25	7,2864	0,2159	80,2717	0,0882
100	5	2,2817	0,1531	5,254	0,0474
	10	3,5458	0,1026	14,6875	0,0284
	15	4,7951	0,0435	30,3097	0,02
	20	6,0449	0,2353	52,1833	0,0893
	25	7,2975	0,238	80,3188	0,0925
200	5	2,2725	0,1806	5,2161	0,0641
	10	3,522	0,0525	14,5889	0,0121
	15	4,7766	0,0837	30,2328	0,0124
	20	6,0278	0,1374	52,1125	0,0287
	25	7,2747	0,1429	80,2252	0,0224

Pada **Tabel 1** terlihat bahwa nilai rata-rata bias dan MSE dari data mengandung pencilan 5% - 25% yang dihasilkan dari metode RKU klasik lebih besar dibandingkan RKU *robust*. Pada RKU klasik, terlihat juga bahwa setiap penambahan persentase dalam satu ukuran sampel maka selalu terjadi peningkatan nilai rata-rata bias dan MSE, bahkan diperoleh nilai rata-rata bias dan MSE > 1 di masing-masing persentase pencilan dan ukuran sampel tersebut. Hal ini dapat diartikan nilai dugaan koefisien RKU klasik semakin buruk. Sedangkan pada RKU *robust* rata-rata bias dan MSE yang dihasilkan selalu lebih kecil atau dibawah nilai rata-rata bias dan MSE RKU klasik yaitu nilai rata-rata bias dan MSE < 1 di masing-masing persentase pencilan dan ukuran sampel tersebut.

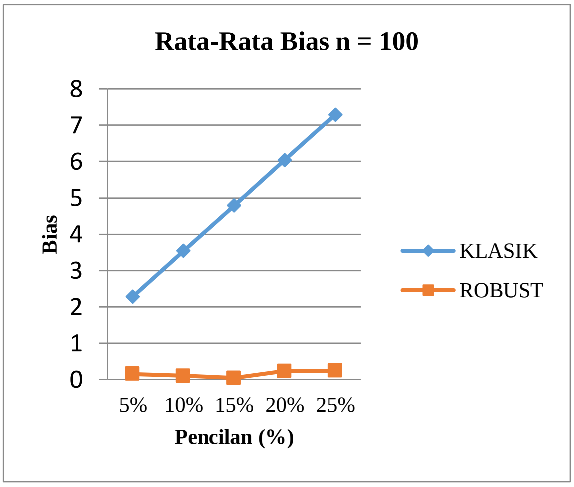
Berdasarkan nilai rata-rata bias dan MSE dengan ukuran sampel yaitu 20, 60, 100, dan 200 serta persentase pencilan yaitu 5%, 10%, 15%, 20%, dan 25% yang disajikan pada **Tabel 1**, maka akan diperoleh grafik perbandingan dari metode RKU klasik dan *robust* sebagai berikut:



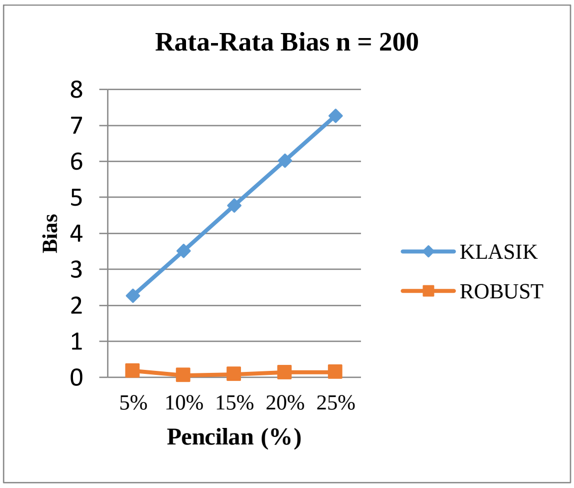
Gambar 1. Grafik Nilai Rata-Rata Bias untuk n = 20.



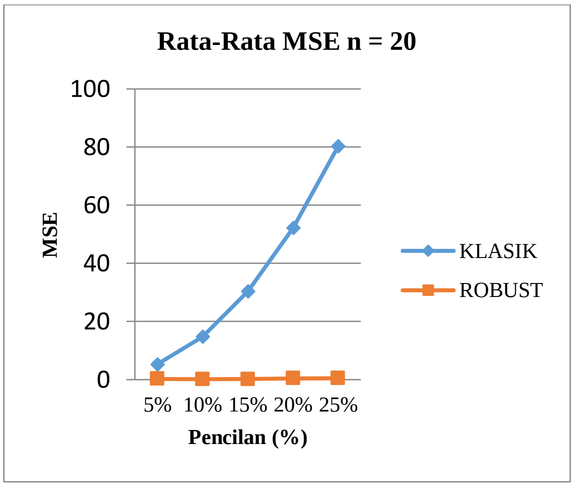
Gambar 2. Grafik Nilai Rata-Rata Bias untuk n = 60.



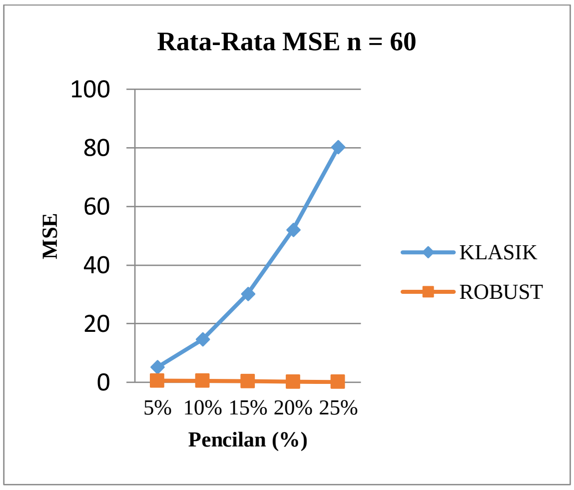
Gambar 3. Grafik Nilai Rata-Rata Bias untuk n = 100.



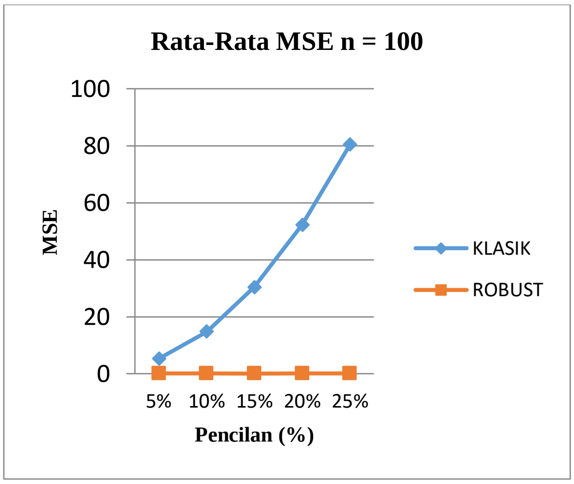
Gambar 4. Grafik Nilai Rata-Rata Bias untuk n = 200.



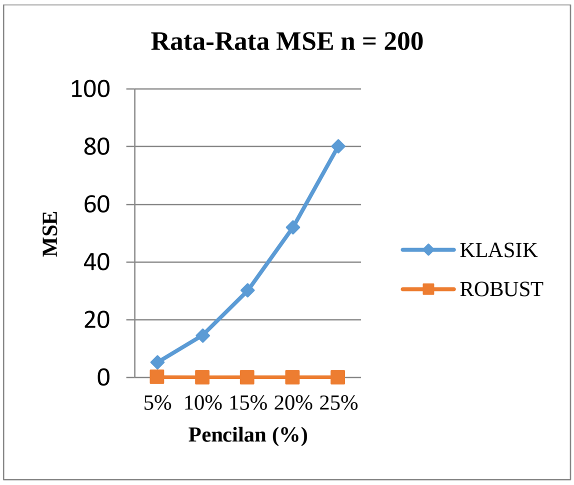
Gambar 5. Grafik Nilai Rata-Rata MSE untuk n = 20.



Gambar 6. Grafik Nilai Rata-Rata MSE untuk n = 60.



Gambar 7. Grafik Nilai Rata-Rata MSE untuk n = 100.



Gambar 8. Grafik Nilai Rata-Rata MSE untuk n = 200.

Berdasarkan **Gambar 1 – Gambar 4**, terlihat bahwa letak garis nilai rata-rata bias RKU klasik dengan ukuran sampel 20, 60, 100, 200 berada diatas garis nilai rata-rata bias RKU *robust*. **Gambar 1 – Gambar 4** juga terlihat bahwa titik-titik yang disambungkan garis untuk nilai rata-rata bias pada RKU klasik di setiap persentase pencilan selalu naik secara konstan sedangkan pada RKU *robust* titik-titik yang disambungkan garis untuk nilai rata-rata bias di setiap persentase pencilan tidak secara konstan naik atau turun. Selanjutnya, berdasarkan **Gambar 5 – Gambar 8** terlihat bahwa letak garis nilai rata-rata MSE RKU klasik dengan masing-masing ukuran sampel tersebut berada diatas garis nilai rata-rata MSE RKU *robust*. **Gambar 5 – Gambar 8** juga terlihat bahwa titik-titik yang disambungkan garis untuk nilai rata-rata MSE pada RKU klasik di setiap persentase pencilan selalu naik secara konstan yaitu berkisar dari angka > 5 hingga angka ≥ 80 sedangkan pada RKU *robust* titik-titik yang disambungkan garis untuk nilai rata-rata MSE di setiap persentase pencilan tidak secara konstan naik atau turun, dapat dikatakan hanya disekitar angka $< 0,5$.

KESIMPULAN

Berdasarkan hasil dan pembahasan, maka diperoleh kesimpulan sebagai berikut:

1. Analisis Komponen Utama *Robust* dapat dilakukan dengan menggunakan matriks kovarian *robust MCD* dan metode regresi *robust* LTS dengan penduga nilai koefisien RKU *robust* sebagai berikut:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{h} y_i - \beta_1 \sum_{i=1}^{h} Q_i}{h}$$

$$\hat{\beta}_1 = \frac{h\sum_{i=1}^{h} y_i Q_i - \sum_{i=1}^{h} y_i \sum_{i=1}^{h} Q_i}{h\sum_{i=1}^{h} {Q_i}^2 - \sum_{i=1}^{h} (Q_i)^2}$$

dimana,

Q = Komponen Utama

h = Banyaknya anggota dalam subhimpunan terbaik dari data yang terpangkas.

2. Pada hasil simulasi diperoleh bahwa pada suatu ukuran sampel dengan presentase pencilan 5%-25%, nilai rata-rata bias dan MSE metode RKU klasik selalu naik secara konstan disetiap penambahan presentase pencilan sedangkan metode RKU *robust* diperoleh nilai rata-rata bias yang lebih kecil dibandingkan RKU klasik. Hal ini menunjukkan bahwa metode RKU *robust* merupakan metode yang kekar terhadap pencilan, sedangkan metode RKU klasik sangat sensitif terhadap danya pencilan. Dengan demikian maka RKU *robust* menggunakan MCD-LTS memberikan hasil yang baik dan merupakan metode efektif dan efisien dalam mengatasi masalah multikolinearitas dan pencilan.

DAFTAR PUSTAKA

[1] N.R. Draper dan H. Smith, *Applied Regression Analysis*, 3rd Edition, Wiley India, 2011.

[2] Notiragayu dan K.Nisa, "Analisis Regresi Komponen Utama *Robust* untuk Data Mengandung Pencilan", *Jurnal Sains MIPA,* vol. 14, no. 1, pp. 45-50, 2008

[3] M.N. Abdel Bary, "Robust Regression Diagnostic for Detecting and Solving Multicollinearity and Outlier Problems: Applied Study by Using Financial Data", *Applied Mathematical Sciences*, vol. 11, no. 13, pp 601 – 622, 2017.

[4] A. Agarwal, D. Shah, D. Shen D. Song, "On Robustness of Principal Component Regression", ArXiv preprint1–57, 2019.

[5] T-H. Oh, Y. Matsushita, I.S. Kweon, D. Wipf, "A Pseudo-Bayesian Algorithm for Robust PCA", proceeding of 30th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 2016.

[6] P.J. Rousseeuw dan P.J., K. Van Driessen, K, "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, vol. 41, no. 3, pp. 212-223, 1999.

[7] N. P. Dayanti, N. L. Suciptawati, N.L. dan M. Susilawati, "Penerapan Bootstrap dalam Metode Minimum Covariance Determinant (MCD) dan Least Median Square (LMS) pada Analisis Regresi Linear Berganda", *E-Jurnal Matematika*, vol. 5 no. 1, pp. 22-26, 2016.

[8] K. Nisa, "Analisis Regresi Robust Menggunakan Metode *Least Trimmed Square* untuk Data Mengandung Pencilan", *Jurnal Ilmiah MIPA*, vol. IX, no. 2, pp.1-9, 2006.

[9] C. Chen, *Robust Regression and Outlier Detection with the ROBUSTREG procedure*, SAS Institue : Cary, NC, pp. 265-27, 2002.

[10] McDonald, G.C. & Galarneau, D.I., "Amonte Carlo Evaluation of Some Ridge-Type Estimators", *J. Amer. Statist. Asoc,* vol. 70, pp. 407-416, 1975.

[11] I.T. Jolliffe, Proncipal Component Analysis second edition, Springer, 2002.