PEBANDINGAN METODE ROBUST *MCD-LMS, MCD-LTS, MVE-LMS,* DAN *MVE-LTS* DALAM ANALISIS REGRESI KOMPONEN UTAMA

Sekar Wulandari, Nur Salam, dan Dewi Anggraini
Program Studi Matematika
Universitas Lambung Mangkurat
Jl. A. Yani Km 36,800 Kampus Unlam Banjarbaru

ABSTRAK

Regresi komponen utama (*Principal Component Regression)* merupakan teknik statistik yang digunakan untuk analisis regresi dengan kolinieritas. Teknik *robust* pada regresi komponen utama sangat diperlukan jika termuat *outlier* didalam data.

Pada penelitian ini dilakukan kombinasi antara Analisis Komponen Utama (*PCA*) *Robust*: *Minimum Covariant determinant* (*MCD*) dan *Minimum Volume Ellipsoid* (*MVE*) dengan Metode Regresi *Robust*: *Least Median Square* (*LMS*) dan *Least Trimmed Square* (LTS), kemudian membandingkan tingkat resistensi metode *MCD-LMS, MCD-LTS, MVE-LMS, MVE-LTS* terhadap *outlier* dengan membandingkan nilai Bias dan *MSE* (*Means Square Error)* pada beberapa ukuran sampel dan persentase *outlier* yang berbeda.

Hasil yang diperoleh menunjukkan bahwa metode *MCD-LMS* lebih baik dari pada metode *MCD-LTS, MVE-LMS, dan MVE-LTS* karena memiliki nilai Bias dan *MSE* yang minimum.

Kata Kunci: *Regresi Robust, MCD, MVE, LMS, LTS*.

ABSTRACT

Principal Component Regression (PCR) is one of the widely used statistical techniques for regression analysis with colinearity. A robust technique on CR required is when data contains outlier is urgently needed.

In this research we consider combination between Robust Principal Ccomponent Analysis (PCA): Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) with Robust Regression methods: Least Median Square (LMS), and Least Trimmed Square (LTS), then compare resistance level of MCD-LMS, MCD-LTS, MVE-LMS and MVE-LTS through the bias and the mean square error on some samples size and outlier's percentage.

The result shows that the MCD-LMS perform better than MCD-LTS, MVE-LMS, and MCD-LTS.

Keywords: *Robust Regression, MCD, MVE, LTS, LMS*.

1. PENDAHULUAN

Regresi komponen utama (RKU) merupakan salah satu metode yang digunakan untuk mengatasi masalah multikolinearitas. Metode ini mengatasi multikolinearitas dengan cara membentuk komponen-komponen utama yang saling bebas dari variabel bebasnya. Selanjutnya komponen-komponen utama yang terbentuk diregresikan dengan peubah respon. Dalam analisis komponen utama klasik, perhitungannya didasarkan pada matriks kovarian (S).

Matriks kovarian ini akan optimal jika data berasal dari suatu distribusi normal multivariat, tetapi sangat sensitif terhadap adanya *outlier* (pencilan), terutama jika data mengandung *outlier* yang ekstrim yang mengakibatkan distribusi data menjadi sangat menjulur (*heavy tailed distribution*), pada kasus seperti ini S akan kehilangan efisiensinya [1].

Untuk mengatasi masalah *outlier* diperlukan suatu metode penduga yang resisten terhadap *outlier* yang disebut sebagai metode *Robust*. Metode *Robust* bagi S yang digunakan adalah metode *Minimum Covariance Determinant* (*MCD*) dan metode *Minimum Volume Elipsoid* (*MVE*). Selanjutnya hasil komponen-komponen utama *Robust* yang terbentuk diregresikan dengan peubah respon menggunakan metode *OLS* (*Ordinary Least Square*).

Metode *OLS* dikenal sebagai metode penduga terbaik dalam analisis regresi, namun metode ini sangat peka terhadap adanya penyimpangan asumsi pada data. Jika data tidak memenuhi salah satu asumsi regresi maka penduga *OLS* tidak lagi efisien. Salah satu asumsi penting dalam analisis regresi yang berkaitan dengan inferensia model adalah asumsi sebaran normal (normalitas). Asumsi normalitas seringkali dilanggar saat data mengandung *outlier*. Jika terdapat *outlier* dalam data, maka bentuk sebaran data tidak lagi simetrik tetapi cenderung menjulur ke arah *outlier* sehingga melanggar asumsi normalitas. Dalam kasus seperti ini, analisis regresi *Robust* merupakan metode yang layak untuk digunakan.

Sampai saat ini berbagai metode *Robust* untuk analisis regresi terus berkembang dan digunakan dalam berbagai bidang, diantaranya adalah *Least Median Square* (*LMS*) dan *Least Trimmed Square* (*LTS*). Metode *LMS* menduga koefisien regresi dari data yang mengandung *Outlier* dengan meminimumkan median dari kuadrat galatnya $\left\{\min median\left(e_i^2\right)\right\}$, sedangkan metode *LTS* dengan melakukan analisis regresi kuadrat terkecil $\left\{\min\left(e_i^2\right)\right\}$ terhadap sebaran data yang sudah terpotong (*trimmed*).

Berdasarkan uraian diatas, penelitian ini akan membandingkan tingkat resistensi antara metode *MCD-LMS, MCD-LTS, MVE-LMS* dan *MVE-LTS* sebagai metode RKU *Robust* dengan menggunakan nilai Bias dan *MSE* (*Means Square Error)* pada beberapa ukuran sampel dan persentase *outlier*.

2. TINJAUAN PUSTAKA

2.1 Data *Outlier*

Outlier ialah data yang tidak mengikuti pola umum model atau lebih jauh dari rata-rata sisaannya (*error*).

2.2 Bias

Bias penduga dari suatu parameter pada simulasi data didefinisikan sebagai jumlah selisih dari penduga parameter pada data yang terdapat outlier dengan penduga parameter pada data yang tanpa *outlier*, dibagi dengan banyaknya perulangan. Hal ini dinotasikan sebagai berikut:

$$Bias(\hat{\beta}_k) = \frac{1}{m}\sum_{s=1}^{m}\left(\hat{\beta}_k{}^{(s)} - \hat{\beta}_k{}^{(0)}\right), \quad \text{k=1, 2, 3} \tag{2.1}$$

[2].

2.3*Means Square Error* (*MSE*)

Nilai *MSE* penduga pada simulasi data adalah jumlah selisih kuadrat dari penduga parameter pada data yang terdapat outlier dengan penduga parameter pada data yang tanpa *outlier*, dibagi dengan banyaknya perulangan. Hal ini dinotasikan sebagai berikut:

$$MSE(\hat{\beta}_k) = \frac{1}{m}\sum_{s=1}^{m}\left(\hat{\beta}_k^{(s)} - \hat{\beta}_k^{(0)}\right)^2 \quad , \quad \text{k=1, 2, 3} \tag{2.2}$$

[2].

2.4 Metode *Minimum Covariance Determinant* (*MCD*)

Misalkan $\mathbf{X} = \{ x_1, \ldots, x_n \}$ merupakan suatu himpunan sampel dari n pengamatan dalam $\mathbf{R}^k$ dengan $h \approx n/2$, maka akan ditentukan subhimpunan **J*** berukuran h sedemikian sehingga:

$$J^* = \min_{J\subset\{x_1,\ldots x_n\}\&|J|=h} \det \hat{S}_J \ , \tag{2.3}$$

dimana $\hat{S}_J$ adalah matriks kovarians berdasarkan pada x_i dengan $i \in J$. Penduga *MCD* diberikan oleh :

$$\hat{\mu} = \bar{x}_{J^*} = \frac{1}{h}\sum_{i\in J^*} x_i \ , \tag{2.4}$$

$$\hat{S} = \hat{S}_{J^*} = \frac{1}{h}\sum_{i\in J^*} (x_i - \hat{\mu})'(x_i - \hat{\mu}) \ , \tag{2.5}$$

[3].

2.5 Metode *Minimum Volume Ellipsoid* (*MVE*)

Minimum Volume Ellipsoid merupakan salah satu metode *robust* yang dapat digunakan untuk mendeteksi terdapatnya *outlier*. Pendeteksian *outlier* merupakan langkah penting dalam analisis data, karena akan sangat berpengaruh terhadap pendugaan. Terdapatnya satu *outlier* saja pada data dapat mengaburkan efek nyata atau menyatakan tidak ada efek dalam pengambilan kesimpulan.

Untuk mengatasi masalah *outlier* ini Rousseeuw memperkenalkan metode *robust* yang resisten terhadap adanya *outlier*, yaitu Metode *Minimum Volume Ellipsoid* (*MVE*) [4].

2.6 Regresi Komponen Utama

Tahap pertama pada prosedur regresi komponen utama yaitu menghitung komponen utama yang merupakan kombinasi linear dari beberapa peubah *X*, dan tahap kedua adalah peubah tak-bebas diregresikan pada komponen utama dalam sebuah model regresi linear. Bentuk persamaan regresi dalam bentuk peubah asli *X* dapat ditulis sebagai:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{2.6}$$

Peubah baru (***K***) sebagai komponen utama adalah hasil transformasi dari peubah asal (***X***) yang modelnya dalam bentuk matriks adalah **K** = ***A X***, dan komponen ke-*j* ditulis:

$$K_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p, \text{ atau}$$

$$K_j = \underline{a_j}' \underline{x}, \tag{2.7}$$

dengan vektor pembobot a_j' diperoleh dengan memaksimumkan keragaman komponen utama ke-*j*, yaitu:

$$\boldsymbol{S_y^2 = \underline{a_j}' S \underline{a_j}}, \tag{2.8}$$

dengan kendala $a_j' a_j = 1$ serta $a_h' a_j = 0$, untuk $h \neq j$. Vektor pembobot $\underline{a_j}'$ diperoleh dari matriks peragam Σ yang diduga dengan matriks ***S***, yaitu:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})', \tag{2.9}$$

Misalkan diberikan notasi K_1, K_2 …, K_m sebagai banyaknya komponen utama dan *Y* sebagai peubah tak-bebas, maka model regresi komponen utama dapat ditulis sebagai:

$$Y = w_0 + w_1 K_1 + w_2 K_2 + \ldots + w_m \mathrm{K_w} + \varepsilon, \tag{2.10}$$

2.7 Metode *Least Median Square* (*LMS*)

Metode *Least Median Square* (*LMS*) merupakan salah satu jenis regresi *robust*. Algortma ini meminimumkan median dari kuadrat residu untuk mendapatkan koefisien regresi β, yaitu:

$$\hat{\beta} = \min_{\beta} median\left(e_i^2\right) = \min_{\beta} median\left(y_i - \hat{y}_i\right)^2, i = 1,2,3\ldots,n \tag{2.11}$$

2.8 Metode *Least Trimmed Square* (*LTS*)

Metode *Least Trimmed Squares* merupakan salah satu metode penaksiran parameter model regresi yang *Robust* terhadap kehadiran nilai *outlier* dengan memangkas data *outlier* terlebih dahulu sebelum diproses dalam penaksiran parameter. Kemudian hasil pemangkasan digunakan untuk mendapatkan parameter dengan meminimalisasi jumlah kuadrat residunya.

$$\hat{\beta} = \min_{\beta}\left(\sum_{i=1}^{h} e_i^{\,2}\right) = \min_{\beta}\left(\sum_{i=1}^{h}(y_i - \hat{y}_i)^{\,2}\right), \quad \frac{(3n + p + 1)}{4} \leq h \leq n \tag{2.12}$$

3. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini bersifat studi literatur, yaitu mengumpulkan bahan atau materi yang berkaitan dengan topik penelitian dan melakukan simulasi data dengan menjalankan program makro simulasinya menggunakan perangkat lunak SAS (*Statistikal Analysis System).*

4. Hasil dan Pembahasan

4.1 Simulasi Data

Pada peneliian ini dilakukan simulasi data dengan banyaknya sampel secara keseluruhan sebanyak 3000 sampel, yaitu kombinasi 3 ukuran sampel (n = 20, 100 dan 200) dengan 5 jenis persentase *outlier* (5%, 10%, 15%, 20%, dan 25%), serta melakukan perulangan sebanyak 200 kali.

4.2 Perbandingan Nilai Bias dan *MSE*

Dari hasil perhitungan yang dilakukan dengan menggunakan program statistik *SAS*, maka didapatkan nilai Bias dan nilai *MSE* keempat metode yang diproses dengan fungsi *CALL* yang disajikan pada Tabel.1 dan Tabel 2, sebagai berikut:

Tabel 1. Nilai Perbandingan Bias pada setiap persentase *outlier* (p) dan ukuran sampel (n)

N	% *outlier*	BIAS			
		MCD-LMS	***MCD-LTS***	***MVE-LMS***	***MVE-LTS***
20	**5**	0.4062211	0.2872522	0.8952579	0.8512144
	10	0.0882421	0.2420568	0.5868438	0.1215413
	15	0.5459741	0.573448	0.7736965	0.1706505
	20	0.3423222	0.1545703	0.452071	0.9651375
	25	0.3842255	1.0148992	0.6225008	0.7024695
100	**5**	0.3755421	0.1419311	0.847831	0.2916521
	10	0.2332863	0.1409138	0.2538935	0.1655292
	15	0.1307667	0.171274	0.1099079	0.7196663
	20	0.0856148	0.2012745	0.3751612	0.101108
	25	0.2457223	0.357398	0.2008044	0.3307221
200	**5**	0.1449731	0.0850971	0.1678687	0.0928668
	10	0.1543644	0.2845202	0.5323483	0.424492
	15	0.1771534	0.3005856	0.4976848	0.1020183
	20	0.1512293	0.4069505	0.2096056	0.1980781
	25	0.360963	0.4332643	0.155888	0.1372264

Berdasarkan tabel hasil perhitungan untuk ukuran data 20 dan persentase *outlier 5 %*, nilai bias *MCD-LMS* 0.4062211, nilai bias *MCD-LTS* 0.2872522, *MVE-LMS* 0.8952579 dan *MVE-LTS* 0.8512144. Untuk persentase *outlier* 10% nilai bias *MCD-LMS* 0.0882421, *MCD-LTS* 0.2420568, *MVE-LMS* 0.5868438, *MVE-LTS* 0.1215413. Untuk persentase *outlier* 15% nilai bias *MCD-LMS* 0.5459741, *MCD-LTS* 0.573448, *MVE-LMS* 0.7736965, *MVE-LTS* 0.1706505. Untuk persentase *outlier* 20% nilai bias MCD-LMS 0.3423222, *MCD-LTS*

0.1545703, *MVE-LMS* 0.452071, *MVE-LTS* 0.9651375. Untuk persentase *outlier* 25% nilai bias *MCD-LMS* 0.3842255, *MCD-LTS* 0.0148992, *MVE-LMS* 0.6225008, *MVE-LTS* 0.7024695.

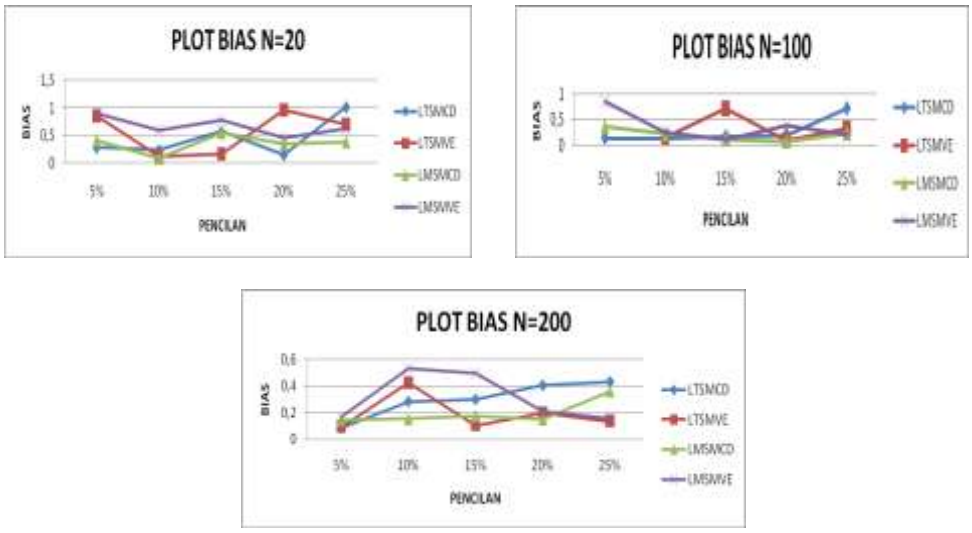
Tabel 2. Nilai Perbandingan *MSE* pada setiap persentase *outlier* (p) dan ukuran sampel (n)

N	% *outlier*	*MSE*			
		MCD-LMS	***MCD-LTS***	***MVE-LMS***	***MVE-LTS***
20	**5**	0.1871799	0.2872522	1.0283926	0.8205123
	10	0.0122856	0.2420568	0.3947104	0.0227195
	15	0.3814278	0.3383103	0.8793922	0.0482091
	20	0.1540165	0.041714	0.2813423	1.0584961
	25	0.2070771	1.477918	0.4207034	0.5171149
100	**5**	0.1774176	0.0278497	0.9875951	0.1164428
	10	0.1003554	0.0275954	0.0913589	0.0611763
	15	0.0209813	0.0369205	0.0122451	0.6378067
	20	0.0075152	0.063531	0.1408097	0.0147666
	25	0.0900214	0.1357883	0.0800948	0.1098245
200	**5**	0.0370406	0.0088453	0.0369071	0.0166624
	10	0.0278838	0.1115916	0.3843337	0.2172198
	15	0.0332465	0.1156163	0.2961968	0.022313
	20	0.0291037	0.1888398	0.0613696	0.0733689
	25	0.1658019	0.2817163	0.0404332	0.0204416

Berdasarkan tabel hasil perhitungan untuk ukuran data 20 dan persentase *outlier 5 %*, nilai *MSE MCD-LMS* 0.1871799, nilai *MSE MCD-LTS* 0.2872522, *MVE-LMS* 1.0283926 dan *MVE-LTS* 0.8205123. Untuk persentase *outlier* 10% nilai *MSE MCD-LMS* 0.0122856, *MCD-LTS* 0.2420568, *MVE-LMS* 0.3947104, *MVE-LTS* 0.0227195. Untuk persentase *outlier* 15% nilai *MSE MCD-LMS* 0.3814278, *MCD-LTS* 0.3383103, *MVE-LMS* 0.8793922, *MVE-LTS* 0.0482091. Untuk persentase *outlier* 20% nilai *MSE MCD-LMS* 0.1540165, *MCD-LTS* 0.41714, MVE-LMS 0.2813423, *MVE-LTS* 1.0584961. Untuk persentase *outlier* 25% nilai *MSE MCD-LMS* 0.2070771, *MCD-LTS* 1.477918, *MVE-LMS* 0.4207034, *MVE-LTS* 0.5171149.

4.3 Plot Bias dan *MSE*

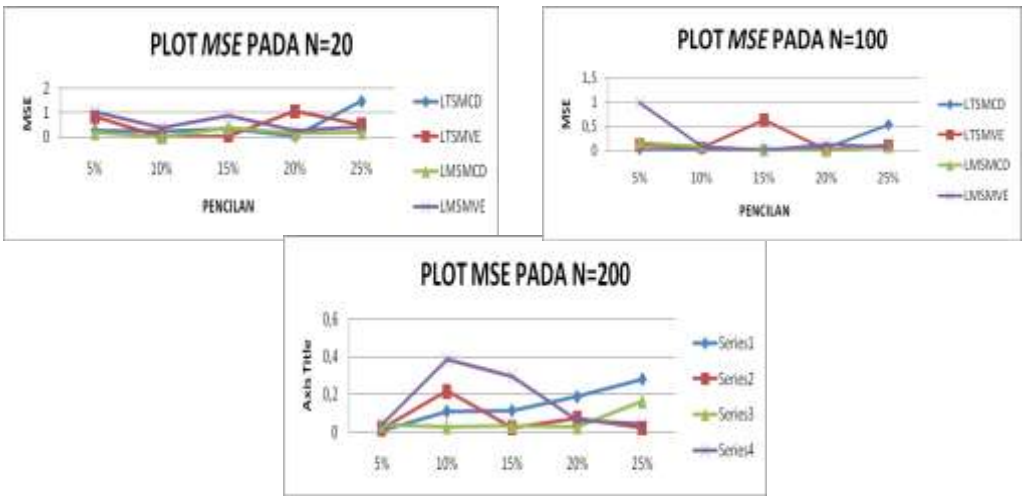
Nilai Bias dugaan koefisien regresi metode *MCD-LMS, MCD-LTS, MVE-LMS, dan MVE-LTS* yang diperoleh dengan menggunakan program *SAS* disajikan dalam bentuk grafik dengan ukuran data 20, 100, 200 dan dengan persentase *outlier* (5%, 10%, 15%, 20%, 25%), sebagai berikut:



Gambar 1. Plot Bias pada n = 20, n = 100, dan n = 200

Dari gambar 1 terlihat bahwa metode *MCD-LMS* dengan garis berwarna hijau mempunyai nilai bias relatif lebih kecil dibandingkan dengan metode *MCD-LTS, MVE-LMS dan MVE-LTS.*

Nilai *MSE* dugaan koefisien regresi metode *MCD-LMS, MCD-LTS, MVE-LMS,* dan *MVE-LTS* yang diperoleh dengan menggunakan program *SAS* yang disajikan dalam bentuk grafik, sebagai berikut:



Gambar 2. Plot MSE pada n = 20, n = 100, dan n = 200

Dari gambar 2 menunjukkan bahwa metode *MCD-LMS* yang digambarkan oleh garis berwarna hijau menunjukkan Nilai *MSE* yang relatif paling kecil dibandingkan dengan metode lainnya.

4.4 Perbandingan Tingkat Resistensi Metode *Robust*

Berdasarkan nilai dugaan koefisien regresi dari 3000 sampel, nilai bias dan *MSE* yang diperoleh dari metode *MCD-LMS, MCD-LTS, MVE-LMS, MVE-LTS* terlihat bahwa metode *MCD-LMS* menghasilkan nilai Bias dan *MSE* yang relatif lebih kecil. Semakin besar ukuran sampel semakin kecil juga nilai Bias dan *MSE* yang diperoleh oleh semua metode.

Nilai Bias dan *MSE* yang disajikan dalam bentuk tabel juga terlihat bahwa metode *MCD-LMS* yang ditunjukkan oleh garis berwarna hijau relatif lebih kecil dibandingkan dengan metode yang lainnya.

5. Kesimpulan

Berdasarkan hasil dan pembahasan dapat disimpulkan bahwa metode *MCD-LMS* memberikan hasil yang lebih baik dibandingkan dengan metode *MCD-LTS, MVE-LMS dan MVE-LTS* dikarenakan perbandingan nilai Bias dan *MSE* pada metode *MCD-LMS* lebih kecil dibandingkan dengan metode *MCD-LTS, MVE-LMS* dan *MVE-LTS.*

DAFTAR PUSTAKA

[1]. Oja, H. 2002. *Robust And Nonparametric Multivariate Methods.* Department of Mathematic and Statistics University of Jyväskylä. Finland.

[2]. Pison, G., Rousseeuw, P. J., P. Filzmoser, & C. Croux. 2001. *Robust Factor Analysis.* Academic Press.
http://www.elsevier.com/locate/jmva.pdf.
Diakses tanggal 15 Februari 2010.

[3]. Critchley F., Schyns ,M., & Haesbroeck, G. 2003. "Smooth Optimization for the MCD estimator". *International Conference On Robust Statistics*, Wed 16 July, Belgium.

[4]. Rousseeuw, R. J., & Leroy, A.M. 1987. Robust Regression and Outlier Detection. New York:Wiley.