PENERAPAN *BOOTSTRAP* DALAM METODE *MINIMUM COVARIANCE DETERMINANT* (MCD) DAN *LEAST MEDIAN OF SQUARES* (LMS) PADA ANALISIS REGRESI LINIER BERGANDA

Ni Putu Iin Vinny Dayanti[§1], Ni Luh Putu Suciptawati[2], Made Susilawati[3]

[1]Jurusan Matematika Fakultas MIPA – Universitas Udayana [Email: vinnyiindayanti@gmail.com]
[2]Jurusan Matematika Fakultas MIPA – Universitas Udayana [Email: putusuciptawati@yahoo.co.id]
[3]Jurusan Matematika Fakultas MIPA – Universitas Udayana [Email: mdsusilawati@unud.ac.id]
[§]*Corresponding Author*

ABSTRACT

Ordinary Least Squares (OLS) Method is a good method to estimate regression parameters when there is no violation in classical assumptions, such as the existence of outlier. Outliers can lead to biased parameters estimator, therefore we need a method that can may not affected by the existence of outlier such as Minimum Covariance Determinant (MCD) and Least Median of Squares (LMS). However, the application of this method is less accurate when it is used for small data. To overcome this problem, it was aplicated bootstrap method in MCD and LMS to determine the comparison of bias in parameters which were produced by both methods in dealing outlier in small data. The used bootstrap method in this study was the residual bootstrap that works by resampling the residuals. By using 95% and 99% confidence level and 5%, 10% and 15% outlier percentage, MCD-bootstrap and LMS-bootstrap give value of parameter estimators which were unbias for all percentage of outlier. We also found that the widht of range which produced by MCD-bootstrap method was shorter than LMS-bootstrap method produced. This indicates that MCD-bootstrap method was a better method than LMS-bootstrap method.

Keywords: outliers, bias, robust, Minimum Covariance Determinant, Least Median of Squares, bootstrap residual

1. PENDAHULUAN

Analisis regresi linier berganda merupakan analisis yang digunakan untuk menyelidiki hubungan linier antara dua atau lebih peubah prediktor terhadap peubah respon yang berskala minimal interval (Neter, et al [1]).

Metode kuadrat terkecil (MKT) merupakan metode penduga parameter regresi yang baik bila tidak terjadi pelanggaran asumsi klasik, seperti adanya pencilan. Pencilan merupakan data yang pengamatannya berada jauh dari sekelompok data amatan lainnya yang menyebabkan penduga parameter bersifat bias (Neter, et al [1]). Metode yang bisa mengatasi pencilan yaitu *Minimum Covariance Determinant* (MCD) dan *Least Median of Squares* (LMS). Namun penggunaan metode MCD dan LMS kurang tepat apabila berhadapan dengan data berukuran kecil. Penelitian ini dilakukan dengan menerapkan *bootstrap* pada kedua metode (MCD-*bootstrap*) dan (LMS-*bootstrap*) untuk mengetahui perbandingan bias pada parameter yang dihasilkan dalam mengatasi pencilan pada data berukuran kecil.

Metode *bootstrap* yang digunakan adalah *bootstrap residual* yang bekerja dengan me*resampling* sisaannya (*residual*) (Efron & Tibshirani [2]).

Metode *Minimum Covariance Determinant* (MCD) memiliki prinsip kerja menggunakan vektor rataan dan matriks kovarians dengan membentuk subsampel H yang berukuran h dari sampel berukuran n amatan yang matriks kovariansnya memiliki determinan terkecil (Hubert & Debruyne [3]). Nilai h diperoleh dari:

$$h = \left\lfloor \frac{n+p+1}{2} \right\rfloor, h \leq n \qquad (1)$$

Selanjutnya dicari vektor rataan $\boldsymbol{V}_{MCD}$ dan matriks kovarians S_{MCD} serta jarak mahalanobis kekar RD dengan menggunakan rumus (Hubert & Debruyne [3]):

$$\boldsymbol{V}_{MCD} = \frac{1}{h}\sum_{i\in H} x_i \quad (2)$$

$$S_{MCD} = \frac{1}{h}\sum_{i\in H}[x_i - \boldsymbol{V}_{MCD}][x_i - \boldsymbol{V}_{MCD}]^T \quad (3)$$

$$RD = \sqrt{(x_i - \boldsymbol{V}_{MCD})^T S_{MCD}^{-1}(x_i - \boldsymbol{V}_{MCD})} \quad (4)$$

Selanjutnya ditentukan *Fast* MCD (Rousseeuw [4]) yaitu terlebih dahulu dengan menentukan subsampel H_1 yang berukuran h kemudian dapat dihitung nilai $\boldsymbol{V}_{MCD}$ dan S_{MCD} dengan misalkan sebagai $\boldsymbol{V_1}$ dan S_1 serta menghitung determinan dari S_1 atau $det\,(S_1)$. Jika $det\,(S_1) \neq 0$ maka dilanjutkan dengan menghitung nilai RD yang diurutkan dari terkecil hingga terbesar. Pada iterasi berikutnya yaitu H_2 akan diambil sebanyak h pengamatan dengan jarak RD terkecil. Demikian seterusnya hingga mencapai konvergen $det\,(S_{i+1}) = det\,(S_1)$. Kemudian pilih himpunan H yang memiliki determinan S_{MCD} terkecil serta menghitung nilai $\boldsymbol{V}_{MCD}$ dan S_{MCD}. Maka selanjutnya data dapat diboboti dengan

$$W_i = \begin{cases} 1, & jika\ (x_i - \boldsymbol{V}_{MCD})^T S_{MCD}^{-1}(x_i - \boldsymbol{V}_{MCD}) \leq \chi^2_{p;1-\alpha} \\ 0, & lainnya \end{cases}$$

Sehingga dapat dibentuk matriks

$$W_{MCD} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

Dan diperoleh penduga MCD

$$\hat{\beta}_{MCD} = (X^T W_{MCD} X)^{-1}(X^T W_{MCD} Y) \quad (5)$$

Least Median of Squares (LMS) merupakan metode yang bekerja dengan meminimalkan median (nilai tengah) dari kuadrat residual (e_i^2) (Rousseeuw [5] yaitu:

$$M_J = \min\{median\ e_i^2\} \quad (6)$$

dilakukan pada urutan nilai residual kuadrat.

Langkah awal metode LMS adalah menentukan kuadrat nilai *error* dari MKT sehingga diperoleh nilai M_1. Selanjutnya dihitung nilai h_1 dengan rumus:

$$h_i = \left\lceil \frac{n+p+1}{2} \right\rceil \quad (7)$$

Kemudian pada iterasi ke-2 (M_2) diambil pengamatan sejumlah h_1 dari M_1 dengan jarak nilai (e_i^2) yang minimum. Demikian seterusnya sampai iterasi berakhir pada iterasi ke-i yaitu saat $h_i = h_{i+1}$ Selanjutnya dapat dihitung bobot w_{ii} dengan rumus:

$$w_{ii} = \begin{cases} 1, & jika\ |e_i/\hat{\sigma}| \leq \alpha \\ 0, & lainnya \end{cases} \quad (8)$$

dengan

$$\hat{\sigma} = 1{,}4826\left[1 + \frac{5}{n-p}\right]\sqrt{M_J} \quad (9)$$

maka dapat dibentuk matriks W:

$$W = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \quad (10)$$

dengan entri matriks $w_{ij} = 0$, dengan $i \neq j$. Penduga parameter regresi LMS dapat dihitung dengan menggunakan rumus:

$$\hat{\beta}_{LMS} = (X^T W X)^{-1}(X^T W Y) \quad (11)$$

Langkah-langkah *bootstrap residual* (Efron &Tibshirani [2]) adalah menentukan nilai $\hat{Y}$ yang dihasilkan oleh model analisis regresi, selanjutnya dapat diperoleh nilai *residual* yaitu, $e = Y - \hat{Y}$. Selanjutnya mengambil sampel *bootstrap* berukuran n dari $e_1, e_2, e_3, \dots e_n$ secara acak dengan pengembalian, sehingga diperoleh sampel bootstrap pertama $e^* = (e_1^*, e_2^*, \dots e_n^*)$. Kemudian hitung nilai *bootstrap* untuk Y^* dengan cara:

$$X\hat{\beta} + e^* = Y^* \quad (12)$$

Lebih lanjut lagi dihitung koefisien regresi untuk sampel *bootstrap* Y^* sehingga diperoleh $\hat{\beta}^*$. Iterasi terus dilakukan sampai pada batas *replikasi* yang diinginkan.

2. METODE PENELITIAN

Penelitian ini menggunakan data simulasi melalui pembangkitan data berdistribusi normal dengan bantuan *software* R i386 3.1.3. Data ini terdiri dari sisaan dan dua peubah prediktor yang akan digunakan untuk menentukan peubah responnya. Persentase pencilan yang diberikan

sebesar 5%, 10% dan 15%. Serta dengan menggunakan alpha (α) sebesar 0,05.

Langkah pembangkitkan data yaitu dengan membangkitkan nilai sisaan (e) berdistribusi $N(0,1)$. Kemudian membangkitkan peubah $X_1 \sim N(50,3)$ dan $X_2 \sim N(80,5)$ sebanyak 40 amatan, dengan memisalkan $\beta_0 = 2$, dan $\beta_1 = \beta_2 = 1$, akan diperoleh nilai Y dengan membentuk persamaan

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Pencilan yang dibangkitkan pada data sisaan dengan $\mu = 5$ dan $\sigma = 0{,}1$ pada tiap persentase pencilan. Selanjutnya menghitung nilai Y yang sudah terkontaminasi pencilan. Kemudian dilakukan uji kenormalan, pendeteksian multikolinearitas, pemeriksaan pencilan dan dilanjutkan menganalisis dengan MKT.

Langkah berikutnya menganalisis dengan metode MCD-*Bootstrap* yaitu menduga nilai β_0, β_1 dan β_2 dari matriks kovarian *robust* yang telah diperoleh dari penduga MCD. *Resampling* sisaan dengan *bootstrap residual* sebanyak 500 dan 1.000 kali dilakukan dengan menggunakan selang kepercayaan 95% dan 99%. Selanjutnya menganalisis dengan metode LMS-*Bootstrap. Resampling* sisaan yang diperoleh dari metode LMS dengan *bootstrap residual* sebanyak 500 dan 1.000 kali dan dilakukan dengan menggunakan selang kepercayaan 95% dan 99%. Kemudian membandingkan hasil yang diperoleh dengan MCD-*bootstrap* dan LMS-*bootstrap*.

3. HASIL DAN PEMBAHASAN

A. Hasil Pengujian Asumsi Kenormalan Data Dengan Uji Anderson-Darling

Berdasarkan hasil pengujian asumsi kenormalan dapat dilihat pada tabel 1 berikut:

Tabel 1. Uji Kenormalan Data

Persentase pencilan	*p-value*	Keterangan
Data awal (tanpa pencilan)	0,780	Normal
5%	0,03635	Tidak normal
10%	<0,005	Tidak normal
15%	<0,005	Tidak normal

Hasil uji kenormalan pada Tabel 1, data dengan pencilan sebesar 5%, 10% serta 15% memiliki nilai *p-value* < α, hal ini menunjukkan data dengan pencilan memiliki sebaran data yang tidak normal.

B. Pendeteksian Multikolinearitas

Untuk melihat masalah multikolinearitas maka dilakukan dengan melihat nilai korelasi yang dihasilkan antara peubah prediktor.

Tabel 2. Korelasi Antarvariabel

Variabel	Y	X_1
X_1	0,309	
	0,052	
X_2	0,873	−0,161
	0,000	0.321

Dari Tabel 2 dapat dilihat bahwa nilai korelasi yang dihasilkan pada X_1 dan X_2 sebesar -0,161 yang menunjukkan peubah X_1 dan X_2 memiliki hubungan yang berlawanan arah namun tidak terjadi masalah multikolinearitas.

C. Pemeriksaan Pencilan atau *Outlier*

Pemeriksaan pencilan dilakukan dengan menggunakan *Robust Distance* (RD) lalu membandingkannya dengan nilai *chi-square*. Dalam pemeriksaan menggunakan RD diperoleh hasil seperti pada Tabel 3:

Tabel 3. Pemeriksaan Pencilan dengan *Robust Distance* (RD)

Data	Persentase pencilan	Data pengamatan ke-		Banyak pencilan
		outlier orthogonal	*bad leverage*	
40	5%	1, 2, 3, 4, 5, 6	31	7
	10%	1, 2, 4, 7, 18, 23, 25	3, 31	9
	15%	3, 7, 18, 23	1, 2, 31	7

Tabel 3 menunjukkan hasil pemeriksaan pencilan yaitu dengan persentase pencilan 5% terdeteksi 7 pengamatan sebagai pencilan dan 9 pengamatan yang merupakan pencilan pada persentase 10% dan pada peresentase 15% terdeteksi 7 pengamatan sebagai pencilan. Pencilan yang terdeteksi merupakan jenis *outlier orthogonal* maupun *bad leverage*.

D. Analisis Data dengan Metode Kuadrat Terkecil (MKT)

Analisis data dengan MKT akan menggunakan selang kepercayaan 95% dan 99%.

Tabel 4. Penduga Parameter dengan MKT

Jumlah Pencilan	Parameter	Estimasi	Selang Kepercayaan 95%		Selang Kepercayaan 99%	
			Selang Kepercayaan	Ket	Selang Kepercayaan	Ket
Data tanpa pencilan	β_1	0.9752	0.8514-1.0991	Tidak bias	0.8092-1.1412	Tidak bias
	β_2	1.0608	0.9952-1.1265	Tidak bias	0.9729-1.1488	Tidak bias
5%	β_1	1.3865	0.9669-1.1462	Bias	0.9059-1.3255	Bias
	β_2	1.0591	0.9641-1.1541	Tidak bias	0.9317-1.1864	Tidak bias
10%	β_1	1.4079	0.9021-1.1182	Bias	0.8286-1.3344	Bias
	β_2	1.1412	0.8732-0.9877	Bias	0.8343-1.0229	Bias
15%	β_1	1.4283	0.8549-1.0999	Bias	0.7715-1.3449	Bias
	β_2	1.1854	0.8816-1.0114	Bias	0.8375-1.1413	Bias

Karena nilai penduga penduga parameter β_1 dan β_2 yang dihasilkan oleh MKT bersifat tidak bias hanya saat pencilan 5% untuk β_2, hal ini berarti MKT mengalami bias saat adanya pencilan. Maka akan dilanjutkan dengan menganalisis dengan metode *Minimum Covariance Determinant* (MCD)-*Bootstrap* dan *Least Median of Squares* (LMS)-*Bootstrap.*

E. Analisis Data dengan *Metode Minimum Covariance Determinant* (MCD)-*Bootstrap*

Berdasarkan hasil analisis dengan metode MCD-*bootstrap* dengan *resampling* 500 dan 1000 kali dapat dilihat pada Tabel 5 dan 6 adalah berikut:

Tabel 5. Pendugaan parameter dengan metode MCD-*bootstrap* dengan B=500 kali *resampling*

Jumlah Pencilan	Parameter	Estimasi	Selang Kepercayaan 95%		Estimasi	Selang Kepercayaan 99%	
			Selang Kepercayaan	Ket		Selang Kepercayaan	Ket
5%	β_1	1.0929	0.9871-1.1938	Tidak bias	1.0908	0.9592-1.2217	Tidak bias
	β_2	0.9676	0.9031-1.0368	Tidak bias	0.9693	0.8841-1.0558	Tidak bias
10%	β_1	1.1929	1.0706-1.3156	Tidak bias	1.1958	1.0243-1.3620	Tidak bias
	β_2	0.9065	0.8275-0.9874	Tidak bias	0.905	0.7970-1.0179	Tidak bias
15%	β_1	1.1406	1.0014-1.2722	Tidak bias	1.1366	0.9587-1.3149	Tidak bias
	β_2	0.9436	0.8593-1.0353	Tidak bias	0.9466	0.8325-1.0623	Tidak bias

Tabel 6. Pendugaan parameter dengan metode MCD-*bootstrap* dengan B=1000 kali *resampling*

Jumlah Pencilan	Parameter	Estimasi	Selang Kepercayaan 95%		Estimasi	Selang Kepercayaan 99%	
			Selang Kepercayaan	Ket		Selang Kepercayaan	Ket
5%	β_1	1.0897	0.9879-1.1930	Tidak bias	1.091	0.9582-1.2227	Tidak bias
	β_2	0.9698	0.9031-1.0369	Tidak bias	0.9689	0.8840-1.0559	Tidak bias
10%	β_1	1.1919	1.0723-1.3139	Tidak bias	1.1937	1.0392-1.3471	Tidak bias
	β_2	0.9074	0.8287-0.9862	Tidak bias	0.9063	0.8070-1.0079	Tidak bias
15%	β_1	1.1354	1.0050-1.2686	Tidak bias	1.1396	0.9574-1.3162	Tidak bias
	β_2	0.9471	0.8618-1.0330	Tidak bias	0.9443	0.8307-1.0541	Tidak bias

Dari Tabel 5 dan 6 diperoleh bahwa penduga parameter yang dihasilkan oleh metode MCD-*bootstrap* bersifat tidak bias dengan *resampling* 500 maupun 1000 kali. Hal ini berarti bahwa penduga parameter β_1 dan β_2 yang dihasilkan oleh metode *bootstrap residual* berada di dalam selang kepercayaan 95% dan 99%.

F. Analisis Data dengan Metode *Least Median of Squares* (LMS)-*Bootstrap*

Berdasarkan hasil analisis dengan metode LMS-*bootstrap* dengan *resampling* 500 dan 1000 kali dapat dilihat pada Tabel 7 dan 8 adalah berikut:

Tabel 7. Pendugaan parameter dengan metode *Least Median of Squares* (LMS)-*Bootstrap* dengan 500 kali *resampling*

Jumlah Pencilan	Parameter	Estimasi	Selang Kepercayaan 95%		Estimasi	Selang Kepercayaan 99%	
			Selang Kepercayaan	Ket		Selang Kepercayaan	Ket
5%	β_1	0.9122	0.8474-1.0577	Tidak bias	0.9079	0.8078-1.0973	Tidak bias
	β_2	1.0854	1.0397-1.1764	Tidak bias	1.088	1.0142-1.2019	Tidak bias
10%	β_1	0.908	0.8355-1.0868	Tidak bias	0.9086	0.7926-1.1297	Tidak bias
	β_2	1.0924	1.0350-1.1979	Tidak bias	1.0915	1.0072-1.2257	Tidak bias
15%	β_1	0.9264	0.6754-0.9689	Tidak bias	0.9334	0.6353-1.0090	Tidak bias
	β_2	1.0827	0.9379-1.1294	Tidak bias	1.0781	0.9127-1.1546	Tidak bias

Tabel 8. Pendugaan parameter dengan metode *Least Median of Squares* (LMS)-*Bootstrap* dengan 1000 kali *resampling*

Jumlah Pencilan	Parameter	Estimasi	Selang Kepercayaan 95%		Estimasi	Selang Kepercayaan 99%	
			Selang Kepercayaan	Ket		Selang Kepercayaan	Ket
5%	β_1	0.9102	0.8456-1.0595	Tidak bias	0.9062	0.8174-1.0877	Tidak bias
	β_2	1.0866	1.0386-1.1775	Tidak bias	1.0891	1.0201-1.1960	Tidak bias
10%	β_1	0.9073	0.8282-1.0941	Tidak bias	0.9132	0.7947-1.1276	Tidak bias
	β_2	1.0927	1.0302-1.2027	Tidak bias	1.0889	1.0086-1.2243	Tidak bias
15%	β_1	0.9314	0.6832-0.9611	Tidak bias	0.9316	0.6341-1.0102	Tidak bias
	β_2	1.0796	0.9436-1.1237	Tidak bias	1.0792	0.9122-1.1551	Tidak bias

Dari Tabel 7 dan 8 diperoleh bahwa dengan menganalisis menggunakan metode LMS-*bootstrap*, selang kepercayaan 95% dan 99% dapat mencakup nilai parameternya. Hal ini berarti hasil yang diperoleh dengan metode LMS-*bootstrap*, nilai penduga parameter β_1 dan β_2 bersifat tidak bias.

G. Perbandingan hasil MCD-*Bootstrap* dan LMS-*Bootstrap*

Perbandingan hasil analisis dengan metode MCD-*bootstrap* dan LMS-*bootstrap* dapat dilihat pada Tabel 9 dan 10 adalah berikut:

Tabel 9. Lebar selang pada selang kepercayaan 95% untuk β_1 dan β_2 pada metode MCD-*bootstrap* dan LMS-*bootstrap*

Parameter	Persentase Pencilan	Metode			
		MCD-*bootstrap*		LMS-*bootstrap*	
		B= 500	*B*= 1000	*B*= 500	*B*= 1000
β_1	5%	0.2067	0.205	0.2102	0.2138
	10%	0.2449	0.2415	0.2512	0.2658
	15%	0.2707	0.2635	0.2935	0.2778
β_2	5%	0.1337	0.1338	0.1367	0.1389
	10%	0.1598	0.1574	0.1629	0.1725
	15%	0.1762	0.1712	0.1914	0.18

Tabel 10. Lebar selang pada selang kepercayaan 99% untuk β_1 dan β_2 pada metode MCD-*bootstrap* dan LMS-*bootstrap*

Parameter	Persentase Pencilan	Metode			
		MCD-*bootstrap*		LMS-*bootstrap*	
		B= 500	*B*= 1000	*B*= 500	*B*= 1000
β_1	5%	0.2625	0.2644	0.2895	0.2703
	10%	0.3376	0.3078	0.3371	0.3329
	15%	0.3562	0.3588	0.3737	0.3761
β_2	5%	0.1716	0.1718	0.1877	0.1759
	10%	0.2208	0.2008	0.2185	0.2157
	15%	0.2297	0.2333	0.2419	0.2428

Dari Tabel 9 dan 10 menunjukkan bahwa dengan selang kepercayaan 95% dan 99%, metode MCD-*bootstrap* menghasilkan nilai lebar selang yang lebih kecil dibandingkan metode LMS-*bootstrap* untuk semua persentase pencilan pada β_1 dan β_2.

4. KESIMPULAN

Metode MCD-*bootstrap* maupun LMS-*bootstrap* merupakan metode yang baik dalam menduga nilai parameter saat data mengandung pencilan. Pada selang kepercayaan 95% dan 99%, metode MCD-*bootstrap* dan LMS-*bootstrap* menghasilkan nilai penduga parameter yang bersifat tidak bias untuk seluruh persentase pencilan. Karena lebar selang kepercayaan yang dihasilkan metode MCD-*bootstrap* lebih pendek dibanding metode LMS-*bootstrap,* maka dapat dikatakan metode MCD-*bootstrap* lebih akurat.

DAFTAR PUSTAKA

[1] Neter, J., Wasserman, W., & Kutner, M. 1997. *Model Linier Terapan Buku II: Analisis Regresi Linier Sederhana.* (Terjemahan Bambang Sumantri). Bandung: Jurusan FMIPA-IPB.

[2] Efron, B., & Tibshirani, R.J. 1993. *An Introduction to the Bootstrap.* New York London: Chapman & Hall.

[3] Hubert, M., & Debruyne, M. 2009. Minimum Covariance Determinant. *WIREs Computational Statistics 2010*, pp 36-43.

[4] Rousseeuw, P.J. 1999. Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, august 1999. Vol. 41, No. 3 American Statistical Association and the American Society for Quality, pp.212-223.

[5] _____,1984. Least Median of Squares Regression. *Journal of the American Statistical Association*, pp. 871-880.