Saintia Matematika
Vol. 1, No. 1 (2013), pp. 73–85.

PERBANDINGAN METODE *LEAST TRIMMED SQUARES* DAN PENAKSIR M DALAM MENGATASI PERMASALAHAN DATA PENCILAN

Sri Wulandari, Sutarman, Open Darnius

Abstrak. *Analisis regresi digunakan untuk mengetahui hubungan antar variabel. Salah satu metode penaksir parameter dalam model regresi ini ialah metode kuadrat terkecil. Di dalam penelitian ini digunakan simulasi terhadap empat kelompok data yang terdiri atas 20 observasi. Tulisan ini bertujuan untuk membanding dua metode regresi robust yakni Least Trimmed Squares (LTS) dan penaksir M. Dari hasil simulasi pada penelitian ini, LTS memberikan hasil perbandingan rata-rata kuadrat sisa yang lebih baik daripada penaksir M dan metode OLS. Sementara itu, penaksir M juga menghasilkan rata-rata kuadrat sisa yang lebih baik daripada metode OLS.*

1. PENDAHULUAN

Secara umum pencilan adalah data yang tidak mengikuti pola umum data [1]. Pencilan dapat menyebabkan munculnya nilai rata-rata dan simpangan baku yang tidak konsisten terhadap mayoritas data. Selain itu, estimasi koefisien garis regresi yang diperoleh tidak tepat, dan pada beberapa analisis inferensia dapat menyebabkan kesalahan dalam pengambilan keputusan dan kesimpulan. Pencilan dapat dideteksi menggunakan beberapa metode.

Received 25-01-2013, Accepted 21-02-2013.
2010 Mathematics Subject Classification: 93E10
Key words and Phrases: Pencilan, metode kuadrat terkecil, regresi *robust*, *least trimmed squares*, dan penaksir M.

Metode-metode tersebut diantaranya ialah metode grafik dan pendeteksian berdasarkan nilai *Leverage, DfFITS, Cook's Distance,* dan *DfBETA(s)* [2].

Jika terdapat pencilan, metode kuadrat terkecil tidak lagi efisien untuk mendapatkan penaksir parameter. Untuk mengatasi masalah ini, salah satu metode yang digunakan ialah metode regresi *robust*. Metode regresi ini dapat mengatasi pencilan dengan mencocokkan model regresi terhadap sebagian besar data. Selanjutnya, mengatasi titik-titik pencilan yang memiliki nilai sisaan sebagai solusi regresi *robust* [3].

Di dalam regresi *robust*, metode estimasi yang bisa digunakan ialah *Least Median Squares* (LMS), *Least Trimmed Squares* (LTS), penaksir M (M estimator), penaksir S, dan penaksir MM [3].

2. LANDASAN TEORI

2.1 Regresi Linier

Analisis regresi digunakan untuk mengetahui hubungan antara variabel terikat (Y) dengan satu atau lebih variabel bebas (X). Salah satu metode penaksir parameter dalam model regresi ini ialah metode kuadrat terkecil. Metode ini menentukan persamaan linier dengan cara meminimumkan jumlah kuadrat sisa. Model regresi untuk satu variabel bebas yaitu model regresi linier sederhana, dinyatakan dalam persamaan berikut [4]:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{1}$$

Model penaksir untuk persamaan (1) ialah:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i. \tag{2}$$

Untuk mendapatkan nilai penaksir β_0 dan β_1, digunakan prinsip metode kuadrat terkecil, yaitu meminimumkan jumlah kuadrat sisaan yang dinyatakan sebagai berikut:

$$Minimum \sum_{i=1}^{n} \varepsilon_i^2.$$

Berdasarkan metode kuadrat terkecil, nilai β_0 dan β_1 dapat ditaksir menggunakan rumus berikut:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{3}$$

dan

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n}}{[-\frac{1}{n}[\sum_{i=1}^{n} X_i]^2 + \sum_{i=1}^{n} X_i^2]}. \tag{4}$$

Kecocokan model dapat didasarkan pada nilai rata-rata kuadrat sisa. Jika nilai rata-rata kuadrat sisa yang dihasilkan semakin kecil maka model tersebut semakin baik. Nilai rata-rata kuadrat sisa dinyatakan dalam rumus berikut [1]:

$$S^2 = \frac{JKS}{n-p} = \frac{JKT - JKR}{n-p} \tag{5}$$

dengan

JKS = Jumlah kuadrat sisa
JKT = Jumlah kuadrat total
= $\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2$
JKR = Jumlah kuadrat regresi
= $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2$
n = Banyak sampel
p = Banyak parameter
Y_i = Data sebenarnya
$\hat{Y}_i$ = Data dugaan
$\bar{Y}_i$ = Rata-rata data sebenarnya.

2.2 Regresi *Robust*

Regresi *Robust* merupakan analisis data yang tidak peka terhadap kehadiran pencilan. Salah satu metode yang populer dalam regresi *robust* ialah *least trimmed squares*. Metode ini menggunakan konsep pengepasan metode kuadrat terkecil (ordinary least square) untuk meminimumkan jumlah kuadrat sisaan [5], dapat dinyatakan dalam rumus berikut:

$$\sum_{i=1}^{h} e_{(i)}^2 \tag{6}$$

dengan
$e^2_{(i)}$= Kuadrat residual (sisaan kuadrat) yang terurut dari terkecil hingga terbesar.
n = Jumlah pengamatan,
p = Jumlah parameter,
$h = \left[\frac{n}{2} + \frac{(p+1)}{2}\right] = \frac{[n+p+1]}{2}$.

Selain itu, penaksir M juga merupakan metode yang sangat populer. Metode ini menggunakan *weighted least square* (WLS) secara ite- rasi untuk meminimumkan $\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$.
Tahapan iterasi dalam penaksiran koefisien regresi ini ialah [6]:

1. Menghitung penaksir β, dinotasikan **b** menggunakan *least square*, sehingga didapatkan $\hat{y}_{i,0}$ dan $\varepsilon_{i,0} = y_i - \hat{y}_{i,0}$, ($i$ = 1,2,...,n) yang diperlakukan sebagai nilai awal (y_i adalah hasil eksperimen).

2. Dari nilai-nilai residual ini dihitung $\hat{\sigma}_0$, dan pembobot awal $w_{i,0} = \frac{\psi(\varepsilon^+_{i,0})}{(\varepsilon^+_{i,0})}$ Nilai $\psi(\varepsilon^*_i)$ dihitung sesuai fungsi Huber, dan $\varepsilon^*_{i,0} = \varepsilon_{i,0}/\hat{\sigma}_0$.

3. Menyusun matriks pembobot berupa matriks diagonal dengan elemen $w_{1,0},\ w_{2,0}, ..., w_{n,0}$ dinamai W_0.

4. Menghitung penaksir koefisien regresi:
$b_{Robust\ ke-1} = (X^T W_0 X)^{-1} X^T W_0 Y$.

5. Dengan menggunakan $b_{Robust\ ke-1}$ dihitung pula $\sum_{i=1}^{n} |y_i - \hat{y}_{i,1}|$ atau $\sum_{i=1}^{n} |\varepsilon_{i,1}|$.

6. Selanjutnya, langkah 2 sampai dengan 5 diulang sampai didapatkan nilai $\sum_{i=1}^{n} |\varepsilon_{i,m}|$ yang konvergen yakni jika selisih antara b_{m+1} dan b_m mendekati 0, dengan m jumlah iterasi.

3. METODE PENELITIAN

Langkah-langkah dalam penelitian ialah sebagai berikut:

a. Menggunakan data simulasi yang mengandung permasalahan pencilan.

b. Menguji data kemudian menggunakan dua metode regresi *robust* yakni *least trimmed squares* dan penaksir M untuk mengatasi pencilan.

c. Mengolah data menggunakan bantuan *software*.

d. Membandingkan hasil penyelesaian dan pengolahan data antara kedua metode.

e. Menyimpulkan hasil perbandingan.

4. HASIL DAN PEMBAHASAN

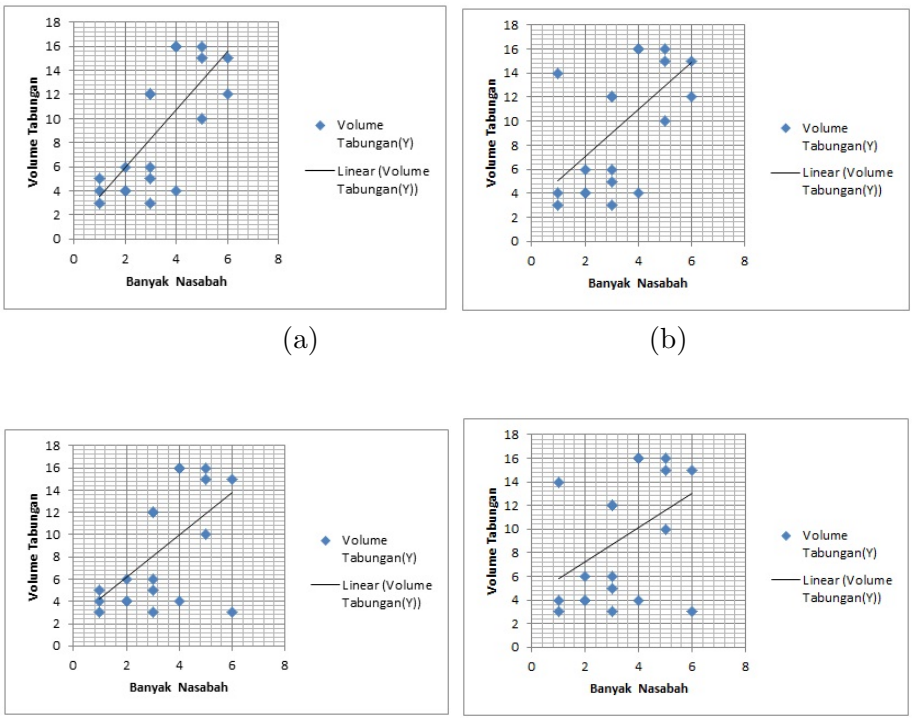
4.1 Pendeteksian Pencilan

Pada bagian ini akan diurai data dengan kehadiran pencilan. Sebagai simulasi dikemukakan empat kelompok data yang terdiri atas 20 observasi dengan satu variabel bebas. Keempat data disajikan pada Tabel 1 berikut:

Tabel 1: Data Simulasi

Hari ke-	DATA 1		DATA 2		DATA 3		DATA 4	
	Banyak Nasabah	Volume Tabungan	Banyak Nasabah	Volume Tabungan	Banyak Nasabah	Volume Tabungan	Banyak Nasabah	Volume Tabungan
1	3	3	3	3	3	3	3	3
2	1	5	1	14	1	5	1	14
3	2	6	2	6	2	6	2	6
4	4	4	4	4	4	4	4	4
5	3	6	3	6	3	6	3	6
6	5	10	5	10	5	10	5	10
7	2	4	2	4	2	4	2	4
8	1	3	1	3	1	3	1	3
9	3	5	3	5	3	5	3	5
10	6	15	6	15	6	15	6	15
11	4	16	4	16	4	16	4	16
12	3	12	3	12	3	12	3	12
13	5	16	5	16	5	16	5	16
14	6	12	6	12	6	3	6	3
15	4	16	4	16	4	16	4	16
16	1	4	1	4	1	4	1	4
17	5	15	5	15	5	15	5	15
18	4	16	4	16	4	16	4	16
19	2	4	2	4	2	4	2	4
20	3	12	3	12	3	12	3	12

Secara grafis, *scatter plot* untuk keempat data ialah sebagai berikut:



Gambar 1: (a) *Scatter plot* Data 1, (b) *Scatter plot* Data 2, (c) *Scatter plot* Data 3, (d) *Scatter plot* Data 4

Nilai-nilai *Leverage*, DfFITS, dan *Cook's Distance* untuk keempat kelompok data disaji pada Tabel 2 berikut:

Tabel 2: Nilai-nilai *Leverage*, DfFITS, $|DfFITS|$, dan *Cook's Distance*

Hari	Kelompok Data 1				Kelompok Data 2			
ke-	*Leverage*	DfFITS	$\|DfFITS\|$	*Cook's*	*Leverage*	DfFITS	$\|DfFITS\|$	*Cook's*
1	0,05263	−0,16852	0,16852	0,06178	0,05263	−0,34293	0,34293	0,05536
2	**0,16864**	0,19427	0,19427	0,01976	**0,16864**	1,17347	**1,17347**	0,52097
3	0,08915	0,00566	0,00566	0,00002	0,08915	−0,07338	0,07338	0,00284
4	0,05908	−0,51509	0,51509	0,11249	0,05908	−0,43481	0,43481	0,08503
5	0,05263	−0,15199	0,15199	0,01194	0,05263	−0,16292	0,16292	0,01367
6	0,10849	−0,31818	0,31818	0,05109	0,10849	−0,24264	0,24264	0,03031
7	0,08915	−0,16852	0,16852	0,01478	0,08915	−0,22338	0,22338	0,02565
8	**0,16864**	−0,06715	0,06715	0,00238	**0,16864**	−0,22662	0,22662	0,02679
9	0,05263	−0,21937	0,21937	0,02424	0,05263	−0,22037	0,22037	0,02445
10	**0,20086**	−0,09025	0,09025	0,00430	**0,20086**	0,01808	0,01808	0,00017
11	0,05908	0,37903	0,37903	0,06704	0,05908	0,30612	0,30612	0,04561
12	0,05263	0,23967	0,23967	0,02867	0,05263	0,16717	0,16717	0,01437
13	0,10849	0,27853	0,27853	0,03959	0,10849	0,26143	0,26143	0,03503
14	**0,20086**	−0,55221	0,55221	0,15068	**0,20086**	−0,36427	0,36427	0,06813
15	0,05908	0,37903	0,37903	0,06704	0,05908	0,30612	0,30612	0,04561
16	**0,16864**	0,06309	0,06309	0,00211	**0,16864**	−0,11453	0,11453	0,00692
17	0,10849	0,17743	0,17743	0,01642	0,10849	0,17577	0,17577	0,01612
18	0,05908	0,37903	0,37903	0,06704	0,05908	0,30612	0,30612	0,04561
19	0,08915	−0,16852	0,16852	0,01478	0,08915	−0,22338	0,22338	0,02565
20	0,05263	0,23967	0,23967	0,02867	0,05263	0,16717	0,16717	0,01437

Tabel 2: Nilai-nilai *Leverage*, DfFITS, $|DfFITS|$, dan *Cook's Distance* (*sambungan*)

Hari	Kelompok Data 3				Kelompok Data 4			
ke-	*Leverage*	DfFITS	$\|DfFITS\|$	*Cook's*	*Leverage*	DfFITS	$\|DfFITS\|$	*Cook's*
1	0,05263	−0,27232	0,27232	0,03640	0,05263	−0,27702	0,27702	0,03757
2	**0,16864**	0,07693	0,07693	0,00313	**0,16864**	0,86500	**0,86500**	0,32549
3	0,08915	−0,01230	0,01230	0,00008	0,08915	−0,07868	0,07868	0,00327
4	0,05908	−0,34778	0,34778	0,05752	0,05908	−0,32125	0,32125	0,04982
5	0,05263	−0,10811	0,10811	0,00611	0,05263	−0,12706	0,12706	0,00840
6	0,10849	−0,15003	0,15003	0,01179	0,10849	−0,11434	0,11434	0,00688
7	0,08915	−0,15304	0,15304	0,01223	0,08915	−0,20791	0,20791	0,02230
8	**0,16864**	−0,13395	0,13395	0,00945	**0,16864**	−0,26889	0,26889	0,03749
9	0,05263	−0,16121	0,16121	0,01339	0,05263	−0,17560	0,17560	0,01581
10	**0,20086**	0,14344	0,14344	0,01084	**0,20086**	0,21268	0,21268	0,02369
11	0,05908	0,34917	0,34917	0,05793	0,05908	0,30485	0,30485	0,04526
12	0,05263	0,20662	0,20662	0,02162	0,05263	0,15685	0,15685	0,01269
13	0,10849	0,33151	0,33151	0,05525	0,10849	0,32228	0,32228	0,05236
14	**0,20086**	−1,6523	**1,65235**	**0,88191**	**0,20086**	−1,28160	**1,28160**	0,62811
15	0,05908	0,34917	0,34917	0,05793	0,05908	0,30485	0,30485	0,04526
16	**0,16864**	−0,02837	0,02837	0,00043	**0,16864**	−0,17137	0,17137	0,01542
17	0,10849	0,24793	0,24793	0,0316	0,10849	0,24662	0,24662	0,03128
18	0,05908	0,34917	0,34917	0,05793	0,05908	0,30485	0,30485	0,04526
19	0,08915	−0,15304	0,15304	0,01223	0,08915	−0,20791	0,20791	0,02230
20	0,05263	0,20662	0,20662	0,02162	0,05263	0,15685	0,15685	0,01269

Dengan memperhatikan nilai-nilai *Leverage*, $|DfFITS|$ dan *Cook's Distance* pada Tabel 2 dan Tabel 3, data yang termasuk pencilan untuk keempat kelompok data ialah nilai yang lebih besar dari $Leverage = \frac{(2p-1)}{n} = \frac{2(2)-1}{20} = \frac{3}{20} = 0{,}15000$. Pada kelompok data 1, 2, 3, dan 4, data yang termasuk pencilan yaitu observasi di hari ke-2, ke-8, ke-10, ke-14, dan ke-16. Selanjutnya, berdasarkan nilai yang lebih besar dari $|DfFITS| = \left|2\sqrt{\frac{p}{n}}\right| = \left|2\sqrt{\frac{2}{20}}\right| = 0{,}63246$, yang termasuk pencilan untuk kelompok data 2 yaitu observasi di hari ke-2 sedangkan pada data 1 tidak terdapat pencilan. Sementara itu, yang termasuk pencilan untuk kelompok data 3 yaitu observasi di hari ke-14 dan pada kelompok data 4 yaitu observasi di hari ke-2 dan ke-14. Selain itu, pendeteksian berdasarkan nilai yang lebih besar dari *Cook's Distance* $= \mathrm{F}(0{,}5;p;n-p) = \mathrm{F}(0{,}5;2;18) = 0{,}72054$, hanya kelompok data 3 yang memiliki pencilan yaitu observasi di hari ke-14.

4.2 Penaksiran Parameter Berdasarkan Metode OLS, LTS, dan Penaksir M

Hasil penaksiran parameter berdasarkan metode kuadrat terkecil untuk keempat kelompok data dengan model penaksir $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ialah sebagai berikut:
Kelompok data 1: $\hat{Y} = 1{,}0966 + 2{,}4189X_i$
Kelompok data 2: $\hat{Y} = 3{,}0687 + 1{,}9646X_i$
Kelompok data 3: $\hat{Y} = 2{,}3631 + 1{,}9066X_i$
Kelompok data 4: $\hat{Y} = 4{,}3351 + 1{,}4522X_i$.

Selain itu, penaksiran parameter berdasarkan *least trimmed squares* ialah dengan mengurutkan nilai sisaan kuadrat $(e^2_{(i)})$ dari terkecil hingga terbesar menjadi sebanyak h. Hasil penaksiran untuk keempat kelompok data dengan model penaksir $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ialah sebagai berikut:
Kelompok data 1: $\hat{Y}_i = 0{,}3333 + 2{,}4722X_i$
Kelompok data 2: $\hat{Y}_i = 1{,}6575 + 2{,}1301X_i$
Kelompok data 3: $\hat{Y}_i = 0{,}9895 + 2{,}2684X_i$
Kelompok data 4: $\hat{Y}_i = 0{,}8214 + 2{,}3929X_i$.

Selanjutnya, penaksiran parameter berdasarkan penaksir M dapat diolah dengan bantuan *software* MINITAB 16 ataupun dengan mengikuti prosedur sebagai berikut:

1. Menghitung koefisien regresi menggunakan metode kuadrat terkecil, didapatkan nilai b dan $\varepsilon_{i,0}$.
2. Menghitung nilai $\hat{\sigma}_0 = 1,5$ (median $|\varepsilon_{i,0}|$) sehingga didapatkan nilai $\varepsilon^*_{i,0}$ dan $|\varepsilon^*_{i,0}|$.
3. Menentukan nilai $\psi(\varepsilon^*_i)$ dan pembobot $w_{i,0}$ sesuai dengan fungsi Huber.

Prosedur berikutnya yaitu:

4. Melakukan perhitungan $b_{Robust\ ke-1}$ sebagai penaksir weighted least square dengan pembobot $w_{i,0}$, diperolehlah koefisien $b_{Robust\ ke-1}$, $\varepsilon_{i,1}$, $\hat{\sigma}_1 = 1,5$ (median $|\varepsilon_{i,0}|$), $\varepsilon^*_{i,0}$, $\psi(\varepsilon^*_i)$ dan pembobot $w_{i,1}$, serta nilai $\sum_{i=1}^{n} |\varepsilon_{i,1}|$.

Berdasarkan *output* program MINITAB 16, diperoleh nilai koefisien regresi untuk keempat kelompok data yang disaji pada Tabel 3 berikut:

Tabel 3: Nilai Koefisien Regresi Penaksir M

Koefisien Regresi	Data 1	Data 2	Data 3	Data 4
	Iterasi ke-7	Iterasi ke-9	Iterasi ke-10	Itersai ke-12
b_0	1,1666	2,0780	1,3155	2,3755
b_1	2,4167	2,2252	2,3472	2,0969
$\sum_{i=1}^{n} \lvert\varepsilon_{i,m}\rvert$	59,2500	69,3387	68,0615	78,4602

Selanjutnya, dari Tabel 3 dapat diperoleh hasil penaksiran untuk keempat kelompok data dengan model penaksir $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ialah sebagai berikut:
Kelompok data 1: $\hat{Y}_i = 1,1667 + 2,4167X_i$
Kelompok data 2: $\hat{Y}_i = 2,0780 + 2,2252X_i$
Kelompok data 3: $\hat{Y}_i = 1,3155 + 2,3472X_i$
Kelompok data 4: $\hat{Y}_i = 2,3755 + 2,0969X_i$.

Secara ringkas, nilai koefisien regresi dan rata-rata kuadrat sisa untuk keempat kelompok data dan ketiga metode dapat dilihat pada Tabel 4 berikut:

Tabel 4: Hasil Estimasi Koefisien Regresi dan Rata-rata Kuadrat Sisa

Metode		OLS	LTS	M
Data 1	b_0	1,0967	0,3333	1,1667
	b_1	2,4189	2,4722	2,4167
Rata-rata Kuadrat Sisa		**13,6017**	**4,2222**	**13,6255**
Data 2	b_0	3,0687	1,6575	2,0780
	b_1	1,9646	2,1310	2,2252
Rata-rata Kuadrat Sisa		**18,8273**	**6,2294**	**15,9880**
Data 3	b_0	2,3631	0,9895	1,3155
	b_1	1,9066	2,2684	2,3472
Rata-rata Kuadrat Sisa		**20,8080**	**4,9968**	**15,7561**
Data 4	b_0	4,3351	0,8214	2,3755
	b_1	1,4522	2,3929	2,0969
Rata-rata Kuadrat Sisa		**25,2795**	**6,3679**	**19,3174**

5. KESIMPULAN

Kesimpulan dari hasil penelitian ialah:

1. Metode *Least Trimmed Squares* (LTS) menggunakan konsep pengepasan metode kuadrat terkecil untuk meminimumkan kuadrat sisa- an dari n residual menjadi h residual.

2. Dari hasil simulasi pada penelitian ini, menunjukkan bahwa LTS memberikan hasil perbandingan lebih baik daripada penaksir M dan metode OLS karena mampu menghasilkan estimasi koefisien regresi yang baik dan rata-rata kuadrat sisa paling kecil.

3. Hasil simulasi juga menunjukkan bahwa penaksir M lebih baik daripada metode OLS terutama dalam mengatasi masalah pencilan karena solusi dari penaksir M yaitu melakukan metode iterasi *weighted least squares* sehingga diperoleh model dan koefisien regresi yang cocok serta rata-rata kuadrat sisa yang lebih kecil.

Daftar Pustaka

[1] R. K. Sembiring. Analisis Regresi. Bandung: Penerbit ITB, (1995)

[2] Soemartini. Pencilan (Outlier). Jatinangor: Penerbit Universitas Padjajaran, (2007)

[3] P. J. Rousseeuw dan A. M. Leroy. Robust Regression and Outlier Detection. Canada, (1987)

[4] Drapper, N. R. dan H. Smith. Applied Regression Analysis. John Willey and Sons Inc: New york, (1992)

[5] M. S. Akbar dan L. Maftukhah. Optimasi kekuatan tourqe pada lampu TL. *Jurnal Ilmiah Sains dan Teknologi* 6(3): hal. 218 229, (2007)

[6] W. S. Winahju. Regresi Robust dengan Program Macro MINITAB, (18 Januari 2012)

SRI WULANDARI: Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Sumatera Utara, Medan 20155, Indonesia
E-mail: wul.cute@yahoo.co.id

SUTARMAN: Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Sumatera Utara, Medan 20155, Indonesia
E-mail: sutarman@usu.ac.id

OPEN DARNIUS: Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Sumatera Utara, Medan 20155, Indonesia
E-mail: open@usu.ac.id