Saintia Matematika ISSN: 2337-9197
Vol. 02, No. 03 (2014), pp. 225–235.

MENENTUKAN KOEFISIEN DETERMINASI ANTARA ESTIMASI M DENGAN TYPE WELSCH DENGAN LEAST TRIMMED SQUARE DALAM DATA YANG MEMPUNYAI PENCILAN

Sabam Daoni Sinambela, Suwarno Ariswoyo,
Henry Rani Sitepu

Abstrak: Koefisien determinasi (R^2) adalah suatu indikator yang digunakan untuk menggambarkan berapa banyak variasi yang dijelaskan dalam model. Berdasarkan nilai R^2 dapat diketahui tingkat signifikansi atau kesesuaian hubungan antara variabel bebas dan variabel tak bebas dalam regresi linier. Pencilan adalah data yang tidak mengikuti sebagian besar pola dan terletak jauh dari pusat data, dapat dideteksi dengan metode *boxplot (Interquartil Range)*, menentukan nilai Leverage, DFFITS dan Cooks Distance. *Least Trimmed Squares* (LTS) yaitu metode penaksiran parameter regresi *robust* yang menggunakan konsep pemangkasan untuk meminimumkan jumlah kuadrat residual. Penaksir M yaitu metode dalam mengatasi pencilan dan dapat menggunakan penaksir *Welsch* dalam mengestimasi parameter regresi. Tujuan penelitian ini adalah membandingkan dua metode regresi *robust* yakni penaksir LTS dan penaksir M type Welsch dalam mengatasi permasalahan data pencilan. Hasil penelitian yang diperoleh yaitu penaksir LTS merupakan metode paling baik karena mampu mengatasi pencilan dan diperoleh bahwa *Least Trimmed Squares* memiliki nilai koefisien determinasi yang paling tinggi dari penaksir M type *Welsch*.

Received 03-10-2013, Accepted 08-04-2014.
2010 Mathematics Subject Classification: 62M10, 62N02
Key words and Phrases: *Estimasi M*, *Type Welsch*, *Least Trimmed Squqres*, *Regresi Robust*, Data Pencilan.

1. PENDAHULUAN

Regresi merupakan suatu metode statistika yang digunakan untuk menyelidiki pola hubungan antara dua atau lebih variabel. Tujuan dari analisis regresi adalah untuk mengestimasi parameter model yang menyatakan pengaruh hubungan antara variabel *predictor* dan variabel *respon*.

Metode estimasi yang digunakan adalah *Ordinary Least Square* (OLS). Namun metode ini mempunyai asumsi yang pada data riil sering tidak dapat dipenuhi. Asumsi tersebut mengenai kenormalan *residual* yang sering dilanggar ketika adanya pengamatan yang bersifat *outlier*.

Outlier tidak dapat dibuang atau dihapus begitu saja dari pengamatan. Adakalanya *outlier* memberikan informasi yang tidak bisa diberikan oleh titik data lainnya, misalnya karena *outlier* timbul dari kombinasi yang tidak biasa dan perlu diselidiki lebih jauh[1].

2.LANDASAN TEORI

Pengertian Regresi Linier

Regresi secara umum adalah sebuah alat statistik yang memberikan penjelasan tentang pola hubungan antara 2 variabel atau lebih. Dalam analisis regresi dikenal 2 jenis variabel yaitu variabel *dependent* yang dinotasikan dengan Y dan variabel *independent* yang dinotasikan dengan X. Tujuan utama regresi linier adalah untuk membuat perkiraan nilai suatu variabel jika nilai variabel yang lain yang berhubungan dengannya sudah ditentukan[2].

Pendeteksian Data Pencilan *Outlier*

Untuk mendeteksi pencilan dapat dilakukan dengan *boxplot* yaitu dengan menentukan *Interquartil Range* (IQR) dan dirumuskan dengan
Q_1: Kuartil 1
Q_3: Kuartil 3

Batas bukan data pencilan adalah data yang kurang dari 1,5x IQR terhadap Q_1 dan lebih dari 1,5x IQR terhadap Q_3.

Metode Kuadrat Terkecil

Metode Kuadrat Terkecil adalah metode yang digunakan ketika terjadi penyimpangan antara nilai yang sebenarnya dengan nilai suatu taksiran. Penyimpangan tersebut dinamakan dengan e_i dan ditaksir dengan $Y_i - \widehat{Y_i}$.

Metode Kudarat Terkecil meminimumkan nilai e_i^2 dan menghasilkan $a = \frac{\sum_{i=1}^{n} Y_i}{n}$ dan $b = \frac{\sum_{i=1}^{n} Y_i X_i - Y_i \frac{\sum_{i=1}^{n} X_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\sum_{i=1}^{n} X_i^2}{n}}$.

Dengan:
Y_i: Variabel terikat ke-i.
X_i: Variabel bebas ke-i.

Dalam perhitungannya, terlebih dahulu dihitung nilai b kemudian nilai tersebut digunakan untuk mendapatkan nilai a.

Regresi *Robust*

Regresi *robust* adalah suatu metode yang digunakan untuk mencari persaman terbaik dalam data yang mengandung *outlier*.

Dalam regresi *robust* banyak metode *estimasi* yang bisa digunakan seperti penaksir *Least Median Square, Least Trimmed Square*, Penaksir M, Penaksir S dan Penaksir MM.

Regresi *Robust* Dengan *Least Trimmed Square*

Least Trimmed Square (LTS) merupakan suatu metode pendugaan parameter regresi *robust* untuk meminimumkan jumlah kuadrat h *residual*.

Tahapan algoritma *Least Trimmed Square* adalah:

1. Menghitung $\widehat{Y}$ berdasarkan nilai parameter.
2. Menghitung *coverage* (h).
3. Menghitung $\sum_{i=1}^{h} r_i^2$.
4. Melakukan estimasi parameter dari h pengamatan.
5. Menentukan $\widehat{Y}$ berdasarkan nilai parameter yang baru.
6. Melakukan iterasi sampai mendapatkan koefisien determinasi yang relatif lebih baik.

Regresi *Robust* Dengan Estimasi *M Type Welsch*

Estimasi M didasarkan ide penggantian *residual* kuadrat sehingga menghasilkan fungsi *residual minimum*.

Residual minimum dirumuskan dengan *minimize* $\sum_{i=1}^{n} \rho(r_i)$, dimana ρ adalah fungsi simetris dengan nilai minimum sama dengan 0 dan memerlukan standarisasi *residual* berupa pendekatan dari sebuah σ yang menghasilkan $\sum_{i=1}^{n} \psi(\frac{r_i}{\widehat{\sigma}})X^T = 0$.

Persamaan dapat ditulis menjadi $\sum_{i=1}^{n} w(\frac{Y_i - X_b}{\widehat{\sigma}})X^T = 0$ atau $X^T W X b = X^T W Y$. W adalah matriks diagonal dari w dengan ukuran $n \times n$.

Kedua ruas dikalikan dengan $(X^T W X)$ menghasilkan nilai $b = (X^T WX)^{-1} X^T W Y$[3].

Algoritma penyelesaian dari estimasi M dengan type *Welsch* adalah:

1. Menghitung $\widehat{Y}_i$ dan ε_i berdasarkan nilai dari masing-masing parameter.
2. Menghitung $\widehat{\sigma}_i$ dari nilai-nilai *residual*.
3. Menyusun matrik pembobot berupa matrik diagonal dengan elemen $w_{1,1}, w_{2,1}, ..., w_{n,1}$ dan dinamai dengan W_0.
4. Menghitung nilai β_{Robust_1} .
5. Menghitung nilai $\sum_{i=1}^{n} abs(Y_i - \widehat{Y}_i)$ atau $\sum_{i=1}^{n} abs(\varepsilon_i)$.
6. Langkah 2 sampai 5 diulang sampai diperoleh $\sum_{i=1}^{n}[\varepsilon_{i,m}]$ yang konvergen.

Koefisien Determinasi

R^2 adalah suatu indikator yang menggambarkan berapa banyak variasi yang dijelaskan dalam model[4]. Nilai dari R^2 dapat dicari dengan menggunakan rumus:

$$R^2 = \frac{b_1 \sum x_1 y + b_2 \sum x_2 y + ... + b_n \sum x_n y}{\sum y^2}$$

3.METODE PENELITIAN

Adapun metode penelitian yang digunakan penulis adalah:

1. Menentukan data.
2. Menentukan ada tidaknya *outlier* dengan metode *boxplot*.
3. Menghitung koefisien determinasi dengan *Least Trimmed Square* dan estimasi M dengan type $Welsch$.
4. Membandingkan kedua koefisien determinasi tersebut untuk memperoleh estimasi yang relatif lebih baik.

4.PEMBAHASAN

Data yang diambil adalah data pengukuran keasinan garam dan arus sungai di *Carolinas Pamlico Sound* Utara dari buku *Robust Regression And Outlier Detection* seperti pada Tabel 1.

Tabel 1: Salinity Data

Index (i)	Lagged Salinity (X_1)	Trend (X_2)	Discharge (X_3)	Salinity (Y)
1	2,00	4	23,01	7,60
2	7,60	5	23,87	7,70
3	4,60	0	26,42	4,30
4	4,30	1	24,86	5,90
⋮	⋮	⋮	⋮	⋮
18	7,70	3	22,69	9,50
19	10,00	0	21,79	12,00
20	12,0	1	22,04	12,60
21	12,10	4	21,03	13,60
22	13,60	5	21,01	14,11
23	15,00	0	25,87	13,52

Sumber: *Robust Regresion And outlier Detection* 1986

Berikut ini adalah pendeteksian masing-masing variabel dengan metode *box-plot* dan masing-masing variabel telah diurutkan dari variabel terkecil ke variabel terbesar.

Pendeteksian data *Lagged Salinity*:

$$\begin{aligned} Q_1 &= \frac{X_7 + X_8}{2} = \frac{7,7 + 8,2}{2} = 7,95 \\ Q_2 &= \frac{X_{21} + X_{22}}{2} = \frac{13 + 13,1}{2} = 13,05 \\ Q_3 - Q_1 &= 13,05 - 7,95 = 5,1 \\ 1,5 \quad IQR &= 1,5 \times 5,1 = 7,65 \end{aligned}$$

Pendeteksian data *Trend, Discharge* dan *Salinity* memiliki nilai Q_1, Q_2, Q_3 dan IQR yang ditunjukkan dalam Tabel 2.

Tabel 2: Tabel IQR

Variabel	Nilai Q_1	Nilai Q_3	Nilai IQR
X_1	7,95	13,05	7,65
X_2	1,00	4,00	1,50
X_3	21,78	24,87	4,62
Y	7,95	13,05	7,65

Penyelesaian dengan *Least Trimmed Square*:

1. Dengan SPSS. 17 diperoleh $\widehat{Y}$= 9,590 + $0,77X_1$ - $0,26X_2$ - $0,295X_3$.

Iterasi I

2. *Coverage* (h)= $\frac{(n+p+1)}{2} = \frac{(28+3+1)}{2} = 16$.

Dengan h = 16, *residual* terkecil sampai yang terbesar adalah:

Tabel 3: Kuadrat Residual

No	Residual Kuadrat	No	Residual Kuadrat
1	0,01	9	0,29
2	0,01	10	0,32
⋮	⋮	⋮	⋮
5	0,19	13	0,89
6	0,20	14	1,09
7	0,28	15	1,14
8	0,29	16	1,15

3. $\hat{\beta}_{new} = \sum_{i=1}^{h_{new}} r_i^2 = 6,78126$.
4. Dengan estimasi paramter diperoleh $\widehat{Y}$= 94,023 + $0,731X_1 + 0,731X_2$ - $0,286X_2 + 0,446X_3$.

Iterasi II

1. *Coverage*= $\frac{(n+p+1)}{2} = \frac{(16+3+1)}{2} = 10$.
2. Seperti iterasi 1 diperoleh nilai $\hat{Y} = 1.498.591 + 0,679X_1 - 0,9X_2 - 0,302X_3$.
3. Dengan SPSS 17 diperoleh koefisien determinasi= 0,945.

Penyelesaian dengan Estimasi M dengan type *Welsch*:

1. Data yang digunakan adalah Tabel 1 (*Salinity Data*).
2. Nilai *residual* masing-masing varibel ditunjukkan pada Tabel 4.

Tabel 4: Nilai Residual

No	X_1	X_2	X_3	Y	$\hat{Y}$	$\varepsilon_{i,0} = Y - \hat{Y}$	$\lvert Y - \hat{Y}\rvert$
1	8,20	4	23,01	7,60	8,13	-0,54	0,54
2	7,60	5	23,87	7,70	7,15	0,55	0,55
3	4,60	0	26,42	4,30	5,37	-1,07	1,07
4	4,30	1	24,87	5,90	5,34	0,57	0,56
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	13,20	1	23,83	12,60	12,56	0,04	0,04
11	12,60	2	25,14	10,40	11,44	-1,04	1,04
12	10,40	3	22,43	10,80	10,27	0,53	0,53
13	10,80	4	21,78	13,10	10,52	2,59	2,59
14	13,10	5	22,38	12,30	11,87	0,43	0,43
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	14,10	5	21,39	15,10	12,93	2,16	2,16
Jumlah				295,50	279,07	16,43	32,59

Iterasi I

3. Nilai $\hat{\sigma}_0$ dengan c = 2,3849 adalah:

   $\widehat{\sigma}_0 = \frac{MAR}{0{,}6745} = \frac{\frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{Y}_i]}{0{,}6475}$.
   $\widehat{\sigma}_0 = \frac{\frac{1}{28}(32{,}58450)}{0{,}6475} = \frac{1{,}1637}{0{,}6745} = 1,7253$.

4. Nilai $\beta_{robust} = (X^T W_0 X)^{-1} X^T W_0 Y$. Dengan *Mathlab* diperoleh nilai
   $$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 10,9825 \\ 0,7686 \\ -0,0725 \\ -0,3476 \end{pmatrix}$$

5. Iterasi selanjutnya dilakukan seperti iterasi I. Nilai masing-masing iterasi ditunjukkan pada Tabel 5.

Tabel 5: Nilai $\beta_0, \beta_1, \beta_2, \beta_3$ dengan Iterasi

	β_0	β_1	β_2	β_3
Iterasi 1	10,98	0,77	-0,073	-0,35
Iterasi 2	14,11	0,75	-0,12	-0,47
Iterasi 3	16,44	0,73	-0,15	-0,55
Iterasi 4	17,53	0,72	-0,17	-0,59
Iterasi 5	17,97	0,72	-0,18	-0,61
Iterasi 6	18,15	0,72	-0,18	-0,62
Iterasi 7	18,23	0,72	-0,19	-0,62
Iterasi 8	18,26	0,72	-0,19	-0,62
Iterasi 9	18,28	0,72	-0,19	-0,62
Iterasi 10	18,28	0,72	-0,19	-0,62
Iterasi 11	18,28	0,72	-0,19	-0,62
Iterasi 12	18,29	0,72	-0,19	-0,62
Iterasi 13	18,29	0,72	0,19	-0,62

6. Koefisien determinasi dihitung dengan menggunakan rumus:

$$
\begin{aligned}
R^2 &= \frac{b_1 \sum x_1 y + b_2 \sum x_2 y + b_3 \sum x_3 y}{\sum y^2} \\
\sum x_1 y &= \sum_{i=1}^{n} X_1 Y_i - \frac{\sum X_1 \sum_{i=1}^{n} Y_i}{n} = 3.272,6100 - 219,4618 \\
&= 157,3102. \\
\sum x_2 y &= \sum_{i=1}^{n} X_2 Y_i - \frac{\sum X_2 \sum_{i=1}^{n} Y_i}{n} = 756,5000 - 738,7500 \\
&= 17,7500. \\
\sum x_3 y &= \sum_{i=1}^{n} X_3 Y_i - \frac{\sum X_3 \sum_{i=1}^{n} Y_i}{n} = 6.904,3464 - 7.013,0909 \\
&= -108,7445.
\end{aligned}
$$

$$\begin{aligned}\sum y^2 &= \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 3.363,2300 - 3.118,5804 \\ &= 244,6496.\end{aligned}$$

$$\begin{aligned}R^2 &= \frac{(0,7168 \times 157,3102) - (0,168 \times 17,7500) +}{244,6496} \\ &\quad \frac{(0,6232 \times 108,7445)}{244,6496}.\end{aligned}$$

$$R^2 = 0,9065 = 90,65\%.$$

Jadi koefisien determinasinya adalah 90,65.

5. KESIMPULAN

Kesimpulan yang dapat diambil adalah:

1. *Least Trimmed Square* merupakan metode yang relatif lebih baik. Hal ini dapat terlihat dari koefisien determinasi yang relatif lebih tinggi meskipun memiliki selisih yang kecil.
2. Iterasi yang lebih banyak menghasilkan koefisien determinasi yang lebih baik.

Daftar Pustaka

[1] Draper, N.R dan H. Smith. 1992. Analisis Regresi Terapan. Bambang Sumantri. Gramedia.Jakarta, (1992).

[2] Hasan Iqbal. Pokok-pokok materi statistik. Penerbit Bumi Aksara. Jakarta, (1999).

[3] Alfigari. Analisis Regresi. Sekolah Tinggi Ilmu Ekonomi YKPN. Yogyakarta, (2002).

[4] Dixon J,Wilfrid dan Massey J.Frank. Pengantar Analisis Statistik. Universitas Gadjah Mada. Yogyakarta, (1991).

[5] Cahmayati Dian dan Tanuji Hadi. Efektivitas Metode Regresi *Robust* Penduga *Welsch* dalam mengatasi Pencilan pada Pemodelan Regresi Linier Berganda. Universitas Sriwijaya, (2009).

Sabam Daoni Sinambela: *Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of North Sumatera*, Medan 20155, Indonesia.
E-mail: sabamdaoni@gmail.com

Suwarno Ariswoyo: *Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of North Sumatera*, Medan 20155, Indonesia.
E-mail: suwarno@usu.ac.id

Henry Rani Sitepu: *Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of North Sumatera*, Medan 20155, Indonesia.
E-mail: henry1@usu.ac.id