

PERBANDINGAN METODE *COEFFICIENT OF DETERMINATION RATIO* DAN REGRESI DIAGNOSTIK DALAM MENDETEKSI *OUTLIER* PADA ANALISIS REGRESI LINIER BERGANDA

Emi, Dadan Kusnandar, Naomi Nussyana Debataraja

INTISARI

Outlier merupakan suatu pengamatan yang keberadaannya dapat mengganggu proses analisis data. Adanya *outlier* juga dapat mengganggu asumsi kenormalan. Oleh karena itu, mendeteksi *outlier* sangat diperlukan sebelum analisis data dilakukan. Penelitian ini membahas metode untuk mendeteksi *outlier* menggunakan metode *coefficient of determination ratio (CDR)* dan regresi diagnostik lainnya seperti: *leverage values*, *studentized deleted residuals*, *Cook's distance*, *difference in fits standardized (DFITS)* dan *covariance ratio*. Metode *coefficient of determination ratio (CDR)* menggunakan nilai koefisien determinasi (R^2) sebagai dasar proses deteksi dan diterapkan pada data hasil simulasi. Hasil perbandingan menunjukkan bahwa metode regresi diagnostik berdasarkan nilai *studentized deleted residuals* atau *R-Student* lebih efektif dalam mendeteksi *outlier* dibandingkan dengan metode *coefficient of determination ratio (CDR)*. Keefektifan dilihat berdasarkan ketepatan dalam mendeteksi data *outlier* dan nilai koefisien determinasi (R^2) yang meningkat. Namun demikian hasil pendeteksian dengan metode *coefficient of determination ratio (CDR)* juga berhasil meningkatkan nilai koefisien determinasi. Oleh sebab itu metode *coefficient of determination ratio (CDR)* dapat digunakan sebagai metode alternatif untuk mendeteksi *outlier*.

Kata kunci: Simulasi, Diagnostik, nilai cut-off

PENDAHULUAN

Analisis regresi linier berganda merupakan suatu analisis statistika yang digunakan untuk mengetahui hubungan antara satu variabel dependen dengan lebih dari satu variabel independen [1]. Untuk memperoleh suatu model regresi yang baik maka model-model tersebut harus memenuhi asumsi-asumsi pada regresi, diantaranya asumsi normalitas, homoskedastisitas, autokorelasi dan kehadiran *outlier*. *Outlier* merupakan suatu pengamatan yang tidak konsisten dengan kumpulan pengamatan lainnya [2], [3]. Zakaria *et al.*, [4] memperkenalkan metode *coefficient of determination ratio (CDR)* sebagai standar pengukuran alternatif untuk mendeteksi *outlier* berdasarkan nilai koefisien determinasi (R^2) dimana penentuan keberadaan *outlier* menggunakan metode CDR hanya berdasarkan tampilan visual.

Tujuan dari penelitian ini adalah untuk membandingkan tingkat efektifitas metode *coefficient of determination ratio (CDR)* dan metode regresi diagnostik. Perbandingan keefektifan dari metode yang digunakan dilihat berdasarkan ketepatan dalam mendeteksi data *outlier* dan nilai koefisien determinasi (R^2). Dalam penelitian ini metode CDR dan regresi diagnostik diterapkan terhadap data hasil simulasi dengan hanya dua variabel independen. Data hasil simulasi dibangkitkan mengikuti model regresi berganda dengan kondisi data tanpa kontaminasi *outlier*, kontaminasi 10% dan 20% *outlier* pada *error*. Untuk setiap simulasi dibangkitkan sebanyak 30 data berdistribusi normal.

REGRESI DIAGNOSTIK

Secara umum model regresi berganda dengan lebih dari satu variabel independen diberikan sebagai berikut [1].

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

atau dapat ditulis dalam notasi matriks yaitu:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

Regresi diagnostik merupakan teknik untuk menyelidiki masalah yang mempengaruhi dalam analisis regresi dan untuk menentukan kelayakan asumsi [5]. Regresi diagnostik digunakan untuk memeriksa tiga karakteristik data yang secara potensial merupakan *outlier* yaitu *leverage*, *discrepancy* dan *influence* [6]. *Leverage* mengukur seberapa jauh nilai suatu kasus pada variabel independen (x). *Discrepancy* mengukur perbedaan antara nilai prediksi variabel dependen ($\hat{y}$) dengan nilai observasi dari variabel hasil (y), yaitu $y_i - \hat{y}$ yang merupakan nilai *residual* (ε_i), dihitung menggunakan metode *externally studentized residual* atau *studentized deleted residuals* atau biasa juga disebut sebagai *R-student*. Sedangkan *influence* menggambarkan besaran dari perubahan koefisien regresi jika *outlier* dihapus dari himpunan data. Secara konseptual, *influence* mempresentasikan perkalian dari *leverage* dan *discrepancy*. Nilai *influence* bisa dihitung dengan metode *Cook's distance*, *difference in fits standardized (DFFITS)* dan *covariance ratio*.

1. Leverage Values (h_{ii})

Leverage values digunakan untuk mendeteksi *outlier* pada variabel independen. Nilai *leverage* berasal dari nilai diagonal matriks *hat* ($\mathbf{H}$) yang berukuran $(n \times n)$.

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3)$$

dengan $\mathbf{X}$ merupakan matriks $n \times (k + 1)$ dengan n merupakan banyaknya data, dan k merupakan banyaknya variabel independen. Diagonal dari $\mathbf{H}$ berisi nilai-nilai *leverage*. *Leverage* untuk kasus ke- i (h_{ii}) merupakan nilai dari baris ke- i dan kolom ke- i dari $\mathbf{H}$ [6].

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad (4)$$

dimana $\mathbf{x}_i$ adalah baris ke- i dari $\mathbf{X}$. Penentuan nilai yang memiliki *leverage* yang besar didasarkan pada nilai *cut-off* $h_{ii} > \frac{2(k+1)}{n}$. Nilai h_{ii} yang melebihi nilai *cut-off* dideteksi sebagai *outlier* [1].

2. Studentized Deleted Residuals (t_i)

Studentized deleted residuals merupakan *error* yang digunakan untuk mendeteksi keberadaan *outlier* pada nilai *residual* yang didefinisikan sebagai berikut [7]:

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1-h_{ii}}} \quad (5)$$

Dimana $s_{(i)}$ adalah simpangan baku yang dihitung tanpa mengikutsertakan pengamatan ke- i , dengan nilai

$$s_{(i)} = \sqrt{\frac{SSE_{(i)}}{n-k-2}} \quad (6)$$

Suatu pengamatan dicurigai sebagai *outlier* apabila pengamatan tersebut memiliki nilai *cut-off* $|t_i| > t_{\frac{\alpha}{2}; n-k-2}$ [5].

3. Cook's Distance (D_i)

Cook (1977) memperkenalkan ukuran "jarak (*distance*)" antara nilai $\hat{\beta}_i$ dengan $\hat{\beta}_{(i)}$, dimana $\hat{\beta}_{(i)}$ diperoleh ketika pengamatan ke- i dihapus. *Cook's distance* merupakan suatu ukuran pengaruh pengamatan ke- i terhadap semua koefisien regresi. *Cook's distance* untuk pengamatan ke- i didefinisikan:

$$D_i = \frac{\varepsilon_i^2}{k+1} \left(\frac{h_{ii}}{1-h_{ii}} \right) \quad (7)$$

dimana $\varepsilon_i' = \frac{\varepsilon_i}{s\sqrt{1-h_{ii}}}$ dan $s = \sqrt{\sum \varepsilon_i^2 / n - k - 1}$ dengan nilai *cut-off* $D_i > \frac{4}{n-(k+1)}$ [5].

4. *Difference in Fits Standardized (DFFITS)*

DFFITS digunakan untuk mendeteksi *outlier* berpengaruh pada nilai estimasi variabel dependen yang diberikan oleh

$$DFFITS_i = t_i \left[\frac{h_{ii}}{1-h_{ii}} \right]^{1/2} \quad (8)$$

Dimana nilai *cut-off* DFFITS adalah jika $|DFFITS_i| > 2 \sqrt{\frac{k+1}{n}}$ [6].

5. *Covariance Ratio (CVR)*

Covariance ratio merupakan suatu ukuran yang menggambarkan pengaruh suatu pengamatan ke-*i* terhadap ketelitian estimasinya. Untuk menghitung nilai *covariance ratio* setiap pengamatan ke-*i* ditentukan oleh

$$CVR_i = \left(\frac{s_{(i)}^2}{s^{2k}} \right)^k \left(\frac{1}{1-h_{ii}} \right) \quad (9)$$

Nilai *covariance ratio* dikatakan berpengaruh apabila *cut-off* $CVR_i > 1 + \frac{3(k+1)}{n}$ atau $CVR_i < 1 - \frac{3(k+1)}{n}$ [1].

METODE COEFFICIENT OF DETERMINATION RATIO (CDR)

Coefficient of determination ratio (CDR) merupakan salah satu metode alternatif yang dapat digunakan untuk mendeteksi *outlier* dengan cara menghapus pengamatan ke-*i* dari himpunan data. Penghapusan pengamatan bertujuan untuk meningkatkan nilai koefisien determinasi (R^2). Cara kerja metode CDR untuk mendeteksi *outlier* berdasarkan nilai *coefficient of determination* (R^2) dari model analisis regresi linier berganda adalah sebagai berikut:

1. Membangun model regresi dan tentukan nilai R^2 .
2. Hitung nilai R^2 saat pengamatan ke-*i* dihapus dari himpunan data ($R_{(i)}^2$).
3. Bandingkan nilai R^2 dengan $R_{(i)}^2$ dengan mengambil rasionya.
4. Metode ini disebut sebagai *metode coefficient of determination ratio*.

Pendeteksian *outlier* menggunakan CDR untuk pengamatan ke-*i* didefinisikan sebagai berikut [4]:

$$CDR_i = \frac{R_{(i)}^2}{R^2}, \quad i = 1, 2, \dots, n \quad (10)$$

Pengamatan dikatakan *outlier* jika nilai *cut-off* untuk metode CDR yaitu nilai batas yang digunakan untuk menentukan apakah suatu pengamatan terdeteksi sebagai *outlier* atau bukan adalah jika $CDR_i > \overline{CDR} + 1,65.Sd_{cdr}$ atau $CDR_i < \overline{CDR} - 1,65.Sd_{cdr}$ dimana $\overline{CDR}$ adalah rata-rata CDR dan Sd_{cdr} adalah simpangan baku CDR.

HASIL DAN PEMBAHASAN

Data yang digunakan merupakan data hasil simulasi, banyak data yang digunakan adalah $n = 30$, nilai parameter regresi yang sudah ditetapkan yaitu $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$ untuk setiap kondisi data. Kondisi data tersebut berhubungan dengan tingkatan dari persentase *outlier*, yaitu data tanpa kontaminasi *outlier*, 10% dan 20% *outlier* pada *error* (ε) yang diasumsikan mengikuti distribusi normal. Untuk data tanpa kontaminasi *outlier* diperoleh dengan cara membangkitkan $n = 30$ pengamatan dari data mengikuti distribusi normal dengan rata-rata nol dan variansi satu. Sedangkan data dengan kontaminasi 10% *outlier* diperoleh dengan cara membangkitkan $10\% \times 30 = 3$ data pengamatan yang berdistribusi normal dengan rata-rata dan variansi tertentu, sehingga didapatkan data

outlier yang jauh lebih besar dari sebaran populasi berdistribusi normal, kemudian data *outlier* digabungkan ke dalam data tanpa kontaminasi *outlier* pada *error*. Penggabungan dilakukan dengan cara mengganti sejumlah pengamatan secara acak pada *error* data asli. Proses yang sama dilakukan untuk mendapatkan data yang lain pada kasus kontaminasi 20% *outlier*. Prosedur secara umum pengerjaan peneliti ini adalah sebagai berikut:

Langkah kerja yang dilakukan pertama kali adalah membangun model regresi linier berganda dari data hasil simulasi dan hitung nilai koefisien determinasi (R^2) untuk setiap kondisi data. Kedua menggunakan metode regresi diagnostik dan CDR untuk mendeteksi keberadaan *outlier* kemudian dibandingkan dengan nilai *cut-off*. Sebelumnya terlebih dahulu dilakukan perhitungan untuk nilai *cut-off* dari setiap metode. Pengamatan yang mempunyai nilai lebih besar dari nilai *cut-off* merupakan *outlier*. Ketiga, membentuk model regresi tanpa mengikutsertakan pengamatan yang terdeteksi sebagai *outlier* dan hitung kembali nilai R^2 . Terakhir, bandingkan nilai R^2 dari sebelum dan sesudah data *outlier* dihapus antara metode regresi diagnostik dengan metode CDR serta menggunakan indeks *plot* untuk memvisualisasikan keberadaan *outlier*.

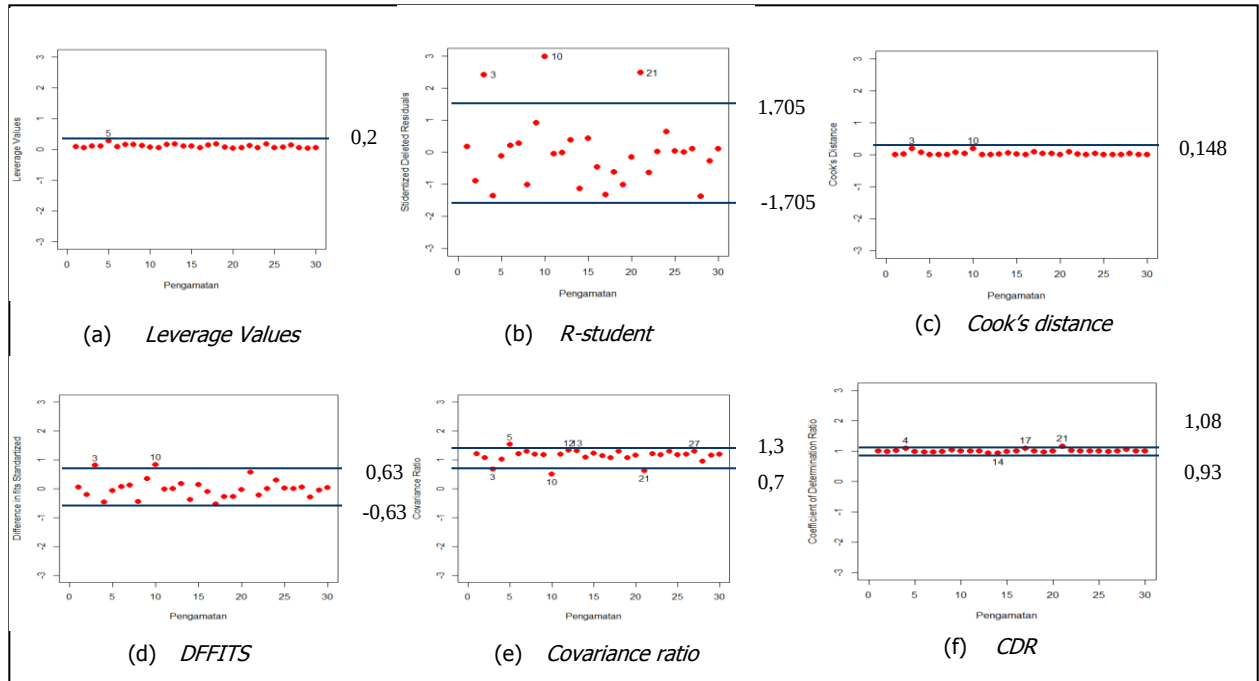
Sebagai ilustrasi disajikan data hasil pendeteksian *outlier* untuk kondisi data kontaminasi 10% *outlier* dari semua metode pada Tabel 1.

Tabel 1. Data hasil pendeteksian untuk kondisi data kontaminasi 10% *outlier*

No	Leverage	R-Student	Cook	DFFITS	Covratio	CDR
1	0,082	0,167	8,55E-04	0,049	1,215	1,001
⋮						
3	0,099	2,420*	1,83E-01*	0,806*	0,676*	1,018
4	0,107	-1,372	7,26E-02	-0,474	1,017	1,090*
5	0,269*	-0,124	1,96E-03	-0,075	1,531*	0,985
⋮						
10	0,073	2,986*	1,80E-01*	0,835*	0,499*	0,999
⋮						
12	0,151	-0,011	7,93E-06	-0,005	1,319*	0,998
13	0,162	0,376	9,43E-03	0,166	1,315*	0,936
14	0,105	-1,139	5,01E-02	-0,390	1,081	0,922*
⋮						
17	0,136	-1,339	9,13E-02	-0,531	1,061	1,085*
⋮						
21	0,049	2,479*	8,86E-02	0,563	0,623*	1,158*
⋮						
27	0,140	0,110	6,79E-04	0,044	1,300*	0,999
⋮						
30	0,055	0,103	2,12E-04	0,025	1,183	0,991

Data hasil pendeteksian *outlier* pada Tabel 1 yang bertanda bintang (*) merupakan pengamatan yang terdeteksi sebagai *outlier* untuk kondisi data kontaminasi 10% *outlier*. Nilai *cut-off leverage* (h_{ii}) yaitu $\frac{2(k+1)}{n} = 0,2$. Dari Tabel 1 terlihat bahwa nilai *leverage* yang terdeteksi sebagai *outlier* yaitu pengamatan ke-5. Untuk mengevaluasi nilai *studentized deleted residuals* (t_i) atau *R-student* pada pengamatan dengan taraf nyata $\alpha = 10\%$ sehingga diperoleh nilai *cut-off* $t_{\frac{\alpha}{2}; n-k-2} = 1,705$ terlihat bahwa t_i untuk pengamatan {3,10,21} lebih besar dari $t_{0,05;26} = 1,705$. Nilai *Cook's distance* (D_i) pada Tabel 1 pengamatan {3,10} terdeteksi sebagai *outlier* karena nilai D_i lebih besar dari nilai *cut-off*nya yaitu $D_i > \frac{4}{n-(k+1)} = 0,148$. Nilai DFFITS yang diidentifikasi sebagai *outlier* adalah pengamatan yang nilai $|DFFITS| > 2\sqrt{(k+1)/n} = 0,632$. Berdasarkan Tabel 1 pengamatan yang terdeteksi *outlier* oleh DFFITS adalah pengamatan {3,10}. Dari Tabel 1 terlihat bahwa pengamatan yang terdeteksi sebagai *outlier* oleh *covariance ratio* (CVR) yaitu pengamatan {3,5,10,12,13,21,27} dengan nilai *cut-off CVR* $> 1 + \frac{3(k+1)}{n} = 1,3$ atau $CVR < 1 - \frac{3(k+1)}{n} = 0,7$. Sedangkan, pengamatan yang terdeteksi oleh CDR sebagai *outlier* dengan nilai $\overline{CDR} = 1,002$ dan $Sd_{cdr} = 0,046$ sehingga diperoleh nilai *cut-off CDR* $> 1,08$ atau $CDR_i < 0,93$ yaitu pengamatan

{4, 14, 17, 21}. Penghapusan pengamatan yang terdeteksi sebagai *outlier* dari himpunan data diharapkan untuk meningkatkan nilai R^2 . Oleh karena itu pengamatan yang diidentifikasi sebagai *outlier* dianggap mempengaruhi analisis regresi. Untuk memvisualisasikan keberadaan *outlier* disajikan pada Gambar 1.



Gambar 1. (a), (b), (c), (d), (e), (f) Indeks *plot* hasil pendeteksian *outlier* pada data kontaminasi 10% *outlier* dengan metode regresi diagnostik dan CDR

Adapun hasil dari pendeteksian data *outlier* dari semua metode dengan kondisi data tanpa kontaminasi *outlier*, kontaminasi 10% dan 20% data *outlier*, serta nilai koefisien determinasi (R^2) dari data asli yaitu data sebelum *outlier* dihapus dan nilai R^2 terkoreksi yaitu setelah data yang terdeteksi sebagai *outlier* dihapus disajikan pada Tabel 2.

Tabel 2. Nilai koefisien determinasi (R^2) data asli dan data terkoreksi

Persentase <i>outlier</i> (%)	Metode	Pengamatan yang terdeteksi sebagai <i>outlier</i>	Nilai R^2 data asli	R^2 data terkoreksi	Peningkatan nilai R^2 setelah data <i>outlier</i> dihapus
0%	<i>Leverage</i>	5	0,639	0,642	0,003
	<i>R-Student</i>	9, 15, 28		0,734	0,095
	<i>Cook's D</i>	9		0,682	0,043
	<i>DFFITS</i>	9, 17		0,722	0,083
	<i>Covratio</i>	5, 12, 18		0,615	-0,024
	<i>CDR</i>	4, 9, 13, 17, 28		0,788	0,149
10%	<i>Leverage</i>	5	0,429	0,423	-0,006
	<i>R-Student</i>	3, 10, 21		0,618	0,189
	<i>Cook's D</i>	3, 10		0,459	0,030
	<i>DFFITS</i>	3, 10		0,459	0,030
	<i>Covratio</i>	3, 5, 10, 12, 13, 21, 27		0,556	0,127
	<i>CDR</i>	4, 14, 17, 21		0,560	0,131
20%	<i>Leverage</i>	5	0,189	0,203	0,014
	<i>R-Student</i>	12, 13, 14, 19, 25		0,424	0,235
	<i>Cook's D</i>	12, 13		0,133	-0,056
	<i>DFFITS</i>	12, 13		0,133	-0,056
	<i>Covratio</i>	5, 18, 19, 24		0,253	0,064
	<i>CDR</i>	13, 14, 19		0,288	0,099

Dari Tabel 2 terlihat bahwa nilai koefisien determinasi (R^2) data asli untuk kondisi data tanpa *outlier* adalah 0,639 yang artinya variasi nilai variabel dependen (y) dapat dijelaskan sebesar 63,9%

oleh variabel independen (x_1, x_2) . Akan tetapi, karena adanya penambahan *outlier* menyebabkan semakin kecil nilai yang dapat direpresentasikan oleh model regresi yaitu hanya sebesar 42,9% untuk data kontaminasi 10% *outlier* dan 18,9% untuk data kontaminasi 20% *outlier*. Berdasarkan Tabel 2 setelah pengamatan yang dideteksi sebagai *outlier* dihapus perubahan terbesar nilai R^2 pada metode CDR untuk kondisi data tanpa kontaminasi *outlier* dimana nilai R^2 meningkat sebesar 0,149 yaitu dari 0,639 menjadi 0,788 dan untuk data kontaminasi 10% *outlier* nilai R^2 meningkat sebesar 0,189 yaitu dari 0,429 menjadi 0,618 yang dideteksi oleh *R-student*. Sedangkan perubahan terbesar nilai R^2 untuk kondisi data kontaminasi 20% *outlier* yaitu sebesar 0,235 dideteksi oleh metode *R-student*. Perbandingan nilai R^2 sebelum dan setelah himpunan data yang dicurigai sebagai *outlier* dihapus pada Tabel 2 menunjukkan bahwa *outlier* memberikan pengaruh terhadap perubahan nilai R^2 .

KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa pendeteksian *outlier* dapat dilakukan menggunakan metode regresi diagnostik diantaranya: *leverage values*, *studentized deleted residuals*, *Cook's distance*, *difference in fits standardized (DFFITS)*, *covariance ratio* dan metode *coefficient of determination ratio (CDR)*. Hasil perbandingan menunjukkan bahwa metode regresi diagnostik berdasarkan nilai *studentized deleted residuals* atau *R-Student* lebih efektif dalam mendeteksi *outlier* dibandingkan dengan metode *coefficient of determination ratio (CDR)*. Keefektifan dilihat berdasarkan ketepatan dalam mendeteksi data *outlier* dan nilai koefisien determinasi (R^2) yang meningkat. Namun demikian hasil pendeteksian dengan metode *coefficient of determination ratio (CDR)* juga berhasil meningkatkan nilai koefisien determinasi. Oleh sebab itu metode *coefficient of determination ratio (CDR)* dapat digunakan sebagai metode alternatif untuk mendeteksi *outlier*.

DAFTAR PUSTAKA

- [1] Yan Xin and Xiao Gang Su. *Linier Regression Analysis*. Singapore: World Scientific; 2009.
- [2] Barnett V and Levis T. *Outliers in Statistical Data*. New York: John Wiley & Sons; 1980.
- [3] Draper NR. and Smith H. *Analisis Regresi Terapan* [Sumantri B]. Jakarta: Ed ke-2 PT Gramedia Pustaka Utama; 1992.
- [4] Zakaria A, Howard NK, and Nkansah BK. On the Detection of Influential *Outliers* in Linear Regression Analysis. *American Journal of Theoretical and Applied Statistics*. 2014; 3(4):100-106.
- [5] Fox J. *Regressions Diagnostics*. Newbury Park: Sage Publications; 1991.
- [6] Cohen J, Cohen P, West SG. and Aiken LS. *Applied Multiple Regression/Correlation Analysis For the Behavioral Sciences*. New Jersey: Thrid Edition Lawrence Erlbaum Associate; 2003.
- [7] Rencher AC and Schaalje GB. *Linier Model in Statistics*, New Jersey: Second Edition John Wiley & Sons; 2008.

EMI	: Jurusan Matematika FMIPA UNTAN, Pontianak, aai_emy@yahoo.com
DADAN KUSNANDAR	: Jurusan Matematika FMIPA UNTAN, Pontianak, dkusnand@yahoo.com
NAOMI NESSYANA DEBATARAJA	: Jurusan Matematika FMIPA UNTAN, Pontianak, naominessyana@gmail.com