Regresi TELBS untuk Mengatasi Masalah Pencilan

Nurul Gusriani[1], Firdaniza[2], Novi Octavianti[3]

[1,2,3]*Departemen Matematika FMIPA Univesitas Padjadjaran*
Jalan Raya Bandung-Sumedang Km. 21 Jatinangor 45363
E-mail : [1]nurul.gusriani@unpad.ac.id, [2]firdaniza@unpad.ac.id, [3]novioctaa@gmail.com

ABSTRAK

Salah satu metode yang digunakan untuk memodelkan regresi linier berganda pada data yang mengandung pencilan adalah metode TELBS. Metode ini menghasilkan model yang dapat mewakili sebagian besar data. Paper ini menggunakan data simulasi dengan membangkitkan data berdistribusi normal (0,1) dan data sekunder mengenai produktivitas primer fitoplankton. Dengan menghitung koefisien determinasi berdasarkan metode kuadrat terkecil dan metode TELBS diperoleh hasil bahwa metode TELBS adalah metode yang tepat digunakan untuk menaksir parameter regresi linier ketika data terkontaminasi oleh pencilan.

Kata Kunci
data pencilan, koefisien determinasi, metode TELBS

1. PENDAHULUAN

Analisis regresi linier dengan menggunakan Metode Kuadrat Terkecil (MKT) merupakan salah satu aplikasi statistika yang sangat terkenal karena mudah dipahami dan mudah dalam perhitungannya. Prinsip dasar dari MKT adalah meminimumkan jumlah kuadrat residual [4]. Sisi lain dari kemudahan dalam MKT adalah metode ini harus memenuhi asumsi Gaussian atau asumsi klasik. Jika asumsi Gaussian dipenuhi, maka MKT menjadi penaksir yang bersifat BLUE (*Best Linier Unbiased Estimation*) [7].

Data pencilan merupakan salah satu permasalahan yang muncul dalam analisis regresi. Pencilan dapat menyebabkan kenormalan pada asumsi Gaussian tidak terpenuhi. Keberadaan pencilan menyebabkan model tidak cocok untuk sebagian besar data, karena gangguan (*error*) yang akan diperoleh dari model akan menjadi besar. Dengan demikian jika digunakan sebagai peramalan, penggunaan MKT akan memperoleh hasil yang kurang baik.

Untuk mengatasi hal tersebut, digunakan metode *robust* sebagai aternatif dari MKT. Prosedur *robust* ditujukan untuk mengakomodasi adanya keanehan data, sekaligus meniadakan identifikasi adanya data *outlier* dan juga bersifat otomatis dalam menanggulangi data *outlier* [2]. Sampai saat ini banyak penelitian yang mengkaji metode penaksiran parameter regresi yang *robust* terhadap pencilan. Masing-masing metode pada umumnya mempunyai kelebihan dan kekurangannya, sehingga, seiring dengan waktu, satu metode menjadi popular pada masanya dan kemudian dipatahkan oleh metode lain yang punya kelebihan. Beberapa metode *robust* terhadap pencilan diantaranya yaitu: penaksir M yang dikemukakan Huber pada tahun 1973, *Least Median Square* (LMS) dan *Least Trimmed Square* (LTS) yang dikemukakan oleh Rousseeuw pada tahun 1984 [6]. Birch memperkenalkan metode pembobotan *robust* yang merupakan generalisasi dari penaksir M [1]. Metode ini dinamakan penaksir *Bounded-Influence* (penaksir B-I). Selanjutnya regresi *Minimum Covariance Determinant* (MCD) dikemukakan oleh Rousseeuw pada tahun 2004 dan lebih lanjut oleh Hubert [6]. Metode terbaru adalah metode TELBS [10]. Metode ini sekaligus dapat menangani masalah pencilan baik dalam ruang X ataupun ruang Y. Penelitian ini akan memperlihatkan keunggulan metode TELBS dalam menaksir parameter regresi dengan menggunakan data simulasi. Selain itu, metode TELBS akan diaplikasikan pada data sekunder mengenai produktivitas fitoplankton dengan menyertakan nilai koefisien determinan untuk menunjukkan kecocokan model pada data.

2. ANALISIS REGRESI

Analisis regresi merupakan aplikasi statistika yang dapat membantu memodelkan hubungan antara variabel bebas (*X*) dengan varibel tak bebas (*Y*). Model yang terbentuk dapat berupa hubungan linier atau non linier. Pada paper ini yang akan dibahas adalah model yang berbentuk linier.

Model sampel regresi linier dapat dinyatakan dalam persamaan:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_j x_{ij} + e_i \qquad (1)$$

dimana $i = 1, 2, \ldots, n$ dan $j = 1, 2, \ldots, p$ dengan y_i dan x_{ij} berturut-turut merupakan pengamatan ke-*i* untuk variabel tak bebas dan variabel bebas ke-*j*. Metode yang umum digunakan untuk menaksir parameter β adalah Metode Kuadrat Terkecil (MKT). Metode MKT merupakan metode yang paling mudah dilakukan dengan asumsi-asumsi tertentu yang disebut asumsi Gaussian [7], yaitu :

1. Normalitas
   Asumsi ini menyatakan bahwa nilai harapan dari ε_i. dengan syarat diketahui X_i mempunyai nilai 0, yaitu : $E(\varepsilon_i|X_i) = 0$

2. Homoskedasitas
   Asumsi ini menyatakan bahwa ε_i dengan syarat diketahui X_i memiliki variansi yang konstan, yaitu : $\text{var}(\varepsilon | X_i) = \sigma^2$
3. Tidak ada multikolinearitas
   Asumsi ini menyatakan bahwa tidak terdapat korelasi atau hubungan antara variabel bebas yang satu dengan variabel bebas yang lainnya.
4. Tidak ada autokorelasi
   Asumsi ini menyatakan unsur *error* pada pengamatan yang satu tidak dipengaruhi oleh unsur *error* yang berhubungan dengan pengamatan lain, yaitu :
   $cov(\varepsilon_i, \varepsilon_j) = 0, \; i \neq j$

Jika semua asumsi Gaussian dipenuhi maka β adalah parameter taksiran yang memenuhi sifat linier, tidak berbias dan memiliki varians minimum.

Persamaan (1) jika dinyatakan dalam bentuk matriks, akan menghasilkan:

$$\mathbf{y} = \mathrm{X}b + \mathbf{e} \qquad (2)$$

Dengan menggunakan Metode Kuadrat Terkecil (MKT), komponen β ditaksir dengan prinsip meminimumkan jumlah kuadrat residual sehingga menghasilkan taksiran sebagai berikut:

$$b = (X^t X)^{-1} X^t \mathbf{y} \qquad (3)$$

2.1 Koefisien Determinasi

Nilai koefisien determinasi (R^2) mencerminkan seberapa besar variasi dari variabel tak bebas (*Y*) dapat dijelaskan oleh variabel bebas (*X*). Nilai R^2 bernilai antara 0 sampai 1, jika nilai R^2 mendekati 1 menunjukkan tingkat ketepatan model yang semakin baik dalam menerangkan variasi data [9]. Koefisien determinasi dinyatakan dalam rumus berikut:

$$R^2 = \frac{\text{Jumlah Kuadrat Regresi}}{\text{Jumlah Kuadrat Regresi}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (4)$$

2.2 Pencilan

Pencilan adalah pengamatan yang jauh dari pusat data yang mungkin berpengaruh besar terhadap koefisien regresi [9]. Keberadaan pencilan dapat menyebabkan *error* dan variansi data menjadi besar. Akibatnya interval taksiran parameternya menjadi besar.

Pencilan baru ditolak jika setelah ditelusuri ternyata mengakibatkan kesalahan-kesalahan pada ukuran atau analisis, ketidaktepatan pencatatan data, dan terjadi kerusakan alat pengukuran. Bila ternyata bukan akibat dari kesalahan-kesalahan semacam itu, penyelidikan yang seksama harus dilakukan [3]. Penyelidikan atau diagnosa dalam analisis regresi adalah salah satu cara untuk memantau masalah yang timbul baik yang berkaitan dengan data ataupun model.

Salah satu cara diagnosa yang berkaitan dengan data yaitu pencilan adalah dengan menggunakan matriks *hat* [7] yang didefinisikan sebagai berikut:

$$H = X(X^t X)^{-1} X^t \qquad (5)$$

Matriks *hat* pada dasarnya mentransformasikan vektor nilai pengamatan *y* ke vektor nilai *y* taksiran. Diagonal utama dari matriks *hat* yaitu h_{ii} akan menunjukkan nilai yang lebih besar dari 2p/*n* jika data merupakan pencilan.

2.3 Metode TELBS

Salah satu metode regresi untuk mengestimasi parameter regresi ketika terdapat pencilan adalah metode TELBS. Estimasi TELBS bekerja lebih baik jika dibandingkan dengan metode kuadrat terkecil, penaksir M dan MM [10]. Menurut Tabatabai et al. [10], regresi *robust* estimasi TELBS dilakukan dengan meminimumkan fungsi objektif :

$$\min_b \sum_{i=1}^{n} \frac{r(t_i)}{L_i} = \min_b \sum_{i=1}^{n} \frac{1 - \text{sech}(w.t_i)}{L_i} \qquad (6)$$

dimana:

$$t_i = \frac{\left(y_i - x_i^t \hat{\beta}\right)\left(1 - h_{ii}\right)}{\sigma} \qquad (7)$$

$$L_i = \sum_{j=1}^{k} \max\left(M_j, |x_{ij}|\right) \qquad (8)$$

$$M_j = \text{median}\left\{|x_{1j}|, |x_{2j}|, \ldots, |x_{nj}|\right\} \qquad (9)$$

Nilai estimator σ dapat diperoleh dari persamaan sebagai berikut:

$$\hat{\sigma} = 1.1926\text{x median}\left\{\left|e_i - \text{median}\left(e_i\right)\right|\right\} \qquad (10)$$

Pemilihan konstanta 1,1926 membuat σ merupakan suatu estimasi yang mendekati tak bias dari sampel yang terbatas [8]. ω adalah bilangan real positif yang disebut sebagai konstanta kesesuaian (*tunning constant*) yang bernilai 0.628.

Untuk meminimumkan persamaan (6), turunan dari ρ terhadap β_0 dan βj disamakan dengan nol, sehingga menghasilkan persamaan:

$$\sum_{i=1}^{n} \frac{\psi\left(t_i\right)}{L_i} \frac{\left(1 - h_{ii}\right)}{s} \frac{\partial(y_i - b_0 - x_{i1} b_1 - x_{i2} b_2 - \ldots - x_{ik} b_k)}{\partial b_j} = 0 \qquad (11)$$

dengan $\psi(x) = \frac{\partial r(x)}{\partial x} = w\text{Sech}(wx)\,\text{Tanh}(wx)$.

Didefinisikan fungsi pembobot w_{ii} [10] adalah:

$$w_{ii} = \frac{\psi\left(t_i\right)\left(1 - h_{ii}\right)}{s e_i L_i} \qquad (12)$$

maka persamaan (11) dapat ditulis menjadi:

$$\sum_{i=1}^{n} w_{ii} e_i \frac{\partial(y_i - b_0 - x_{i1} b_1 - x_{i2} b_2 - \ldots - x_{ik} b_k)}{\partial b_j} = 0 \qquad (13)$$

Persamaan (13) jika diturunkan terhadap β_0 dan βj akan menghasilan bentuk matriks sebagai berikut :

$$\left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)b = \mathbf{X}^t\mathbf{W}\mathbf{y}$$

sehingga diperoleh penaksir metode TELBS yaitu:

$$\hat{\beta}_{TELBS} = \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{y} \qquad (14)$$

dimana **W** adalah matriks bujursangkar dengan elemen diagonalnya adalah w_{ii} pada persamaan (12) dan entri matriks $w_{ij} = 0, i \neq j$.

Pada metode TELBS nilai taksiran tidak langsung diperoleh sekali, tetapi dengan melakukan iterasi pada matriks W. Iterasi berhenti jika kekonvergenan tercapai.

Koefisien determinasi pada analisis regresi *robust* dengan metode estimasi TELBS didefinisikan sebagai berikut [10]:

$$R^2 = 1 - \left(\frac{\underset{i:1\le i\le n}{Median}\left|e_i\right|}{\underset{i:1\le i\le n}{Median}\left|y_i - \underset{i:1\le i\le n}{Median}\,\hat{y}_i\right|} \right)^2 \qquad (15)$$

3. DATA

Data yang digunakan adalah data simulasi dan data sekunder. Data simulasi diperoleh dengan membangkitkan distribusi normal (0,1) sebanyak 30 dan membuat model regresi linier yang melibatkan variabel tak bebas dan satu variabel bebas dengan memasukkan beberapa nilai yang ekstrim sebagai data pencilan.

Data sekunder merupakan hasil penelitian mengenai produktifitas primer fitoplankton dengan faktor-faktor fisika-kimia perairan pada budidaya perikanan menggunakan jala terapung, dengan produktifitas primer sebagai variabel tak bebas (*Y*) dan tiga buah variabel bebas yaitu intensitas cahaya (X_1), PH (X_2), dan kerapatan fitoplankton (X_3) [5].

4. HASIL DAN PEMBAHASAN

4.1 Hasil Analisis Data Simulasi

Simulasi dilakukan untuk membuktikan ketepatan metode pada data yang sudah dipersiapkan sebelumnya dengan mengkondisikan adanya pencilan di beberapa titik. Simulasi dilakukan dengan membangkitkan data berdistribusi normal dengan rata-rata satu dan varians nol sebanyak 30 dengan menggunakan software minitab. Hasilnya kemudian dimasukkan ke dalam persamaan (1) dimana variabel bebas yang dilibatkan hanya satu dan ditentukan secara sembarang. Komponen β_0 dan β_1 ditentukan dengan nilai berturut-turut dua dan 0.5. Hasil data hasil simulasi disajikan pada Tabel 1 sebagai berikut:

Tabel 1. Data Simulasi

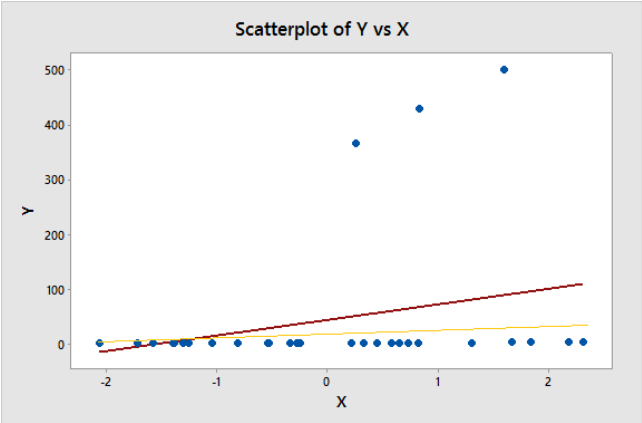
No.	X	Y	No.	X	Y
1	-2.05717249	0.971413755	16	0.829027346	2.414513673
2	1.304155024	2.652077512	17	0.335044018	2.167522009
3	0.732088968	2.366044484	18	-1.577341256	1.211329372
4	0.581938642	2.290969321	19	-1.035620495	1.482189752
5	0.255500679	**365.993**	20	0.649518012	2.324759006
6	-0.802987286	1.598506357	21	1.595841046	**500.34**
7	1.841209351	2.920604675	22	0.448437659	2.224218829
8	-0.338244042	1.830877979	23	2.315792729	3.157896364
9	2.182363536	3.091181768	24	-0.263915929	1.868042035
10	-0.274428675	1.862785663	25	-1.717873427	1.141063287
11	-1.299646311	1.350176844	26	0.219487165	2.109743582
12	-0.522699012	1.738650494	27	1.667805461	2.833902731
13	-0.538984787	1.730507607	28	-1.395247202	1.302376399
14	0.830870675	**428.6473**	29	-1.25639101	1.371804495
15	-1.386180966	1.306909517	30	-0.248020358	1.875989821

Data pada Tabel 1 diberikan pencilan pada data ke-5,14, dan 21, kemudian dianalisis dengan menggunakan MKT dan metode TELBS sebagai berikut:

Tabel 2. Hasil Data Simulasi dengan MKT dan Metode TELBS

Komponen	MKT	TELBS
β_0	43.9265	2
β_1	28.2803	0.5
R^2	0.0678	0.9997

Pada Tabel 2 dapat dilihat bahwa hasil taksiran dengan menggunakan metode TELBS memperlihatkan nilai koefisien determinasi yang menghasilkan nilai hampir sempurna. Untuk lebih menjelaskan bahwa metode TELBS menghasilkan model terbaik dapat dilihat pada Gambar 1. Pada Gambar 1 terlihat tiga titik yang berada di luar dari sebagian besar data yang ada. Garis berwarna merah adalah garis regresi dengan menggunakan MKT. Terlihat bahwa garis merah tertarik ke atas ke arah pencilan sehingga *error* yang dihasilkan menjadi besar. Garis berwarna kuning adalah garis regresi dengan menggunakan metode TELBS. Garis regresi dengan metode TELBS lebih mewakili sebaran data yang ada sehingga *error* yang dihasilan menjadi kecil. Hal ini sesuai dengan hasil koefisien determinasi yang lebih besar daripada MKT, yang menunjukkan bahwa model yang dihasilkan dengan metode TELBS lebih baik daripada MKT.



Gambar 1. Garis regresi untuk data simulasi

4.2 Hasil Analisis Data Sekunder

Data sekunder yang diperoleh mengenai produktivitas primer fitoplankton disajikan pada Tabel 3 berikut:

Tabel 3. Nilai Produktivitas Primer Fitoplanktoin dengan Faktor-faktor Fisika-Kimia Perairan

X1	X2	X3	Y
6485.2	7.8	148	0.2595
7030.5	7.38	194	0.7294
6551.3	7.48	180	1.5671
6140.3	7.52	134	0.5391
7243.2	7.59	102	0.7814
1245.2	7.8	110	0.2541
1391.3	7.39	226	0.2854
1296.3	7.45	209	0.7835
1656.8	7.53	126	0.2854
1956.4	7.56	94	0.4618
239.1	7.8	69	0.0577
275.3	7.39	189	0.1586
256.6	7.43	220	0.0326
447	7.36	61	0.1269
527.3	7.57	55	0.1412
49.9	7.8	103	0.0577
54.5	7.39	260	0.0951
50.8	7.98	110	0.1306
120.6	7.32	58	0.0000
142.3	7.57	41	0.0355

Dari data di atas kemudian didiagnosa dengan menentukan matriks *hat* (persamaan 5) untuk menentukan ada tidaknya pencilan. Nilai diagonal utama pada matriks *hat* disajikan pada Tabel 4. Terlihat bahwa data ke-17, 18, dan 19 adalah data pencilan yang ditunjukkan oleh nilai yang lebih besar dari $2p/n$=0.3. Jika dipaksakan dengan mengunakan MKT akan menghasilkan nilai koefisien determinasi sebesar 0.57762, artinya hanya 57.762% variabilitas produksi primer fitoplankton dapat dijelaskan oleh faktor-faktor yang mempengaruhinya.

Tabel 4. Nilai h_{ii}

Nilai h_{ii}	0.2889	0.2612	0.1960	0.1674	0.2663
	0.1463	0.1755	0.1336	0.0542	0.0727
	0.1736	0.1425	0.1874	0.2537	0.1433
	0.1689	**0.3106**	**0.3684**	**0.3083**	0.1815

Oleh karena itu, data kemudian dianalisis dengan menggunakan metode TELBS. Model yang diperoleh berdasarkan metode TELBS adalah sebagai berikut:

$$\hat{y}_i = -0.4259 + 1.75.10^{-4} x_{i1} + 0.0584 x_{i2} + 2.232.10^{-4} x_i \quad (16)$$

Koefisien determinasi dengan metode TELBS menghasikan nilai sebesar 0.8781. Artinya 87.81% variabilitas produksi primer fitoplankton dapat dijelaskan oleh faktor-faktor yang mempengaruhinya, sisanya sebesar 12.19% ditentukan oleh faktor-faktor lain, yang tidak dimasukkan dalam penelitian. Nilai ini lebih tinggi jika dibandingkan dengan koefisien determinasi dengan MKT.

4. KESIMPULAN

Metode TELBS dapat menghasilkan taksiran parameter regresi ketika terdapat data pencilan. Model yang diperoleh dengan menggunakan metode TELBS menjadi model yang mewakili sebagian besar data dengan ditandai oleh nilai koefisien determinasi yang tinggi.

5. SARAN

Paper ini hanya membahas pembentukan model regresi linier ketika pencilan terdeteksi, disarankan untuk penelitian selanjutnya dilakukan pengujian koefisien regresi.

UCAPAN TERIMA KASIH

Ucapan terima kasih kepada pihak Universitas Padjadjaran yang telah mendanai penelitian dalam skema Hibah Internal Universitas Padjadjaran.

DAFTAR PUSTAKA

[1] Birch, J.B., Estimation and Inference in Multiple Regression Using Robust Weight: A Unifield Approach, *Technical Report* 92-2, Departemen od Statistics Virginia Polytechnic Institute and State University, Blackburg Virginia, 1992.

[2] Cahyawati, D. Efektifitas Metode Regresi Robust Penduga Welsch dalam Mengatasi Pencilan pada Pemodelan Regresi Linear Berganda, *Jurnal Penelitian Sains*, Vol.12, no.1(A), Unsri, Sumatera Selatan, 2009.

[3] Chandraningtyas, S., dkk., Regresi Robust MM-*Estimator* untuk Penanganan Pencilan pada Regresi Linier Berganda. *Jurnal Gaussian*, Vol. 2, no.4, 395-404, Undip, Semarang, 2013.

[4] Gujarati, D.N., *Basic Econometrica*, 2 nd Edition, New York, McGraw-Hill Inc., 1988

[5] Gusriani, N., Firdaniza, Ardelina, D., Kajian Penaksir *Bounded-Influence* dan Metode *Minimum Covariance Determinant* pada Analisis Regresi Linier Berganda untuk Kasus Pencilan, Laporan Penelitian, Jurusan Matematika FMIPA Unpad, Bandung, 2011.

[6] Hubert, Mia et al., High-Breakdown Robust Multivariate Methods. *Statistical Science*, Vol. 23 No.1 (online: http://arxiv.org/pdf/0808.0657), 2008.

[7] Myers, R.H, *Classical and Modern Regression With Applications*. 2 nd edition. Boston, PSW-KENT Publishing Company, 1990.

[8] Rousseeuw, P.J and Croux, Alternative to the Median Absolute Deviation. *American Statistical Association*. 1993, Vol. 88, No. 424, 1993.

[9] Sembiring, R.K., *Analisis Regresi*, Edisi 2. Bandung: ITB, 2003

[10] Tabatabai, M.A. et al., TELBS Robust Linear Regression Method, Open Acces Medical Statistics, USA, Dove Medical Press, 2012.