J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 6, No.2, 92-106, Januari 2010
ISBN:1858-1382

Analisis Regresi Robust Menggunakan Kuadrat Terkecil Terpangkas untuk Pendugaan Parameter

Anisa, Raupong, Sarmiati Zainuddin *

Abstrak

Prosedur regresi robust ditujukan untuk mengakomodasi adanya keanehan data, sekaligus meniadakan identifikasi adanya data pencilan dan juga bersifat otomatis dalam menanggulangi data pencilan. Adanya pencilan pada suatu data biasanya menyebabkan ketidakakuratan pengambilan kesimpulan akhir. Untuk memperbaiki ketidakakuratan yang ada, penelitian ini akan mengkaji metode regresi robust untuk mengurangi pengaruh pencilan. Metode pendugaan parameter regresi robust yang digunakan adalah metode Kuadrat Terkecil Terpangkas. Metode kuadrat terkecil terpangkas dalam pendugaan parameternya menggunakan persamaan metode kuadrat terkecil biasa yang persamaannya dibentuk berdasarkan sub himpunan data sebanyak $\binom{n}{h}$ dan dipilih berdasarkan jumlah kuadrat sisaan terkecil.

Kata Kunci: *Analisis regresi, pencilan, robust, metode kuadrat terkecil, metode kuadrat terkecil terpangkas.*

1. Pendahuluan

Analisis regresi merupakan analisis yang mempelajari adanya keterkaitan antara satu variabel tak bebas (respon) dengan satu atau lebih variabel bebas, mempelajari bagaimana membangun sebuah model fungsional dari data untuk dapat menjelaskan ataupun meramalkan suatu fenomena alami atas dasar fenomena yang lain. Untuk itu dibutuhkan sekumpulan data prediktor untuk dapat menjelaskan data respon (Draper & Smith, 1992).

Hal pertama yang dilakukan dalam setiap analisis data adalah tahap persiapan data yang meliputi pengumpulan dan pemeriksaan data. Proses pengumpulan data dapat dilakukan dengan cara sensus atau sampling. Tahap selanjutnya adalah pemeriksaan data. Hal ini dilakukan untuk menghindari hal-hal yang tidak diinginkan, misalnya kekeliruan atau ketidakcocokan tentang data (Soemartini, 2007).

Pada data yang diperoleh bukan dari angket, tidak jarang ditemukan satu atau beberapa data yang jauh dari pola kumpulan data keseluruhan, yang lazim didefinisikan sebagai data pencilan (*outlier*), dimana suatu pengamatan terhadap suatu keadaan tidak menutup kemungkinan diperoleh suatu nilai pengamatan yang berbeda dengan nilai pengamatan lainnya. Hal ini mungkin disebabkan oleh kesalahan pada saat persiapan data atau terdapat peristiwa yang ekstrim yang mempengaruhi data (Soemartini, 2007). Pada regresi, pencilan adalah pengamatan dengan nilai sisaan yang besar, artinya pada pengamatan tersebut nilai variabel bebas tidak sesuai dengan nilai yang diberikan oleh variabel tak bebas (Sembiring, 1995).

* *Jurusan matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin, email: nkalondeng@yahoo.com*

Bila ternyata hasil identifikasi menunjukkan adanya pencilan, maka yang dapat dilakukan adalah identifikasi lanjut terhadap pencilan tersebut. Jika memberikan pengaruh setelah dilakukan pengujian, identifikasi lanjut bisa dilakukan dengan melihat hasil analisis jika pencilan tersebut dibuang/dihilangkan dari data atau tidak. Karena bagaimanapun juga keberadaan data pencilan mengganggu proses pengambilan kesimpulan.

Dalam penaksiran model regresi, baik pada analisis regresi linier sederhana maupun analisis regresi linier berganda, dilakukan metode penaksiran titik tertentu, diantaranya diperoleh dengan menggunakan metode Kuadrat Terkecil Biasa (*Ordinary Least Square*) dan Metode Kemungkinan Maksimum (*Maksimum Likelihood*).

Metode kuadrat terkecil biasa diketahui rentan terhadap pengaruh data pencilan/outliers (Rahmatul, 2006). Oleh karena itu diperlukan metode lain yang bersifat robust atau tahan terhadap pengaruh pencilan. Metode robust yang dimaksud antara lain Metode Kuadrat Terkecil Terboboti (*Weighted Least Square*/WLS), Metode Simpangan Mutlak Terkecil (*Least Absolute Value*/LAV), Metode Median Kuadrat Terkecil (*Least Median Square/*LMS), Metode Deviasi Mutlak Terkecil (*Least Absolute Deviations*/LAD), Penduga M, Penduga S, dan Metode Kuadrat Terkecil Terpangkas (*Least Trimmed Square*/LTS). Inti metode robust adalah memberikan bobot yang berbeda pada setiap pengamatan, meskipun metode ini memiliki kelemahan dalam teknik pemberian bobot pada setiap observasi. Pada metode kuadrat terkecil, setiap data diberi bobot sama yaitu 1, sedangkan pada metode robust, setiap data diberi bobot yang berbeda. Untuk pencilan, Rahmatul (2006) menyatakan untuk memberi bobot lebih kecil dari 1 bahkan 0 (terpangkas). Namun demikian, pada penelitian ini akan dikaji satu metode saja yaitu LTS, yang akan dibandingkan dengan metode klasik OLS.

Dalam penelitian ini digunakan data sekunder yang diperoleh dari skripsi tentang pengaruh berat tubuh mencit yang meminum obat, berat hati mencit yang meminum obat, dan dosis obat yang diminum terhadap konsentrasi obat dalam hati mencit, dengan judul ”Analisis Pencilan Pada Kecocokan Model Regresi” oleh Muhammad Hardoyo tahun 1997. Data ini telah diketahui mengandung pencilan.

Adapun asumsi awal dalam penelitian ini adalah metode regresi robust yang digunakan dapat mengurangi pengaruh suatu pencilan. Tujuan yang akan dicapai pada penelitian ini yaitu menduga parameter masing-masing metode regresi robust yang digunakan dan menentukan model terbaik bersesuaian dengan data yang digunakan dengan metode regresi robust, menentukan variabel bebas (prediktor) yang berpengaruh terhadap variabel respon untuk kedua metode robust yang digunakan.

2. Analisis Regresi

Model regresi yang mengandung satu variabel atau peubah bebas *X*, peubah respon *Y* dan fungsi regresinya linier, disebut model regresi linier sederhana. Pola hubungan antara *X* dan *Y* dikatakan linier bila besar perubahan nilai *Y* yang diakibatkan oleh *X* adalah konstan. Model tersebut dapat ditulis sebagai berikut :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

dengan :

Y_i adalah nilai peubah respon ke-*i*

β_0 dan β_1 adalah parameter

X_i adalah nilai peubah penjelas *X* pada amatan ke-*i* (konstanta yang diketahui)

ε_i adalah sisaan dari data ke-*i* yang bersifat acak (faktor acak)

Umumnya persoalan tentang regresi memerlukan lebih dari satu peubah bebas dalam modelnya. Misalkan bentuk data regresi berganda $\{Y_i, X_{i1}, X_{i2}, \cdots, X_{ik}\}$, *i*=1, 2, 3,…., *n*, dan $n \geq p$, Y_i menyatakan respon ke-*i* dari *k* peubah bebas $X_1, X_2, \cdots, X_k$ yang memenuhi persamaan

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \text{ , } i = 1,2,3,\ldots,n \qquad (2)$$

Persamaan (2) dapat dituliskan dalam bentuk sederhana menjadi :

$$Y_i = \beta_o + \sum_{i=1}^{n}\sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i, \qquad i = 1,2,3,\ldots,n \quad , \quad k = 1,2,\ldots, p \qquad (3)$$

Dalam bentuk matriks persamaan di atas dapat dituliskan menjadi :

$$Y = X\beta + \varepsilon \qquad (4)$$

dengan : Y = sebuah vektor pengamatan

β = vektor parameter-parameter

X = matriks konstan

ε = vektor acak berdistribusi normal sehingga saling bebas dengan $E(\varepsilon) = 0$, $\sigma^2(\varepsilon) = \sigma^2 I$

Asumsi sisaan yang harus dipenuhi oleh model regresi ialah $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, artinya sisaan ε_i berdistribusi normal independen dan identik, dengan mean dan variansi masing-masing bernilai *0* dan σ^2. Vektor acak *Y* memiliki $E(Y) = X\beta$ dan matriks $\sigma^2(Y) = \sigma^2 I$. Sebagai contoh, Sembiring (1995) mengambil *n* pengamatan, sehingga persamaan (4) dapat ditulis dalam bentuk matriks sebagai berikut :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ 1 & \vdots & \vdots & \cdots & \vdots \\ \vdots & & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (5)$$

3. Pencilan (*Outliers*)

Pencilan ialah data yang tidak mengikuti pola umum model, atau secara kasar sisaan atau errornya berjarak tiga kali simpangan baku atau lebih jauh lagi dari rata-rata sisaannya. Pencilan merupakan suatu keganjilan yang ada pada data dan menandakan suatu titik yang sama sekali tidak tipikal dibandingkan data lainnya. Berbagai kaidah telah diajukan untuk menolak pencilan, dengan kata lain mencoba menyisihkan amatan tersebut dari data, untuk kemudian menganalisis kembali tanpa amatan tersebut. Penolakan begitu saja suatu pencilan bukanlah prosedur yang bijaksana. Ada kalanya pencilan memberikan informasi yang tidak bisa diberikan oleh data lainnya, misalnya karena pencilan timbul dari kombinasi keadaan yang tidak biasa yang mungkin saja sangat penting dan perlu diselidiki lebih jauh (Draper & Smith, 1992).

Beberapa definisi pencilan menurut beberapa pakar :

1. Ferguson (1961), data yang menyimpang dari sekumpulan data yang lain.
2. Barnett (1981), pengamatan yang tidak mengikuti sebagian besar pola dan terletak jauh dari pusat data.
3. Weissberg (1985), jika terdapat masalah yang berkaitan dengan pencilan, maka diperlukan alat diagnosis yang dapat mengidentifikasi masalah pencilan, salah satunya dengan menyisihkan pencilan dari kelompok data kemudian menganalisis data tanpa pencilan (Soemartini, 2007).

Keberadaan data pencilan akan mengganggu dalam proses analisis data dan harus dihindari dalam banyak hal. Dalam kaitannya dengan analisis regresi, pencilan dapat menyebabkan hal-hal berikut :

1. Sisaan yang besar dari model yang terbentuk atau $E(\varepsilon) \neq 0$
2. Variansi pada data tersebut menjadi lebih besar
3. Taksiran interval parameternya memiliki rentang yang lebar

Kriteria pengambilan keputusan adanya pencilan dapat dilihat dengan menggunakan grafik, plot residual, dan beberapa nilai kriteria pencilan yang diberikan pada tabel berikut ini.

Tabel 1. Kriteria Pencilan

Leverage Values	>	$\frac{2p-1}{n}$
DfFITS	>	$2\sqrt{\frac{p}{n}}$
Cook's Distance	>	$F(0{,}5; p, n-p)$
DfBETA(s)	>	$\frac{2}{\sqrt{n}}$

Sumber : Soemartini (2007).

dengan : n = jumlah observasi (sampel)
p = jumlah parameter ($p = k+1$)
k = banyaknya variabel bebas

4. Metode Kuadrat Terkecil Terpangkas

Sebelum membicarakan mengenai Metode Kuadrat Terkecil Terpangkas, sebelumnya akan dibahas Metode Kuadrat Terkecil (Least Squares).

Metode Kuadrat Terkecil (*Least Square*)

Metode kuadrat terkecil merupakan teknik yang sangat popular digunakan untuk menduga parameter dan pencocokan suatu data. Draper dan Smith (1992) menyatakan metode kuadrat terkecil ini pertama kali ditemukan secara terpisah oleh Carl Friedrich Gauss asal Jerman (1777-1855), dan Adrien Marie Legendre asal Prancis (1752-1833).

Dewasa ini, metode kuadrat terkecil sangat luas digunakan untuk menemukan atau menduga nilai numerik parameter pencocokan sebuah fungsi data khusus untuk macam-macam penduga statistik, menganalisis data, dan mengambil kesimpulan yang bermakna tentang hubungan kebergantungan yang ada (Herve, 2003).

Draper & Smith (1992) menyatakan persamaan fungsi linier dengan satu variabel bebas dan variabel respon dapat dilihat pada persamaan (1) dan pendugaan parameter diberikan pada persamaan berikut :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (6)$$

Persamaan di atas terdiri dari dua parameter bebas, dimana $\hat{\beta}_0$ merupakan intersep/titik potong, dan $\hat{\beta}_1$ merupakan slope/kemiringan garis regresinya. Metode kuadrat terkecil menjelaskan bahwa pendugaan parameter persamaan ini sama dengan meminimumkan jumlah kuadrat diantara ukuran dan model prediksinya. Metode kuadrat terkecil diperlukan untuk membandingkan jumlah dari n kuadrat simpangan. Nilai minimum ini dijelaskan sebagai berikut :

$$L = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad (7)$$

dimana ε_i (error) merupakan jumlah nilai yang akan diminimumkan atau sisaan yang sifatnya acak dan merupakan penyimpangan model dari keadaan sesungguhnya.

Regresi Robust

Analisis regresi robust telah digunakan selama ratusan tahun, akan tetapi tidak serius ditangani akhir-akhir ini. Regresi robust merupakan metode yang digunakan menganalisis data yang mengandung pencilan. Metode tersebut dapat digunakan menciptakan suatu keadaaan yang stabil dalam membentuk model terbaik pada suatu kasus, dimana asumsi yang digunakan bahwa data yang ada tidak berdistribusi normal (Kutner, 2004).

Penyimpangan terhadap asumsi ideal vektor sisa ε yaitu vektor tersebut menyebar $N\left(0, I\sigma^2\right)$ sering terjadi. Bila penyimpangan nilai sisaannya terjadi cukup serius, perlu dilakukan penyesuaian seperlunya terhadap model. Untuk mengatasi penyimpangan-penyimpangan serta kemungkinan kekurangan yang lain, dapat menggunakan metode regresi robust sebagai pengganti prosedur metode kuadrat terkecil (Draper & Smith, 1992).

Beberapa teknik yang biasa digunakan dalam regresi robust yaitu Metode WLS, Metode LTS, Penduga S, Penduga–MM, Metode LAV, Metode LMS. Berikut ini akan dijelaskan metode regresi robust LTS yang akan dikaji dalam penelitian ini.

Metode Kuadrat Terkecil Terpangkas (*Least Trimmed Square*/LTS)

LTS merupakan suatu metode pendugaan parameter regresi *robust* yang tahan terhadap adanya pencilan dengan meminimumkan jumlah kuadrat sisaan sub himpunan data berukuran *h*. Adapun tujuan yang ingin dicapai adalah menduga nilai parameter model regresi yang robust terhadap adanya nilai pencilan (Fox, 2002).

Metode ini dikembangkan oleh Rousseeuw dan Leroy (1987). Ketika menggunakan alat–alat analisis, biasanya langkah pertama adalah mencoba menghilangkan pencilan kemudian mencocokkan data yang sudah bagus dengan menggunakan metode kuadrat terkecil, tetapi analisis robust mencocokkan model regresi dengan sebagian besar data dan kemudian mengatasi titik–titik pencilan yang memiliki nilai sisaan yang besar sebagai solusi robust tersebut (Soemartini, 2007). Jadi metode ini tidak membuang bagian dari data melainkan menemukan model fit dari mayoritas data.

Misalkan model regresi berganda pada persamaan (2), maka model taksirannya adalah :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ik} \qquad (8)$$

dan nilai sisaannya adalah :

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Adapun model pendugaan parameter metode LTS disajikan sebagai berikut :

$$\hat{\beta} = \min_{\beta}\left(\sum_{i=1}^{n}\varepsilon_i^2\right) = \min_{\beta}\left(\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\right) \qquad (9)$$

dengan :

$\varepsilon_{(i)}^2$: Kuadrat sisaan yang diurutkan dari terkecil ke terbesar $\varepsilon_{(1)}^2 \le \varepsilon_{(2)}^2 \le \cdots \le \varepsilon_{(n)}^2$

k : Banyaknya variabel bebas

p : Banyaknya parameter

Solusi dari $\hat{\beta}$ pada persamaan di atas dapat diperoleh dengan menggunakan turunan, seperti pada penyelesaian metode OLS, hanya pada LTS persamaan tersebut dihitung pada sub himpunan terbaik yang berukuran *h*. Banyaknya sub himpunan yang dibentuk sebanyak $\binom{n}{h}$ sub himpunan data, dimana *h* nilainya terletak antara $\left[\frac{n}{2}+1\right] \le h \le \left[\frac{3n+p+1}{4}\right]$ namun untuk mendapatkan nilai *h* dalam maksimum *breakdown* (proporsi minimal dari banyaknya pencilan dibandingkan seluruh data) mencapai 50% maka $h = \left[\frac{3n+p+1}{4}\right]$, *n* banyaknya pengamatan, dan ε sisaan. Sub himpunan *h* yang diperoleh merupakan sebaran data yang sudah terpangkas. Kemudian model dengan jumlah kuadrat sisaan yang terkecil dijadikan sebagai model fit (Soemartini, 2007; Notiragayu, 2008).

5. Pengujian Parsial Parameter Regresi

Untuk melihat apakah peubah *X* berpengaruh terhadap peubah *Y* dapat diuji dengan menggunakan uji *t*-student. Uji ini digunakan karena variansi (σ^2) populasi data tidak diketahui. Hipotesis dari pernyataan di atas adalah :

$H_0 : \hat{\beta} = 0$

$H_1 : \hat{\beta} \neq 0$

Statistik uji dari hipotesis di atas dituliskan sebagai berikut :

$$t - hitung = \frac{\hat{\beta}}{\sqrt{KTG \dfrac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}} \qquad (10)$$

Teknik penarikan kesimpulan dari uji *t*-student adalah jika *t*-hitung > $t\text{-tabel}_{\alpha/2;db=n-2,}$ maka H_0 ditolak atau tidak ada alasan yang cukup untuk menerima H_0. Kesimpulan yang diambil adalah $\hat{\beta} \neq 0$ atau $\hat{\beta}$ signifikan (Draper & Smith, 1992).

6. Data Penelitian

Adapun jenis data yang digunakan dalam penelitian ini adalah data sekunder tentang pengaruh berat tubuh mencit yang meminum obat, berat hati mencit yang meminum obat, dan dosis obat yang diminum terhadap konsentrasi obat dalam hati mencit. Data ini diperoleh dari skripsi dengan judul "Analisis Pencilan Pada Kecocokan Model Regresi" oleh Muhammad Hardoyo (NIM-89 03 020) tahun 1997. Indikator/parameter yang diamati dalam penelitian ini diberikan pada tabel berikut.

Tabel 2. Indikator/Parameter yang diukur.

No.	Variabel yang diamati	Satuan	Keterangan
1	Konsentrasi Obat dalam Hati (Y)	orang	Variabel Respon
2	Berat Tubuh (X_1)	mm^6	
3	Berat Hati atau Liver (X_2)	%	Variabel Bebas
4	Dosis Relatif Obat (X_3)	Tahun	

7. Hasil dan Pembahasan

Deskripsi variabel respon dan variabel prediktor/penjelas untuk data yang digunakan diberikan pada tabel berikut.

Dari tabel tersebut terlihat bahwa rata-rata konsentrasi obat dalam hati adalah 0,3353 mg. Variansi konsentrasi obat dalam hati adalah 0,0885. Konsentrasi obat dalam hati maksimum 0,56 mg dan minimum 0,21 mg. Distribusi jumlah pasien menurut berat tubuh, rata-rata berat tubuh adalah 171,53 kg dengan variansi 16,49. Berat tubuh maksimum yaitu 200 kg dan minimum 146 kg. Dari Tabel 4 ditunjukkan bahwa distribusi mencit menurut berat hati, rata-rata berat hati adalah 7,811 g dengan variansi 1,223. Berat hati maksimum 10 gram dan minimum 5,2 g. Distribusi mencit menurut dosis relatif obat, rata-rata adalah 0,8621 mg dengan variansi 0,0858. Dosis relatif obat maksimum 1 mg dan minimum 0,73 mg.

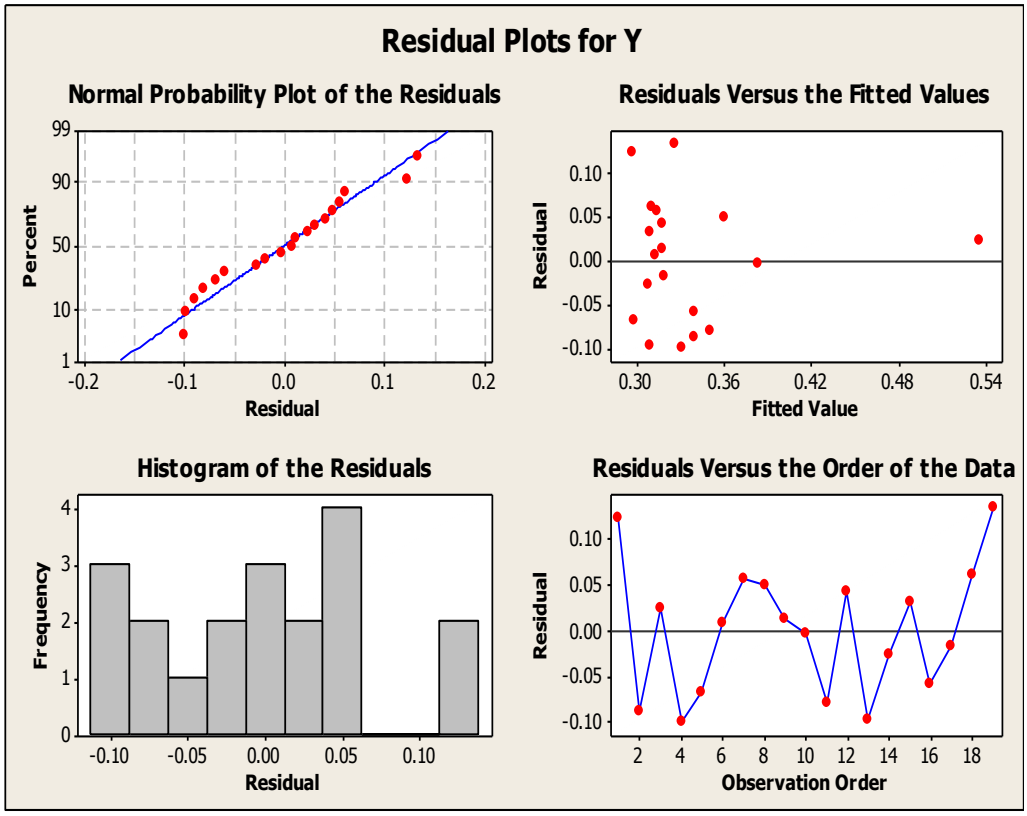
Tabel 3. Deskripsi Variabel untuk Data Konsentrasi Obat Dalam Hati

No.	Nama Variabel	Rata-Rata	Variansi	Minimum	Maksimum	Keterangan
1	Konsentrasi obat dalam hati (Y)	0,3353	0,0885	0,21	0,56	Variabel Respon
2	Berat tubuh (X_1)	171,53	16,49	146	200	Variabel Bebas
3	Berat hati/liver (X_2)	7,811	1,223	5,2	10	Variabel Bebas
4	Dosis Relatif Obat (X_3)	0,8621	0,0858	0,73	1	Variabel Bebas

Sumber : Data Analisis 2009

Analisis Data Pencilan

Untuk mendeteksi asumsi kenormalan data dapat diketahui melalui plot sisaan terhadap nilai amatan *Y* (*Residual Plots for Y*) yang diperlihatkan pada gambar berikut ini :



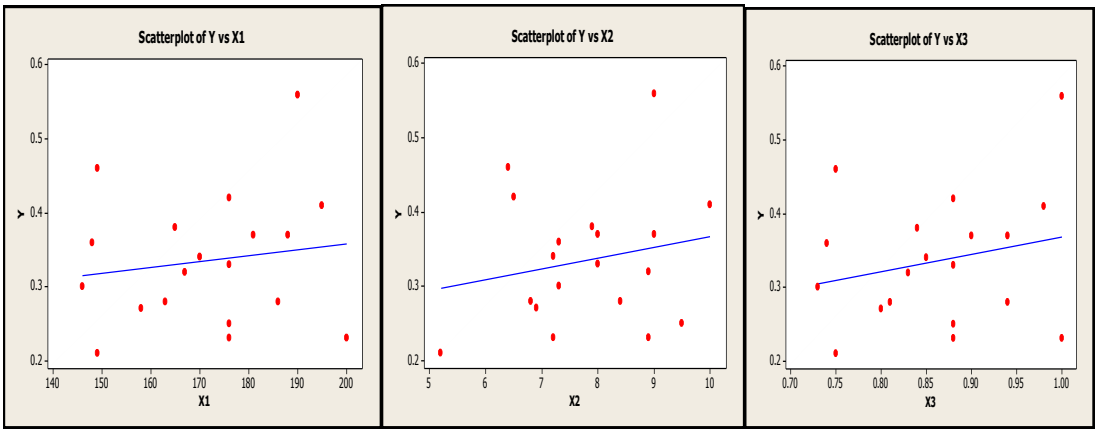
Gambar 1. Plot Residual/Error terhadap Y

Plot pertama terletak di kiri atas, merupakan plot peluang normal terhadap error (*Normal Probability Plot of the Errors*), mendeteksi kenormalan error. Nilai titik-titik error yang menempel atau sangat dekat dengan garis biru menunjukkan error tersebut berdistribusi normal, namun terdapat beberapa titik yang terletak agak jauh dari garis yang diketahui tidak berdistribusi normal, berarti tidak memenuhi asumsi $N(0,\sigma^2)$. Plot kedua pada bagian kiri bawah, merupakan histogram error yaitu plot histogram terhadap sisaan (*Histogram of The Residuals*). Secara visual histogram ini tidak menunjukkan error berdistribusi normal. Plot ketiga merupakan plot sisaan terhadap nilai pendugaan *Y (*Residuals *Versus the Fitted Values*), terletak di kanan atas. Titik-titik error tak ada yang bernilai di atas 2 atau di bawah -2, namun tampak tidak random. Kondisi ini tidak menggambarkan error bersifat *identik*. *Plot keempat* untuk membuktikan kondisi error yang saling bebas (*independent*). Plot ini merupakan plot sisaan terhadap urutan dari data (*Residuals Versus the Order of the Data*) dan terletak di kanan bawah. Titik-titik error tampak tidak acak, membentuk pola, ini berarti urutan pelaksanaan eksperimen atau urutan data ada hubungannya dengan nilai error. Ini berarti error dependen (tidak saling bebas).

Terdapat beberapa kriteria pengambilan keputusan adanya pencilan yaitu diberikan sebagai berikut :

1. Grafik

   Hubungan antara variabel bebas *X*1 (berat tubuh), *X*2 (berat hati/liver), dan variabel bebas *X*3 (dosis relatif obat) terhadap variabel respon *Y* (konsentrasi obat dalam hati) disajikan pada gambar berikut ini :



Gambar 2. Scatterplot *Y* terhadap *X*1, *X*2, dan *X*3.

Dari ketiga grafik model regresi linier di atas terlihat bahwa data yang ada tidak semuanya mengikuti pola umum dari data. Terdapat beberapa data yang jauh dari pusat kumpulan data yang ada. Data yang terletak jauh dari pusat data (garis regresi) tersebut yang dicurigai sebagai data pencilan.
Untuk data di atas diketahui :

$p = k+1 = 3+1=4, \qquad n = 19$

sehingga untuk mengetahui pencilan tersebut terletak pada data ke-*n*, dapat diketahui dari beberapa kriteria pencilan lain yang bisa digunakan berikut .

2. Leverage values > $\frac{2p}{n}$, dimana nilai Leverage values > 0,4210.
3. DFFITS > $2\sqrt{\frac{p}{n}}$, dimana DFFITS > 0,.9177.
4. Cook's Distance > $F(0{,}5; p, n-p)$, dimana Cook's Distance > 3,06.
5. DFBETAS > $\frac{2}{\sqrt{n}}$, dimana DFBETAS > 0,6324

Adapun nilai-nilai kriteria pencilan untuk masing-masing observasi data diberikan pada tabel berikut ini.

Data Berpengaruh

Nilai parameter-parameter yang diperoleh dengan menggunakan metode OLS berdasarkan kombinasi data yang dicurigai sebagai data pencilan disajikan sebagai berikut.

Tabel 4. Nilai Pendugaan Parameter Menggunakan Metode OLS

		All	**3**	**13**	**19**	**3+13**	**3+19**	**13+19**	**3+13+19**
OLS	β_0	0.266	0.311	0.409	0.116	0.468	0.164	0.240	0.300
	β_1	-0.021	-0.008	-0.022	-0.019	-0.007	-0.005	-0.021	0.004
	β_2	0.014	0.009	0.002	0.019	-0.005	0.013	0.009	0.002
	β_3	4.178	1.485	4.352	3.944	1.250	0.972	4.103	0.889
R^2		**0.364**	**0.021**	**0.390**	**0.457**	**0.050**	**0.071**	**0.442**	**0.007**

Dari hasil Tabel 4, diperlihatkan berapa besar pengaruh penghilangan observasi data yang dicurigai sebagai pencilan yaitu observasi ke-3, ke-13 dan ke-19 terhadap kecocokan model. Pertama-tama akan dicoba berapa besar kontribusi pengaruh observasi ke-3 terhadap perubahan koefisien determinasi R^2, dimana diketahui bahwa pada observasi penuh tanpa penghilangan R^2=0,364, sedangkan pada penghilangan observasi ke-3 diketahui bahwa R^2=0,021. Berarti pengaruh penghilangan observasi ke-3 tidak memberikan kontribusi yang baik terhadap kecocokan model. Selanjutnya penghilangan observasi ke-13, dimana R^2=0,39 memberikan kontribusi cukup baik dibandingkan dengan observasi keseluruhan yaitu R^2=0,364. Artinya kecocokan model ketika dilakukan penghilangan terhadap data ke-13 mengalami peningkatan (membaik). Sedangkan untuk penghilangan observasi ke-19 memberikan kontribusi yang lebih baik dengan R^2=0,457. Dari penghilangan ketiga observasi yang dicurigai sebagai pencilan, penghilangan terhadap data ke-19 yang memberikan kontribusi paling baik terhadap kecocokan model.

Penghilangan observasi gabungan yaitu observasi ke-3+ke-13 dan observasi ke-3+ke-19 diperoleh nilai R^2=0,05 dan R^2=0,071 yang membuat kecocokan model regresi makin buruk. Selanjutnya penghilangan observasi gabungan yaitu observasi ke-13+ke-19 diperoleh nilai R^2=0,442 yang membuat kecocokan model regresi makin membaik. Penghilangan observasi gabungan yaitu observasi ke-3+ke-13+ke-19 diperoleh nilai R^2=0,007 yang membuat kecocokan model regresi makin buruk. Dari keempat kombinasi penghilangan observasi, penghilangan observasi ke-13+ke-19 yang memberikan kontribusi paling baik.

Metode Regresi Robust dengan Metode Kuadrat Terkecil Terpangkas (*Least Trimmed Square*)

Cara mendapatkan model regresi dengan menggunakan metode *LTS* yaitu dengan menentukan sub himpunan terbaik dari data dengan menggunakan persamaan berdasarkan nilai $h = \left[\frac{3n + p + 1}{4} \right]$ dan membentuknya sebanyak $\binom{n}{h}$ sub himpunan data. Maka nilai *h* untuk keseluruhan data adalah $h = \frac{(3)(19) + 4 + 1}{4} = 16$ dengan sub himpunan data sebanyak

$\binom{19}{16} = \frac{19*18*17}{3*2*1} = 969$, nilai *h* untuk penghilangan data ke-13 dan ke-19 adalah

$h = \frac{(3)(18) + 4 + 1}{4} = 15$ dengan sub himpunan data sebanyak $\binom{19}{15} = \frac{19*18*17*16}{4*3*2*1} = 3876$, dan nilai *h* untuk penghilangan data ke-13+19 adalah

$h = \frac{(3)(17) + 4 + 1}{4} = 14$ dengan sub himpunan data sebanyak $\binom{19}{14} = \frac{19*18*17*16*15}{5*4*3*2*1} = 11628$.

Adapun nilai pendugaan parameter metode LTS dipilih berdasarkan jumlah kuadrat sisaan terkecil dari banyaknya sub himpunan data yang terbentuk dan kombinasi data berpengaruh yang telah diperoleh di atas, disajikan pada Tabel 5 berikut .

Tabel 5. Nilai Pendugaan Parameter Metode LTS

		All	**13**	**19**	**13+19**
LTS	β_0	0,045	0,068	0,086	0,068
	β_1	-0,020	-0,020	-0,021	-0,020
	β_2	0,049	0,047	0,051	0,047
	β_3	3,818	3,846	4,048	3,846
$\boldsymbol{R^2}$		**0,610**	**0,518**	**0,622**	**0,518**

Sumber : Data Analisis 2009

Dari Tabel 5 terlihat bahwa nilai R^2 untuk keseluruhan data sebesar 0,610. Terjadi peningkatan R^2 atau kecocokan model dengan data menjadi lebih baik saat penghilangan data ke-19 dengan R^2=0,622. Penghilangan data ke-13 dan data ke-13+ke-19 dengan R^2=0,518 mengakibatkan kecocokan model dengan data lebih buruk.

Selanjutnya, dilakukan pendugaan parameter untuk dua metode pendugaan, yaitu OLS dan LTS, dan melihat bagaimana kebaikan metode-metode tersebut untuk data yang bukan merupakan data pencilan. Secara acak, dilakukan pemodelan dengan kedua metode pendugaan jika amatan ke 1, 5, 10, dan 15 dihilangkan dari data. Hasil lengkapnya diberikan pada tabel berikut.

Tabel 6 memperlihatkan berapa besar pengaruh penghilangan amatan/observasi data yang bukan pencilan, yaitu amatan ke-1, 5, 10, dan 15 dengan dua metode OLS dan LTS. Selanjutnya akan dilihat berapa besar perubahan nilai $\boldsymbol{R^2}$ dan $\boldsymbol{R^2}$***-Adj***, yang merupakan indikator kesesuaian antara data dengan model, dengan penghilangan amatan-amatan tersebut. Dari tabel terlihat bahwa penghilangan amatan ke-1 terjadi perubahan besar pada nilai $\boldsymbol{R^2}$ dan $\boldsymbol{R^2}$***-Adj***, dari 0,364 dan 0,237 untuk keseluruhan data menjadi 0,468 dan 0,353. Sedangkan untuk amatan yang lain tidak terjadi perubahan yang cukup signifikan. Sedangkan untuk metode LTS, terlihat bahwa penghilangan observasi ke-1 dan ke-15 mengakibatkan kecocokan antara data dengan model menjadi lebih baik dibandingkan jika semua amatan digunakan dalam model. Dari tabel di atas terlihat bahwa nilai $\boldsymbol{R^2}$ untuk keseluruhan data dengan metode LTS adalah 0,610, sedangkan penghilangan amatan ke-1 dan 15 menghasilkan nilai $\boldsymbol{R^2}$ masing-masing sebesar 0,622 dan 0,672. Pengaruh penghilangan amatan ke-5 dan 10 malah menurunkan nilai kesesuaian antara data dengan model yang digunakan. Dari hasil ini dapat dikatakan bahwa perlu penelitian yang lebih mendalam terhadap amatan ke-1 dan ke 15, meskipun dalam pemeriksaan awal amatan-amatan tersebut bukan merupakan pencilan.

Tabel 6. Nilai Pendugaan Parameter untuk Data Bukan Pencilan Menggunakan Metode OLS dan LTS

		All	**1**	**5**	**10**	**15**
OLS	β_0	0,266	0,273	0,165	0,267	0,262
	β_1	-0,021	-0,24	-0,018	-0,021	-0,022
	β_2	0,014	0,025	0,005	0,014	0,015
	β_3	4,178	4,52	3,764	4,193	4,238
$\boldsymbol{R^2}$		**0,364**	**0,468**	**0,365**	**0,354**	**0,371**

R^2-Adj		**0,237**	**0,353**	**0,228**	**0,216**	**0,237**
LTS	β_0	0,045	0,0859	-0,0167	0,0446	0,0338
	β_1	-0,0197	-0,0212	-0,0176	-0,0197	-0,0204
	β_2	0,049	0,0505	0,042	0,049	0,0512
	β_3	3,875	4,0478	3,5345	3,8183	3,9397
R^2		**0,610**	**0,622**	**0,590**	**0,542**	**0,672**

Sumber : Data Analisis 2010

Pengujian Parsial Parameter Regresi

Nilai *t*-hitung dari masing-masing parameter yang dihitung disajikan pada tabel di bawah ini.

Tabel 7. Nilai *t*-hitung untuk Pendugaan Parameter dengan Metode OLS dan LTS

			All		**13**		**19**		**13+19**	
NILAI *t*-hitung	**OLS**	β_0	1,37		1,99		0,62		1,12	
		β_1	-2,66		-2,94		-2,7		-2,86	
		β_2	0,83		0,1		1,21		0,53	
		β_3	2,74	*	3	*	2,87	*	3,01	*
	LTS	β_0	105,65		25,75		-72,96		31,92	
		β_1	-46,77		-7,54		18,01	**	-9,34	
		β_2	116,34	**	17,65	**	-42,89		21,88	**
		β_3	9063,56	**	1456,48	**	-3437,94		1805,37	**

Ket : * = signifikan (berpengaruh)
**= sangat signifikan

Hipotesis yang akan diuji adalah : $H_0 : \hat{\beta} = 0$ dan $H_1 : \hat{\beta} \neq 0$ dengan t_{tabel} (0,025;17) = 2,110, t_{tabel} (0,025;16) = 2,120, t_{tabel} (0,025;15) = 2,131. Selanjutnya dari Tabel 5, dapat dilihat bahwa berdasarkan hasil untuk penghitungan nilai *t*-hitung, variabel bebas yang berpengaruh terhadap variabel respon untuk metode OLS dan metode regesi robust dengan LTS yang digunakan disajikan sebagai berikut :

1. Metode OLS

- Keseluruhan data. Adalah variabel *X*3, karena *t*-hitung=2,74> t_{tabel} = 2,110. Artinya variabel *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).
- Penghilangan data ke-13. Adalah variabel *X*3, karena *t*-hitung=3,00> t_{tabel} = 2,110. Artinya variabel *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).
- Penghilangan data ke-19. Adalah variabel *X*3, karena *t*-hitung=2,87> t_{tabel} = 2,120. Artinya variabel *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).

- Penghilangan data ke-13+ke-19. Adalah variabel *X*3, karena *t*-hitung=3,01 > t_{tabel} = 2,131. Artinya variabel *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).

2. Metode LTS

- Keseluruhan data. Adalah variabel *X*2 dan *X*3, karena *t*-hitung = 116,34 dan *t*-hitung = 9063,56 > t_{tabel} = 2,110. Artinya variabel *X*2 (berat hati/liver) dan *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).

- Penghilangan data ke-13. Adalah variabel *X*2 dan *X*3, karena *t*-hitung=17,65 dan *t*-hitung=1456,48 > t_{tabel} = 2,110. Artinya variabel *X*2 (berat hati/liver) dan *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).

- Penghilangan data ke-19. Adalah variabel *X*1, karena *t*-hitung=18,01 > t_{tabel} = 2,120. Artinya variabel *X*1 (berat tubuh) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).

- Penghilangan data ke-13+ke-19. Adalah variabel *X*2 dan *X*3, karena *t*-hitung=21,88 dan *t*-hitung=1805,37 > t_{tabel} = 2,131. Artinya variabel *X*2 (berat hati/liver) dan *X*3 (dosis relatif obat) berpengaruh terhadap variabel respon *Y* (konsentrasi obat dalam hati).

3. Metode OLS dan LTS untuk Amatan yang Bukan Pencilan

- Untuk amatan yang bukan pencilan, perlu perhatian lebih mendalam terhadap amatan ke-1 dan ke 15, karena dengan metode OLS dan LTS yang dicobakan ternyata penghilangan amatan-amatan tersebut meningkatkan nilai kesesuaian antara data dengan model.

8. Kesimpulan dan Saran

Beberapa kesimpulan yang dapat ditarik dalam penelitian ini adalah dari nilai pendugaan parameter diperoleh model terbaik untuk metode OLS dan kedua metode regresi robust diberikan sebagai berikut :

a. Model terbaik untuk keseluruhan data

$\hat{Y} = 0{,}045 - 0{,}02X1 + 0{,}049X2 + 3{,}818X3$, R^2= 0,610 (LTS)

b. Model terbaik untuk penghilangan data ke-19

$\hat{Y} = 0{,}086 - 0{,}021X1 + 0{,}051X2 + 4{,}048X3$, R^2= 0,622 (LTS)

c. Model terbaik untuk penghilangan data ke13+19

$\hat{Y} = 0{,}068 - 0{,}02X1 + 0{,}047X2 + 3{,}846X3$, R^2= 0,518 (LTS)

Dari kedua metode regresi robust yang digunakan, diperoleh nilai R^2 yang lebih baik dibandingkan dengan menggunakan metode OLS. Dari segi kecocokan model secara keseluruhan berdasarkan nilai R^2, penggunaan metode LTS lebih baik terutama dari segi penghilangan pencilan. Perlu perhatian lebih seksama terhadap amatan ke-1 dan ke-15, yang bukan merupakan pencilan, karena penghilangan kedua amatan tersebut meningkatkan nilai kesesuaian antara data dengan model yang digunakan. Dengan menggunakan uji-*t*, variabel-variabel bebas yang berpengaruh terhadap variabel respon untuk kedua metode regresi robust adalah variabel X_2 (berat hati) dan X_3 (dosis relatif obat).

Adapun saran dari penulis supaya penelitian selanjutnya dapat mengkaji lebih dalam lagi berbagai metode regresi robust yang dapat digunakan untuk mengurangi pengaruh pencilan.

Daftar Pustaka

[1] Draper, N. dan Smith, H., 1992, *Analisis Regresi Terapan Edisi II*, Gramedia Pustaka Utama, Jakarta.

[2] Fox, J., 2002, *Robust Regression. Appendix to An R and S-PLUS. Companion to Applied Regression*. [Diakses 28 Februari].

[3] Hardoyo, M., 1997, *Analisis Pencilan Pada Kecocokan Model Regresi*, UNHAS, Ujung Pandang.

[4] Herve, A., 2003, *Least Square*, The University of Texas, Dallas.

[5] Kutner, M.H, dkk., 2004, *Applied Linier Regression Models Edisi IV*, McGraw-Hill Education, Singapore.

[6] Myers, R. H., 1989, *Classical and Modern Regression with Applications*, PWS-KENT, USA.

[7] Notiragayu, 2008, *Perbandingan Beberapa Metode Analisis Regresi Komponen Utama Robust*, FMIPA, Lampung.

[8] Nur, B., 2008, *Penduga Parameter Terbaik*, [Diakses 16 Februari 2009]

[9] Sembiring, R.K., 1995, *.Analisis Regresi*, ITB Press, Bandung.

[10] Soemartini, 2007, Pencilan (outlier*), Jurnal Penelitian Universitas Padjajaran*, Jatinangor.

[11] GoogleNet, Tanpa Tahun, Weighted Least Square regression, Engineering Statistics Handbook. *http://*www.google.com.weightedleastsquare [Diakses 16 Februari 2009]

[12] *SAS Institute*, 2004, *SAS Institute Inc., Cary*.USA.