

# Pendeteksian Data Pencilan dan Pengamatan Berpengaruh pada Beberapa Kasus Data Menggunakan Metode Diagnostik

Sally Indra<sup>1</sup>, Dodi Vionanda<sup>2</sup>, Riry Sriningsih<sup>3</sup>

<sup>1</sup>Student of Mathematics Department State University of Padang, Indonesia

<sup>2,3</sup>Lecturers of Mathematics Department State University of Padang, Indonesia

<sup>1</sup>sallindra@gmail.com

<sup>2</sup>dodi\_matunp@yahoo.co.id

<sup>3</sup>srirysriningsih@yahoo.com

**Abstract** — Outlier is an observation which can disturb the process of data analysis. It also could be seen as unusual observation. It causes the violation of normality error assumption in regression analysis, so that it should be detected using diagnostic method. Generally, there are several cases of outlier such as upper outlier, lower outlier, outlier occur near the center of data and upper-lower outlier. Each of those cases gives different effect to linear regression model parameter estimates. Because the unusual observation may be detected as outlier, high leverage point or influential observation. Based on the result, outlier and influential observation give significant effect toward intercept, slope,  $R^2$  and  $s^2$ .

**Keywords** — Regression, Outliers, Influential Observation, Diagnostic

**Abstrak** — Pencilan merupakan suatu pengamatan yang keberadaannya dapat mengganggu proses analisis data. Pencilan menyebabkan asumsi kenormalan galat dalam analisis regresi tidak terpenuhi, sehingga perlu dilakukan pendeteksian keberadaan pencilan ini menggunakan metode diagnostik. Secara umum, terdapat beberapa kasus pencilan, di antaranya pencilan atas, pencilan bawah, pencilan yang mendekati pusat data serta pencilan atas-bawah. Masing-masing dari kasus pencilan tersebut memberikan pengaruh yang berbeda-beda terhadap pendugaan parameter dalam model regresi linier. Karena bisa jadi pencilan tersebut terdeteksi sebagai *outlier*, *high leverage point* maupun pengamatan berpengaruh. Berdasarkan hasil yang diperoleh, pencilan dan pengamatan berpengaruh memberikan pengaruh yang signifikan terhadap perubahan nilai *intercept*, *slope*,  $R^2$  maupun  $s^2$ .

**Kata Kunci** — Regresi, Pencilan, Pengamatan berpengaruh, Diagnostik

## PENDAHULUAN

Masalah yang sering muncul dalam analisis regresi adalah ditemukannya satu atau beberapa titik data berada jauh dari pola data pada umumnya atau yang biasa disebut sebagai pencilan. Pencilan biasanya timbul karena kesalahan pada sistem pengukuran, kesalahan dalam menginputkan data atau pun karena data tersebut memang merupakan suatu kejadian yang tidak biasa. Keberadaan pencilan dapat mengganggu proses analisis data terutama karena asumsi kenormalan galat tidak lagi terpenuhi, sehingga perlu dilakukan pendeteksian keberadaan pencilan ini menggunakan metode diagnostik karena bisa jadi pencilan tersebut terdeteksi sebagai *outlier*, *high leverage point* maupun pengamatan berpengaruh [6].

Secara umum ada beberapa kasus pencilan yang terjadi di dalam data. Beberapa kasus itu antara lain pencilan yang terjadi di sekitar pusat data, pencilan atas, pencilan bawah, serta pencilan atas-bawah [1]. Masing-masing dari kasus pencilan tersebut memberikan pengaruh yang berbeda-beda terhadap pendugaan parameter dalam model regresi linier. Seberapa besar pengaruh pencilan tersebut dapat diketahui dengan

melakukan eksplorasi untuk masing-masing kasus pencilan.

Diketahui model regresi linier adalah:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

dimana  $\mathbf{y}$  merupakan vektor peubah respons berukuran  $n \times 1$ ,  $\mathbf{X}$  adalah matriks peubah bebas berukuran  $n \times p$  untuk  $n$  menyatakan jumlah pengamatan dan  $p$  menyatakan jumlah parameter,  $\boldsymbol{\beta}$  adalah vektor parameter berukuran  $p \times 1$  dan  $\boldsymbol{\varepsilon}$  adalah vektor galat berukuran  $n \times 1$  dengan rata-rata nol dan ragam  $\sigma^2$ .

Dengan menggunakan metode kuadrat terkecil diperoleh dugaan parameter untuk  $\boldsymbol{\beta}$  adalah:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (2)$$

Misalkan penduga dari  $\mathbf{y}$  adalah  $\hat{\mathbf{y}}$ , maka:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \quad (3)$$

Jika persamaan (2) disubstitusikan ke persamaan (3), maka:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ \hat{\mathbf{y}} &= \mathbf{H}\mathbf{y} \end{aligned} \quad (4)$$

dimana  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ .  $\mathbf{H}$  disebut sebagai matriks topi yang mempunyai ukuran  $n \times n$ .  $\mathbf{H}$  digunakan untuk mendeteksi keberadaan *high leverage point* dimana suatu pengamatan ke- $i$  dapat dicurigai sebagai *high leverage point* apabila  $h_{ii} > 2p/n$  untuk  $p$  menyatakan jumlah parameter dan  $n$  menyatakan jumlah pengamatan [3]. Besarnya nilai  $h_{ii}$  adalah:

$$h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad (5)$$

untuk  $\mathbf{x}'_i$  menyatakan setiap baris ke- $i$  dari  $\mathbf{X}$ .

Sisaan pada model regresi linier didefinisikan sebagai:

$$e_i = y_i - \hat{y}_i \quad (6)$$

Sisaan yang digunakan untuk mendeteksi keberadaan outlier adalah *externally studentized residual* atau biasa juga disebut sebagai *R-student* yang didefinisikan sebagai:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} \quad (7)$$

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

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}}$$

untuk nilai  $s^2 = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-p}$ .

Suatu pengamatan dicurigai sebagai *outlier* apabila pengamatan tersebut memiliki nilai  $|t_i| > t_{\alpha/2, n-p-1}$  pada taraf nyata  $\alpha$  [2].

Pendeteksian pengamatan berpengaruh ditentukan oleh ukuran nilai *DFFITs*, *DFBETAS*, *Cook's Distance* dan *Covratio*. *DFFITs* digunakan untuk mengetahui pengaruh suatu pengamatan ke- $i$  terhadap model regresi yang ditinjau dari nilai fitnya. Besarnya nilai *DFFITs* adalah:

$$DFFITs_i = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}} \quad (8)$$

Suatu pengamatan ke- $i$  dikatakan berpengaruh terhadap nilai fitnya apabila pengamatan tersebut memiliki nilai  $|DFFITs_i| > 2\sqrt{p/n}$  [1].

*DFBETAS* digunakan untuk menyatakan pengaruh suatu pengamatan ke- $i$  terhadap koefisien ke- $j$ . Besarnya nilai *DFBETAS* adalah:

$$(DFBETAS)_{j,i} = \frac{r_{j,i}}{\sqrt{\mathbf{r}'_j \mathbf{r}_j}} \frac{e_i}{\sqrt{1-h_{ii}}} (R-student)_i \quad (9)$$

dimana  $\mathbf{r}'_j$  adalah baris ke- $j$  dari  $\mathbf{R}$  untuk  $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Suatu pengamatan ke- $i$  dikatakan berpengaruh terhadap koefisien ke- $j$  apabila pengamatan tersebut memiliki nilai  $|DFBETAS_i| > 2/\sqrt{n}$  [1].

*Cook's D* merupakan suatu ukuran pengaruh pengamatan ke- $i$  terhadap semua dugaan koefisien regresi. Pada *Cook's D*, pengaruh pengamatan ke- $i$  diukur oleh:

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{-i})'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \mathbf{b}_{-i})}{ps^2} \quad (10)$$

dimana  $\mathbf{b}$  adalah vektor koefisien penduga regresi dan  $\mathbf{b}_{-i}$  adalah vektor koefisien penduga regresi tanpa pengamatan ke- $i$  [3]. Suatu pengamatan ke- $i$  akan berpengaruh pada model regresi linier jika nilai  $D_i > F_{\alpha, p, n-p}$  dengan taraf nyata  $\alpha$ .

*COVRATIO* merupakan suatu ukuran yang menggambarkan pengaruh suatu pengamatan ke- $i$  terhadap ketelitian estimasinya [2]. Untuk menghitung nilai *COVRATIO* setiap pengamatan ke- $i$  ditentukan oleh:

$$COVRATIO_i = \frac{(s_{-i}^2)^p}{KTG^p} \left( \frac{1}{1-h_{ii}} \right) \quad (11)$$

dimana *KTG* adalah kuadrat tengah galat yang besar

nilainya adalah  $\sum_{i=1}^n e_i^2 / (n-p)$ .

Di antara ke empat ukuran tersebut tidak ada ukuran yang lebih peka dibandingkan yang lainnya dalam menentukan pengamatan berpengaruh. Semuanya tergantung pada sudut pandang dan tujuan dalam menentukan pengamatan yang berpengaruh [4].

## METODE

Data yang digunakan dalam penelitian ini merupakan data simulasi yang dibangkitkan untuk setiap kasus pencilan. Banyaknya pengamatan yang digunakan untuk setiap kasus adalah  $n = 30$  dan banyaknya parameter adalah  $p = 2$ .

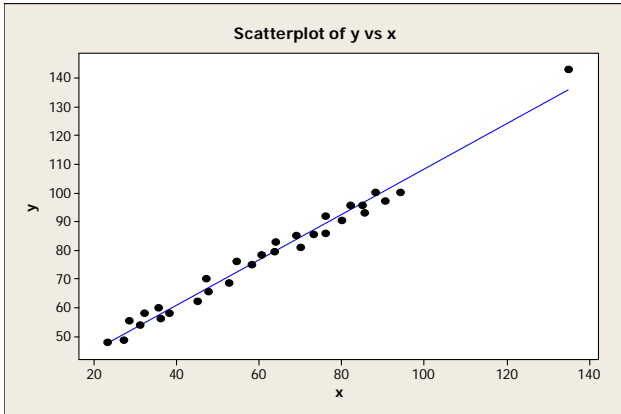
Langkah kerja yang dilakukan pertama kali adalah membangun model regresi linier untuk setiap kasus pencilan. Kedua, menggunakan metode diagnostik untuk mendeteksi keberadaan *outlier*, *high leverage point* dan pengamatan berpengaruh. Ketiga, membentuk model regresi dengan tidak mengikutsertakan pengamatan yang terdeteksi sebagai *outlier*, *high leverage point* dan pengamatan berpengaruh. Melihat perubahan yang terjadi antara sebelum dan setelah menggunakan metode diagnostik, kemudian menentukan pengaruh pencilan dan pengamatan tersebut terhadap pendugaan parameter dalam model regresi linier.

## HASIL DAN PEMBAHASAN

### A. Kasus Pencilan Atas

Pada kasus pencilan atas, pengamatan yang dibangkitkan sebagai pencilan merupakan pengamatan yang nilai peubah respons dan peubah bebasnya jauh lebih besar dibandingkan pengamatan yang lain. Pencilan berada di ujung mendekati garis regresi, jauh dari sebaran data lainnya serta berpengaruh terhadap pendugaan

parameter dalam model regresi linier. Plot data untuk kasus pencilan atas adalah sebagai berikut:



Gambar.1 Plot Data Untuk Kasus Pencilan Atas

Dugaan model awalnya adalah:

$$\hat{y}_i = 29,353 + 0,7899x$$

dengan  $R^2 = 98,1\%$  dan  $s^2 = 8,1950$ .

Selanjutnya melakukan pendeteksian pencilan pengamatan berpengaruh menggunakan metode diagnostik yang hasilnya dapat dilihat pada Tabel I.

Jika nilai  $|t_i| > t_{\alpha/2; n-p-1}$ , maka pengamatan ke- $i$  adalah pengamatan *outlier*. Dengan mengambil taraf nyata  $\alpha = 0,05$ , maka batasan nilai untuk  $t_{0,025; 2; 27}$  berdasarkan tabel- $t$  adalah 2,052. Pada kasus pencilan atas, pengamatan yang nilai  $|t_i|$  besar dari 2,052 adalah pengamatan ke-30 dengan  $|t_{30}| = 3,5171$ . Jika nilai  $h_{ii} > 2p/n = 0,1333$ , maka pengamatan ke- $i$  adalah pengamatan *high leverage point*. Pengamatan yang nilainya melebihi dari 0,1333 adalah pengamatan ke-30

dengan  $h_{30} = 0,3193$ . Selanjutnya, jika batasan nilai untuk  $|DFFITS_i| > 2\sqrt{p/n} = 0,5164$  maka pengamatan ke- $i$  merupakan pengamatan berpengaruh. Berdasarkan tabel pengamatan yang nilainya besar dari 0,5164 adalah pengamatan ke-30.

Dengan demikian, pengamatan ke-30 merupakan pengamatan yang terdeteksi sebagai *outlier, high leverage point* sekaligus pengamatan berpengaruh. Untuk melihat seberapa besar pengaruh yang diberikan oleh pengamatan ke-30 terhadap model regresi dapat diketahui dengan cara menghilangkannya dari data. Adapun hasil yang diperoleh antara sebelum dan setelah pengamatan ke-30 dihilangkan dapat dilihat pada Tabel II.

Dugaan model untuk data tanpa mengikutsertakan pengamatan ke-30 adalah:

$$\hat{y}_i = 31,479 + 0,7499x$$

Dikeluarkannya pengamatan ke-30 dari data memberikan pengaruh terhadap  $b_1$  dan  $s^2$  karena perubahan yang terjadi pada keduanya lebih besar dibandingkan pada  $b_0$  dan  $R^2$ .  $R^2$  pada model regresi data lengkap adalah 98,1% , setelah pengamatan ke-30 dihilangkan nilai ini menurun menjadi 97,9%. Sedangkan  $s^2$  pada data lengkap adalah 8,195 dan setelah pengamatan ke-30 dikeluarkan menurun menjadi 5,8283. Penurunan yang lebih besar pada  $s^2$  dibandingkan pada  $R^2$  menunjukkan bahwa model regresi untuk data tanpa pengamatan ke-30, yaitu pengamatan yang terdeteksi sebagai *outlier, high leverage point* dan pengamatan berpengaruh lebih baik dari model regresi dengan data lengkap untuk kasus pencilan atas.

TABEL I  
HASIL PERHITUNGAN METODE DIAGNOSTIK UNTUK KASUS PENCILAN ATAS

i	x	y	$t_i$	$h_{ii}$	DFFITS	i	x	y	$t_i$	$h_{ii}$	DFFITS
1	23.0	48.0	0.1743	0.1123	0.0620	16	63.5	79.5	-0.0049	0.0335	-0.0009
2	27.0	48.5	-0.7965	0.0968	-0.2607	17	64.0	83.0	1.1026	0.0336	0.2057
3	28.5	55.5	1.3512	0.0914	0.4286	18	69.0	85.0	0.4000	0.0363	0.0776
4	31.0	54.0	0.0570	0.0830	0.0171	19	70.0	81.0	-1.3157	0.0371	-0.2582
5	32.0	58.0	1.2385	0.0798	0.3646	20	73.0	85.5	-0.5345	0.0402	-0.1095
6	35.5	60.0	0.9411	0.0694	0.2571	21	76.0	86.0	-1.2214	0.0444	-0.2632
7	36.0	56.0	-0.6411	0.0681	-0.1733	22	76.0	92.0	0.9310	0.0444	0.2006
8	38.0	58.0	-0.4878	0.0628	-0.1263	23	80.0	90.5	-0.7284	0.0513	-0.1694
9	45.0	62.0	-1.0398	0.0479	-0.2333	24	82.0	95.5	0.4864	0.0555	0.1179
10	47.0	70.0	1.2718	0.0446	0.2748	25	85.0	95.5	-0.3543	0.0624	-0.0914
11	47.5	65.5	-0.4845	0.0439	-0.1038	26	85.5	93.0	-1.4314	0.0637	-0.3733
12	52.5	68.5	-0.8231	0.0377	-0.1629	27	88.0	100.0	0.4041	0.0704	0.1112
13	54.5	76.0	1.2945	0.0360	0.2502	28	90.5	97.0	-1.4230	0.0777	-0.4130
14	58.0	75.0	-0.0591	0.0340	-0.0111	29	94.0	100.0	-1.3388	0.0891	-0.4187
15	60.5	78.5	0.4750	0.0334	0.0883	30	135.0	143.0	3.5171	0.3193	2.4090

TABEL II  
HASIL ANALISIS KETIKA PENGAMATAN KE-30 DIHILANGKAN PADA  
KASUS PENCILAN ATAS

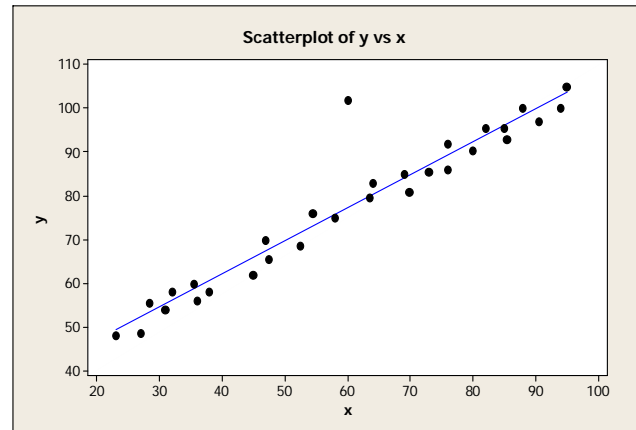
Pembeda	Sebelum Pengamatan ke-30 Dihilangkan	Setelah Pengamatan ke-30 Dihilangkan
<i>Outlier</i>	30	-
<i>High Leverage Point</i>	30	-
Pengamatan Berpengaruh	30	-
<i>Intercept</i> ( $b_0$ )	29,353	31,479
<i>Slope</i> ( $b_1$ )	0,78994	0,74985
$s^2$	8,1950	5,8283
$R^2$	98,1%	97,9%

### B. Kasus Pencilan yang Mendekati Pusat Data

Pada kasus pencilan yang mendekati pusat data, pengamatan yang dibangkitkan sebagai pencilan merupakan pengamatan yang nilai peubah bebasnya adalah nilai rataannya sendiri atau yang mendekati nilai rataannya. Sedangkan nilai peubah bebasnya jauh lebih besar dibandingkan pengamatan yang lain. Pengamatan seperti ini merupakan pencilan yang terdeteksi sebagai *outlier* namun bukan *high leverage point*. Gambar.2 menunjukkan plot data untuk kasus pencilan yang mendekati pusat data. Dugaan model awalnya adalah:

$$\hat{y}_i = 32,0 + 0,755x$$

dengan  $R^2 = 91,3\%$  dan  $s^2 = 28,1595$ .



Gambar.2 Plot Data Untuk Kasus Pencilan yang Mendekati Pusat Data

Hasil pendeteksian pencilan dan pengamatan berpengaruh menggunakan metode diagnostik untuk kasus pencilan yang mendekati pusat data dapat dilihat pada Tabel III. Karena jumlah data yang digunakan pada kasus pencilan yang mendekati pusat data ini sama dengan kasus pencilan atas, maka batasan nilai untuk  $|t_i|$ ,  $h_{ii}$  dan  $|DFFITS_i|$  yang digunakan juga sama. Pengamatan yang terdeteksi sebagai *outlier* adalah pengamatan ke-15 dengan  $|t_{15}| = 10,3223$  karena nilai  $|t_i|$  nya besar dari 2,052. Sedangkan pengamatan yang terdeteksi sebagai *high leverage point* tidak ada karena tidak ada nilai  $h_{ii}$  yang besar dari 0,1333. Pengamatan yang terdeteksi sebagai pengamatan berpengaruh adalah pengamatan ke-15 dengan  $|DFFITS_{15}| = 1,9169$  karena nilai  $|DFFITS_i|$  nya besar dari 0,5164. Dengan demikian,

TABEL III  
HASIL PERHITUNGAN METODE DIAGNOSTIK UNTUK KASUS PENCILAN ATAS

i	x	y	$t_i$	$h_{ii}$	DFFITS	i	x	y	$t_i$	$h_{ii}$	DFFITS
1	23.0	48.0	-0.2799	0.1288	-0.1077	16	63.5	79.5	-0.0909	0.0341	-0.0171
2	27.0	48.5	-0.7791	0.1094	-0.2731	17	64.0	83.0	0.4994	0.0343	0.0941
3	28.5	55.5	0.3796	0.1027	0.1284	18	69.0	85.0	0.1634	0.0386	0.0328
4	31.0	54.0	-0.2818	0.0922	-0.0898	19	70.0	81.0	-0.7420	0.0399	-0.1513
5	32.0	58.0	0.3489	0.0882	0.1085	20	73.0	85.5	-0.3137	0.0446	-0.0677
6	35.5	60.0	0.2223	0.0755	0.0635	21	76.0	86.0	-0.6543	0.0505	-0.1508
7	36.0	56.0	-0.6243	0.0738	-0.1762	22	76.0	92.0	0.4924	0.0505	0.1135
8	38.0	58.0	-0.5264	0.0674	-0.1415	23	80.0	90.5	-0.3709	0.0602	-0.0939
9	45.0	62.0	-0.7708	0.0493	-0.1756	24	82.0	95.5	0.2977	0.0660	0.0791
10	47.0	70.0	0.4703	0.0454	0.1026	25	85.0	95.5	-0.1372	0.0756	-0.0392
11	47.5	65.5	-0.4570	0.0445	-0.0986	26	85.5	93.0	-0.6978	0.0773	-0.2020
12	52.5	68.5	-0.6036	0.0375	-0.1191	27	88.0	100.0	0.2953	0.0864	0.0908
13	54.5	76.0	0.5325	0.0356	0.1023	28	90.5	97.0	-0.6602	0.0964	-0.2157
14	58.0	75.0	-0.1565	0.0337	-0.0292	29	94.0	100.0	-0.5942	0.1119	-0.2109
15	60.0	102.0	10.3223	0.0333	1.9169	30	95.0	105.0	0.2441	0.1166	0.0887

pengamatan ke-15 merupakan pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh. Untuk melihat seberapa besar pengaruh yang diberikan oleh pengamatan ke-15 terhadap model regresi dapat diketahui dengan cara menghilangkannya dari data.

Dugaan model untuk data tanpa mengikutsertakan pengamatan ke-15 adalah:

$$\hat{y}_i = 31,173 + 0,7553x$$

Adapun hasil yang diperoleh antara sebelum dan setelah pengamatan ke-15 dihilangkan dapat dilihat pada Tabel IV.

TABEL IV  
HASIL ANALISIS KETIKA PENGAMATAN KE-15 DIHILANGKAN PADA KASUS PENCILAN MENDEKATI PUSAT DATA

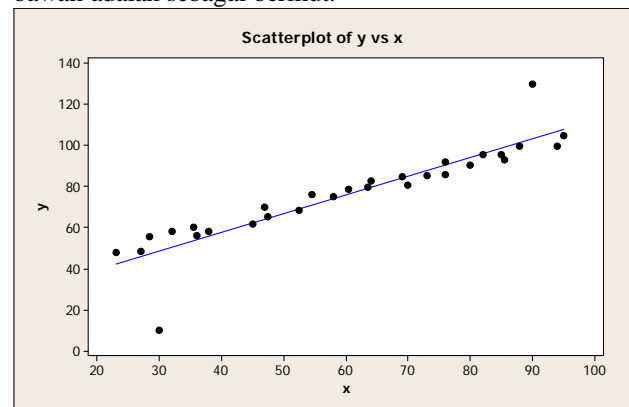
Pembeda	Sebelum Pengamatan ke-15 Dihilangkan	Setelah Pengamatan ke-15 Dihilangkan
<i>Outlier</i>	15	-
<i>High Leverage Point</i>	-	-
Pengamatan Berpengaruh	15	-
Intercept ( $b_0$ )	32,048	31,173
Slope ( $b_1$ )	0,75488	0,75529
$s^2$	28,1595	5,8283
$R^2$	91,3%	98,1%

Dikeluarkannya pengamatan ke-15 dari data memberikan pengaruh yang besar terhadap  $s^2$  dan  $R^2$ .  $R^2$  pada model regresi data lengkap adalah 91,3% , setelah pengamatan ke-15 dihilangkan nilai ini naik menjadi 98,1%. Sedangkan  $s^2$  pada data lengkap adalah 28,1595 dan setelah pengamatan ke-15 dikeluarkan menurun menjadi 5,8283. Penurunan yang besar pada  $s^2$  dan kenaikan pada  $R^2$  menunjukkan bahwa model regresi

untuk data tanpa pengamatan ke-15, yaitu pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh lebih baik dari model regresi dengan data lengkap untuk kasus pencilan yang mendekati pusat data.

### C. Kasus Pencilan Atas-Bawah

Pada kasus pencilan atas-bawah, pengamatan yang dibangkitkan sebagai pencilan merupakan pengamatan yang nilai peubah responsnya satu lebih besar dan satunya lagi lebih kecil dibandingkan pengamatan lainnya. Pencilan seperti ini merupakan pengamatan pencilan yang sepenuhnya mempengaruhi dugaan parameter dalam model regresi linier. Plot data untuk kasus pencilan atas-bawah adalah sebagai berikut:



Gambar.3 Plot Data Untuk Kasus Pencilan Atas-Bawah

Dugaan model awalnya adalah:

$$\hat{y}_i = 21,4 + 0,914x$$

dengan  $R^2 = 82,0\%$  dan  $s^2 = 95,0543$ .

Hasil pendeteksian pencilan dan pengamatan berpengaruh menggunakan metode diagnostik untuk kasus pencilan yang mendekati pusat data dapat dilihat pada Tabel V.

TABEL V  
HASIL PERHITUNGAN METODE DIAGNOSTIK UNTUK KASUS PENCILAN ATAS-BAWAH

i	x	y	$t_i$	$h_{ii}$	DFFITs	i	x	y	$t_i$	$h_{ii}$	DFFITs
1	23.0	48.0	0.6104	0.1285	0.2343	16	63.5	79.5	0.0121	0.0341	0.0023
2	27.0	48.5	0.2633	0.1091	0.0921	17	64.0	83.0	0.3247	0.0343	0.0612
3	28.5	55.5	0.8725	0.1024	0.2947	18	69.0	85.0	0.0610	0.0387	0.0122
4	30.0	10.0	-6.7075	0.0960	-2.1861	19	70.0	81.0	-0.4457	0.0399	-0.0909
5	32.0	58.0	0.7890	0.0880	0.2451	20	73.0	85.5	-0.2642	0.0446	-0.0571
6	35.5	60.0	0.6544	0.0753	0.1867	21	76.0	86.0	-0.4986	0.0505	-0.1150
7	36.0	56.0	0.1824	0.0736	0.0514	22	76.0	92.0	0.1239	0.0505	0.0286
8	38.0	58.0	0.1998	0.0672	0.0536	23	80.0	90.5	-0.4123	0.0603	-0.1044
9	45.0	62.0	-0.0496	0.0492	-0.0113	24	82.0	95.5	-0.0816	0.0660	-0.0217
10	47.0	70.0	0.5906	0.0453	0.1287	25	85.0	95.5	-0.3701	0.0756	-0.1058
11	47.5	65.5	0.0758	0.0444	0.0164	26	85.5	93.0	-0.6854	0.0773	-0.1984
12	52.5	68.5	-0.0854	0.0374	-0.0168	27	88.0	100.0	-0.1860	0.0865	-0.0572
13	54.5	76.0	0.4988	0.0356	0.0958	28	90.0	130.0	3.3156	0.0944	1.0703
14	58.0	75.0	0.0659	0.0337	0.0123	29	94.0	100.0	-0.7831	0.1119	-0.2779
15	60.5	78.5	0.1906	0.0333	0.0354	30	95.0	105.0	-0.3393	0.1166	-0.1233

TABEL VI  
HASIL ANALISIS KETIKA PENGAMATAN KE-4 DAN PENGAMATAN KE-28 DIHILANGKAN PADA KASUS PENCILAN ATAS-BAWAH

Pembeda	Sebelum Pengamatan ke-4 dan ke-28 Dihilangkan	Setelah Pengamatan ke- <i>i</i> Dihilangkan		
		4 saja	28 saja	4 dan 28
<i>Outlier</i>	4 dan 28	28	4	-
<i>High Leverage Point</i>	-	1	-	1
Pengamatan Berpengaruh	4 dan 28	28	4	-
<i>Intercept</i> ( $b_0$ )	21,369	28,159	23,993	31,060
<i>Slope</i> ( $b_1$ )	0,9136	0,8245	0,8539	0,7600
$s^2$	95,0543	36,9702	70,0532	5,9329
$R^2$	82,0%	90,3%	84,0%	97,9%

Pengamatan yang terdeteksi sebagai *outlier* adalah pengamatan ke-4 dan pengamatan ke-28 yang memiliki nilai  $|t_4| = 6,7075$  dan nilai  $|t_{28}| = 3,3156$  yang keduanya lebih besar dari 2,052. Sedangkan pengamatan yang terdeteksi sebagai *high leverage point* pada simulasi data ketiga ini tidak ada karena tidak ada nilai  $h_{ii}$  yang lebih besar dari 0,1333. Pengamatan yang terdeteksi sebagai pengamatan berpengaruh adalah pengamatan ke-4 dan pengamatan ke-28 dengan nilai  $|DFFITs_4| = 2,1861$  dan nilai  $|DFFITs_{28}| = 1,0703$  dimana keduanya lebih besar dari 0,5164. Dengan demikian, pengamatan ke-4 dan pengamatan ke-28 merupakan pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh, tetapi tidak terdeteksi sebagai *high leverage point*.

Untuk melihat seberapa besar pengaruh yang diberikan oleh pengamatan ke-4 dan pengamatan ke-28 terhadap model regresi dapat diketahui dengan cara menghilangkannya dari data. Adapun hasil yang diperoleh antara sebelum dan setelah pengamatan ke-4 dan pengamatan ke-28 dihilangkan dapat dilihat pada Tabel VI.

Dugaan model untuk data tanpa mengikutsertakan pengamatan ke-4 dan ke-28 adalah:

$$\hat{y}_i = 31,060 + 0,760x$$

Dikeluarkannya pengamatan ke-4 dan pengamatan ke-28 dari data memberikan pengaruh yang besar terhadap  $b_0$ ,  $s^2$  maupun  $R^2$ . Sebelum pengamatan ke-4 dan pengamatan ke-28 dihilangkan nilai  $s^2$  nya adalah 95,0543. Namun setelah dihilangkan nilai  $s^2$  nya berkurang menjadi 5,9329. Jika pada model dengan data lengkap  $R^2$  nya adalah 82%, namun pada model setelah pengamatan ke-4 dan ke-28 dihilangkan nilainya berkurang menjadi 97,9%. Adanya penurunan nilai yang lebih besar pada  $s^2$  dibandingkan pada  $R^2$  menunjukkan bahwa model regresi untuk data tanpa pengamatan ke-4 dan pengamatan ke-28, yaitu pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh lebih baik dari model regresi dengan data lengkap untuk kasus

pencilan atas-bawah.

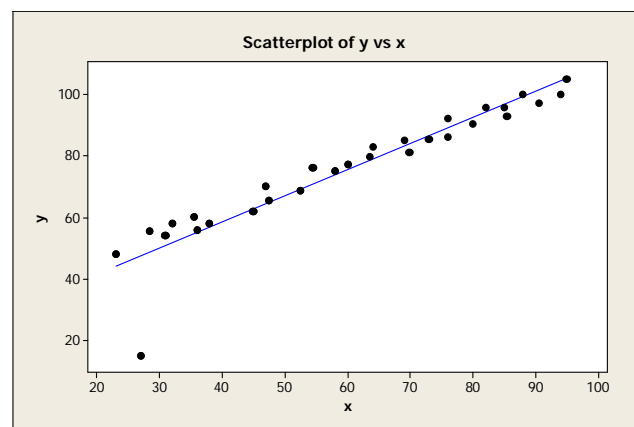
#### D. Kasus Pencilan Bawah

Pada kasus pencilan bawah, pengamatan yang dibangkitkan sebagai pencilan merupakan pengamatan yang nilai peubah respons jauh lebih kecil dibandingkan pengamatan lainnya. Pengamatan pencilan berada di ujung bawah garis regresi. Pengamatan seperti ini memuat informasi penting dalam membangun model regresi linier.

Plot data untuk kasus pencilan atas-bawah terdapat pada Gambar.4 dimana dugaan model awalnya adalah:

$$\hat{y}_i = 25,455 + 0,8320x$$

dengan  $R^2 = 82,0\%$  dan  $s^2 = 95,0543$ .



Gambar.4 Plot Data Untuk Kasus Pencilan Bawah

Selanjutnya melakukan pendeteksian pencilan pengamatan berpengaruh pada kasus pencilan bawah menggunakan metode diagnostik yang hasilnya dapat dilihat pada Tabel VII. Pengamatan yang terdeteksi sebagai *outlier* adalah pengamatan ke-2 dengan  $|t_2| = 14,8492$  karena nilai  $|t_i|$  nya besar dari 2,052. Sedangkan pengamatan yang terdeteksi sebagai *high leverage point* tidak ada karena tidak ada nilai  $h_{ii}$  yang lebih besar dari 0,1333. Pengamatan yang terdeteksi sebagai pengamatan berpengaruh adalah pengamatan ke-2 dengan  $|DFFITs_2| = 5,2044$  karena nilai  $|DFFITs_i|$  nya besar dari 0,5164.

TABEL VII  
HASIL PERHITUNGAN METODE DIAGNOSTIK UNTUK KASUS PENCILAN BAWAH

i	x	y	$t_i$	$h_{ii}$	DFFITs	i	x	y	$t_i$	$h_{ii}$	DFFITs
1	23.0	48.0	0.5162	0.1288	0.1985	16	63.5	79.5	0.1740	0.0341	0.0327
2	27.0	15.0	-14.8492	0.1094	-5.2044	17	64.0	83.0	0.6196	0.0343	0.1168
3	28.5	55.5	0.9560	0.1027	0.3234	18	69.0	85.0	0.3074	0.0386	0.0616
4	31.0	54.0	0.4077	0.0922	0.1299	19	70.0	81.0	-0.3874	0.0399	-0.0790
5	32.0	58.0	0.8846	0.0882	0.2752	20	73.0	85.5	-0.0990	0.0446	-0.0214
6	35.5	60.0	0.7401	0.0755	0.2114	21	76.0	86.0	-0.3884	0.0505	-0.0895
7	36.0	56.0	0.0869	0.0738	0.0245	22	76.0	92.0	0.4805	0.0505	0.1108
8	38.0	58.0	0.1355	0.0674	0.0364	23	80.0	90.5	-0.2195	0.0602	-0.0556
9	45.0	62.0	-0.1288	0.0493	-0.0293	24	82.0	95.5	0.2658	0.0660	0.0706
10	47.0	70.0	0.7924	0.0454	0.1728	25	85.0	95.5	-0.0982	0.0756	-0.0281
11	47.5	65.5	0.0758	0.0445	0.0164	26	85.5	93.0	-0.5278	0.0773	-0.1528
12	52.5	68.5	-0.0907	0.0375	-0.0179	27	88.0	100.0	0.1962	0.0864	0.0603
13	54.5	76.0	0.7528	0.0356	0.1446	28	90.5	97.0	-0.5575	0.0964	-0.1821
14	58.0	75.0	0.1848	0.0337	0.0345	29	94.0	100.0	-0.5490	0.1119	-0.1948
15	60.0	77.0	0.2330	0.0333	0.0433	30	95.0	105.0	0.0761	0.1166	0.0277

Dengan demikian, pengamatan ke-2 merupakan pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh. Untuk melihat seberapa besar pengaruh yang diberikan oleh pengamatan ke-2 terhadap model regresi dapat diketahui dengan cara menghilangkannya dari data.

Dugaan model untuk data tanpa mengikutsertakan pengamatan ke-2 adalah:

$$\hat{y}_i = 31,783 + 0,7474x$$

Adapun hasil yang diperoleh antara sebelum dan setelah pengamatan ke-4 dan pengamatan ke-28 dihilangkan dapat dilihat pada Tabel VIII.

TABEL VIII  
HASIL ANALISIS KETIKA PENGAMATAN KE-2 DIHILANGKAN PADA KASUS PENCILAN BAWAH

Pembeda	Sebelum Pengamatan ke-2 Dihilangkan	Setelah Pengamatan ke-2 Dihilangkan
<i>Outlier</i>	2	-
<i>High Leverage Point</i>	-	1
Pengamatan Berpengaruh	2	-
<i>Intercept</i> ( $b_0$ )	25,455	31,783
<i>Slope</i> ( $b_1$ )	0,83196	0,74736
$s^2$	48,7742	5,5179
$R^2$	88%	98%

$R^2$  pada model regresi data lengkap pada kasus pencilan bawah ini adalah 88%, setelah pengamatan ke-2 dihilangkan nilai ini naik menjadi 98%. Sedangkan  $s^2$  pada data lengkap adalah 48,7742, namun setelah pengamatan ke-2 dikeluarkan nilai  $s^2$  nya menurun

menjadi 5,5179. Penurunan yang besar pada  $s^2$  dan kenaikan pada  $R^2$  menunjukkan bahwa model regresi untuk data tanpa pengamatan ke-2, yaitu pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh lebih baik dari model regresi dengan data lengkap untuk kasus pencilan bawah.

#### SIMPULAN

Berdasarkan hasil eksplorasi dari beberapa kasus data menggunakan metode diagnostik pada bab sebelumnya, maka dapat diambil kesimpulan sebagai berikut:

1. Pada kasus pencilan atas, pengamatan yang terdeteksi sebagai *outlier*, *high leverage point* dan pengamatan berpengaruh memberikan pengaruh terhadap *intercept* sebesar 7,24% dan  $s^2$  sebesar 28,88%.
2. Pada kasus pencilan yang mendekati pusat, pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh memberikan pengaruh terhadap  $s^2$  sebesar 79,3% dan  $R^2$  sebesar 6,8%.
3. Pada kasus pencilan atas-bawah, pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh memberikan pengaruh terhadap *intercept* sebesar 45,35%, *slope* sebesar 16,81%,  $s^2$  sebesar 93,76% dan  $R^2$  sebesar 15,9%.
4. Pada kasus pencilan bawah, pengamatan yang terdeteksi sebagai *outlier* dan pengamatan berpengaruh memberikan pengaruh terhadap *intercept* sebesar 24,86%, *slope* sebesar 10,17%,  $s^2$  sebesar 88,69% dan  $R^2$  sebesar 10%.

## REFERENSI

- [1] Belsley, David A., Kuh, Edwin & Welsch, Roy E. 2004. *Regression Diagnostics*. New York: John Wiley & Sons.
- [2] Montgomery, C. Douglas, Elizabeth, A. Peck & Geofferey, Vining. 2006. *Introduction to Linear Regression Analysis Fourth Edition*. New York: John Wiley & Sons.
- [3] Myers, Raymond H. 1990. *Classical and Modern Regression with Applications*. Boston: PWS-KENT Publishing Company.
- [4] Anggeraini, Nova Diah. 2005. *Analisis Kepekaan Dalam Menentukan Pengamatan Berpengaruh Pada Analisis Regresi Melalui Simulasi Data*. Tugas Akhir. Padang: UNP.
- [5] Indra, Sally. 2013. *Pendeteksian Pencilan dan Pengamatan Berpengaruh pada Beberapa Kasus Data Menggunakan Metode Diagnostik*. Tugas Akhir. Padang: UNP.
- [6] Seber, George A. F. & Lee, Alan J. 2003. *Linear Regression Analysis*. New York: John Wiley & Sons.