II. TINJAUAN PUSTAKA

2.1 Definisi Pencilan

Dalam proses pengumpulan data, peneliti sering menemukan nilai pengamatan yang bervariasi (beragam). Keberagaman data ini, di satu sisi sangat dibutuhkan dalam analisa stastistika, namun di sisi yang lain keberagaman data menyebabkan adanya nilai pengamatan yang berbeda dengan nilai pengamatan lainnya. Dengan kata lain terdapat beberapa data yang berbeda dengan pola keseluruhan data. Penyebabnya mungkin terdapat kesalahan pada pengamatan, pencatatan, maupun kesalahan yang lain. Data yang berbeda ini disebut sebagai *outlier* atau data pencilan.

Pencilan adalah suatu pengamatan yang menyimpang cukup jauh dari pengamatan lainnya sehingga menimbulkan kecurigaan bahwa pengamatan tersebut berasal dari distribusi data yang berbeda. Distribusi pertama yaitu sebagai "distribusi dasar (*Basic distribution*)" yang menghasilkan pengamatan "baik". Distribusi kedua disebut "distribusi kontaminan (*Contaminating Distribution*)" yang menghasilkan pengamatan "tidak baik" (Sujatmiko, 2005).

Pencilan adalah data yang muncul memiliki karakteristik unik yang terlihat sangat jauh berbeda dari observasi-observasi lainnya dan muncul dalam bentuk nilai

ekstrim baik untuk sebuah variabel tunggal atau variabel kombinasi (Hair, dkk. dalam Soemartini, 2007).

Pencilan adalah data yang berprilaku menyimpang dari kelompok mayoritas datanya, atau bila digambarkan secara grafik data tersebut akan terletak di luar mayoritas datanya. Keberadaan data pencilan akan mengganggu dalam proses analisis data dan harus dihindari dalam banyak hal. Namun, membuang pencilan dalam suatu gugus data bukanlah prosedur yang tepat karena adakalanya pencilan memberikan informasi yang tidak bisa diberikan oleh data lain. Pencilan baru ditolak jika setelah ditelusuri ternyata akibat dari kesalahan-kesalahan, seperti kesalahan mencatat amatan yang bersangkutan atau kesalahan ketika menyiapkan peralatan (Draper dan Smith, 1992).

Pencilan adalah pengamatan yang jauh dari pusat data yang mungkin berpengaruh besar terhadap koefisien regresi (Sembiring dalam Soemartini, 2007).

2.2 Penyebab Munculnya Pencilan

Data pencilan mucul disebabkan karena berbagai kemungkinan antara lain:

1. Kesalahan prosedur dalam memasukan data atau mengkoding
2. Karena keadaan yang benar-benar khusus, seperti pandangan responden terhadap sesuatu yang menyimpang
3. Karena ada sesuatu alasan yang tidak diketahui penyebabnya oleh peneliti
4. Muncul dalam range nilai yang ada, tetapi bila dikombinasi dengan variabel lain menjadi ekstrim (disebut multivariat *outliers*).

2.3 Pendeteksian Pencilan

1. Pencilan dapat dilakukan dengan diagram kotak garis (box plot), bilamana terdapat data titik di luar batas pagar (dalam output *software* komputer) umumnya dilambangkan dengan * mengindikasikan terdapat data pencilan (*outlier*). Cara lainnya adalah dengan melihat mean dan standard deviationnya (untuk data interval dan ratio) yaitu bilamana standar devasi > mean berarti terdapat data pencilan (Solimun dalam Krisna, 2009).
2. Pengujian univariat *outlier* dapat dilakukan dengan menentukan nilai ambang batas yang akan dijadikan pencilan dengan cara mengkonversi nilai data penelitian ke dalam *standard score* atau Z-score (Ferdinand, 2002). Nilai terstandar memiliki rata-rata (mean) nol dengan standar deviasi (SD) sebesar satu. Batas nilai Z-score berada pada rentang 3-4 (Hair, dkk., 1998).
3. Pemeriksaan terhadap multi *outlier* dapat dilakukan dengan uji jarak Mahalanobis pada tingkat $p < 0{,}001$ (Solimun dalam Krisna, 2009). Jarak Mahalanobis dievaluasi dengan menggunakan χ^2 pada derajat kebebasan (df) sejumlah variabel yang digunakan dalam penelitian (Ferdinand dalam Krisna, 2009). Data tidak memiliki multi *outlier* apabila jarak Mahalanobis tidak lebih besar dari χ^2.

2.4 Dampak Pencilan

Pencilan berpengaruh terhadap proses analisa data, salah satunya terhadap nilai mean dan standar deviasi. Oleh karena itu, dalam suatu pola data keberadaan

pencilan harus dihindari. Dalam kaitannya dengan analisis regresi, pencilan dapat menyebabkan hal-hal berikut:

1. Residual yang besar dari model yang terbentuk atau $E[e] \neq 0$
2. Varians pada data tersebut menjadi lebih besar
3. Taksiran interval data dan range memiliki rentang yang lebar
4. Mean tidak dapat menunjukkan nilai yang sebenarnya (bias)
5. dan pada beberapa analisa inferensia pencilan dapat menyebabakan kesalahan dalam pengambilan keputusan dan kesimpulan.

2.5 Analisis Regresi

Analisis Regresi merupakan suatu studi mengenai hubungan antar variabel-variabel yang dipisahkan ke dalam dua jenis variabel, yaitu variabel bebas (X, *independent*) dan variabel tak bebas (Y, *dependent*). Regresi di samping digunakan untuk mengetahui bentuk hubungan antar peubah regresi, juga dapat digunakan untuk maksud-maksud peramalan. Biasanya hubungan antar variabel tersebut digambarkan dalam bentuk model matematis, seperti $Y = \beta_0 + \beta_1 X$; di mana Y adalah variabel tak bebas, dan X adalah variabel bebas. Aspek yang sangat penting dari analisis regresi adalah pengumpulan data karena kesimpulan dari analisis sangat tergantung pada data yang dikumpulkan. Pengumpulan data yang baik akan memberikan banyak manfaat, termasuk penyederhanaan analisis dan membangun model yang secara umum dapat dipergunakan dan dipertanggungjawabkan (Usman dan Warsono, 2001).

Dalam analisis regresi, terdapat dua model regresi, yaitu :

1. Model Regresi Linier Sederhana

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{..... (1)}$$

2. Model Regresi Linier Berganda

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ ... + \beta_k X_{ki} + \varepsilon_i$$

Di mana:

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ = Koefisien Regresi

X_i = Variabel bebas (*Regressor*)

Y_i = Variabel tak bebas (*Regressand*)

ε_i = Galat atau *error*

Jadi, dalam regresi linier sederhana ini yang akan diduga adalah $\beta_0$dan β_1.

Persamaan linier untuk pendugaan garis regresi linier ditulis dalam bentuk:

$\hat{y}_i = b_0 + b_1 x_i$

Dengan:

$\hat{y}_i$ = nilai dugaan variabel terikat pengamatan ke-i

x_i = nilai variabel bebas pengamatan ke-i

b_0 = titik potong garis regresi pada sumbu-y atau nilai dugaan $\hat{y}$ bila x=0

b_1 = gradien garis regresi (perubahan nilai dugaan $\hat{y}$ per satuan perubahan nilai x)

Model regresi linier sederhana dapat juga ditulis dalam bentuk matriks yaitu:

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ dengan $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2 I})$

Dengan:

$\boldsymbol{Y}$ = vektor nx1 variabel tak bebas

$\boldsymbol{X}$ = vektor nx2 variabel bebas

$\boldsymbol{\beta}$ = vektor 2x1 parameter

$\boldsymbol{\varepsilon}$ = vektor nx1 sisaan(galat) dengan $E(\varepsilon) = 0$ dan matriks ragam peragam $\sigma^2(\varepsilon) = \sigma^2 I$

Asumsi-asumsi pada analisis regresi adalah sebagai berikut :

1. Galat menyebar normal.

   $\varepsilon_i \sim N(0, \sigma^2)$

2. Ragam galat homogen.

   $Var(\varepsilon_i) = \sigma^2 \ ; i = 1, 2, \dots, n$

3. Nilai ε_i adalah bebas satu dengan yang lainnya.

   $E(\varepsilon_i) = 0$ dan $E = (\varepsilon_i^2) = \sigma^2$

4. Nilai tengah dari Y adalah fungsi linier dari X, yaitu jika dihubungkan titik-titik dari nilai tengah yang berbeda, maka akan diperoleh garis lurus.

2.6 Metode Kuadrat Terkecil (MKT)

Persamaan (1) merupakan model regresi linier sederhana dengan satu peubah bebas dan satu peubah respon dan untuk memperkirakan parameter-parameter β_0 dan β_1 dapat digunakan Metode Kuadrat Terkecil (*Least Square Method*) atau sering juga disebut dengan metode OLS (*Ordinari Least Square*) sedemikian rupa sehingga jumlah kuadrat kesalahan memiliki nilai terkecil.

Hines dan Montgomery (1990) menjelaskan bahwa jumlah kuadrat kesalahan pada pengamatan-pengamatan garis regresi sebenarnya adalah:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad \dots\dots\dots\dots (2)$$

Sehingga fungsi kuadrat terkecilnya adalah:

$$S = \sum_{i=1}^{n}\left[\left(Y_1 - \beta_0' - \beta_1(X_i - \bar{X})\right)\right]^2 \quad \ldots\ldots\ldots\ldots (3)$$

Estimator β_0 dan β_1 yang dinotasikan dengan $\hat{\beta}_0$ dan $\hat{\beta}_1$ harus memenuhi:

$$\frac{\partial S}{\partial \beta_0'} = -2\sum_{i=1}^{n}\left[Y_i - \hat{\beta}_0' - \hat{\beta}_1(X_i - \bar{X})\right] = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}\left[Y_i - \hat{\beta}_0' - \hat{\beta}_1(X_i - \bar{X})\right](X_i - \bar{X}) = 0$$

Penyelesaian untuk persamaan normal tersebut adalah:

$$\hat{\beta}_0' = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y} \quad \ldots\ldots\ldots\ldots (4)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \ldots\ldots\ldots\ldots (5)$$

$\hat{\beta}_0'$ dan $\hat{\beta}_1$ adalah estimator untuk intercept (titik potong) dan slope (kemiringan).

Estimator model regresi linier sederhana adalah:

$$\hat{Y} = \hat{\beta}_0' + \hat{\beta}_1(X - \bar{X}) \quad \ldots\ldots\ldots\ldots (6)$$

Untuk menyajikan hasil-hasil dalam susunan intercept yang asli β_0 maka $\hat{\beta}_0 = \hat{\beta}_0' - \hat{\beta}_1\bar{X}$ sehingga perkiraan yang cocok untuk model regresi adalah:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad \ldots\ldots\ldots\ldots (7)$$

2.7 *Mean Square Error* (MSE)

Jika $\hat{\beta}$ penduga tak bias dari β, maka $E\left(\left(\hat{\beta}\right) - \beta\right)^2$ sama dengan ragam penduga $\hat{\beta}$. Tetapi, jika suatu $\hat{\beta}$ penduga yang bias dari β, maka $E\left(\left(\hat{\beta}\right) - \beta\right)^2$ disebut *Mean Square Error* (MSE) atau kuadrat tengah galat dari penduga $\hat{\beta}$.

$$\text{MSE}\left(\hat{\beta}\right) = E\left(\left(\hat{\beta}\right) - \beta\right)^2$$

Bukti:

$$\text{MSE}\left(\hat{\beta}\right) = E\left(\hat{\beta}^2 - 2\hat{\beta}\beta + \beta^2\right)$$

$$= E\left(\hat{\beta}^2\right) - 2E\left(\hat{\beta}\right)\beta + \beta^2$$

$$= \left\{E\left(\hat{\beta}^2\right) - 2E\left(\hat{\beta}\right)\beta + \beta^2\right\} + \left\{\left(E\left(\hat{\beta}\right)\right)^2 - \left(E\left(\hat{\beta}\right)\right)^2\right\}$$

$$= \left\{E\left(\hat{\beta}^2\right) - \left(E\left(\hat{\beta}\right)\right)^2\right\} + \left\{\left(E\left(\hat{\beta}\right)\right)^2 - 2E\left(\hat{\beta}\right)\beta + \beta^2\right\}$$

$$= \left\{E\left(\hat{\beta}^2\right) - \left(E\left(\hat{\beta}\right)\right)^2\right\} + \left\{\left(E\left(\hat{\beta}\right) - \beta\right)^2\right\}$$

$$= var\left(\hat{\beta}\right) + \left(bias\left(\hat{\beta}\right)\right)^2$$

2.8 *Robust*

Regresi *Robust* diperkenalkan oleh Andrews (1972) dan merupakan metode regresi yang digunakan ketika asumsi normalitas homogenitas tidak terpenuhi dan atau adanya beberapa pencilan yang berpengaruh pada model. Metode ini merupakan alat penting untuk menganalisa data yang dipengaruhi oleh pencilan, sehingga dihasilkan model yang *Robust* atau *resistance* terhadap pencilan.

Prosedur regresi *Robust* dirancang untuk mengurangi pengaruh dari pengamatan-pengamatan yang mempunyai pengaruh tinggi jika metode kuadrat terkecil digunakan. Oleh karena itu, prosedur regresi cenderung untuk mengabaikan sisaan-sisaan yang berhubungan dengan pencilan-pencilan yang besar. Di samping tidak sensitif jika terdapat kasus pencilan, prosedur regresi *Robust* mempunyai tingkat efisien yang sama 90%-95% dibanding kuadrat terkecil jika di bawah distribusi normal (Montgomery & pek,1992). Beberapa metode penduga dalam regresi *Robust* diantaranya Penduga-M, *Least Trimmed Square* (LTS), Penduga-MM, Penduga-S, dan *Least Median of Square* (LMS).

2.9 Penduga-M

Penduga-M (*M-Estimator*) diperkenalkan oleh Huber pada tahun 1964. Penduga-M merupakan metode regresi *robust* yang sering digunakan. Penduga-M dipandang dengan baik untuk mengestimasi parameter yang disebabkan oleh x-*outlier* dan memiliki *breakdown point* 1/n. Penduga-M termasuk jenis penduga *Maximum Likelihood* (Hampel, 1986). Penduga-M menggunakan pendekatan yang sederhana antara komputasi dan teoritis. Prinsip dasar Penduga-M adalah meminimumkan fungsi objektif :

$$\sum_{i=1}^{n} \rho(e_i^*) = \sum_{i=1}^{n} \rho(e_i/\hat{\sigma})$$

$$= \sum_{i=1}^{n} \rho((y_i - x_i b)/\hat{\sigma}) \quad \ldots\ldots\ldots\ldots (1)$$

Dengan:

e_i^* = residual ke-i

$\rho(e_i)$ = fungsi simetris dari residual atau fungsi yang memberikan kontribusi pada masing-masing residual pada fungsi objektif.

$\hat{\sigma}$ = *scale*

Nilai $\hat{\sigma}$ diperoleh melalui iterasi (Chen, 2002):

$$\hat{\sigma}^{(l)} = med_{i=1}^{l} \left|y_i - x_i b^{(l-1)}\right|/\beta_0 \quad \ldots\ldots\ldots\ldots (2)$$

Dengan l (l=0, 1, …) adalah iterasi dan $\beta_0 = (0.6745)$

Dengan $\psi = \rho'$ adalah *derivative* dari ρ, maka untuk meminimumkan persamaan (1):

$$\sum_{i=1}^{n} \psi((y_i - x_i b)/\hat{\sigma}) x_i = 0 \quad \ldots\ldots\ldots\ldots (3)$$

$\psi(\cdot)$ merupakan fungsi *influence* yang digunakan dalam memperoleh bobot (*weight*). Dengan fungsi pembobot $w_i = \frac{\psi(e_i^*)}{e_i^*}$. Untuk penelitian ini digunakan fungsi pembobot *Tukey Biweight Function*, dengan bentuk fungsi sebagai berikut:

$w = (1 - (x/c)^2)^2$; $|x| < c$; $x = e_i/\hat{\sigma}$; $c = 4.685$

$= 0$; lainnya,

maka persamaan (3) menjadi:

$\sum_{i=1}^{n} w_i((y_i - x_i b)/\hat{\sigma})x_i = 0$ ..... (4)

Persamaan (4) dinotasikan ke dalam matriks:

$\boldsymbol{X^T W X b = X^T W y}$ ..... (5)

Persamaan (5) disebut *weighted least squares* yang meminimumkan $\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$. Regresi terboboti tersebut dapat digunakan sebagai alat untuk mendapatkan Penduga-M. Sehingga estimasi parameter menjadi:

$\boldsymbol{b = \left(X^T W X\right)^{-1} X^T W y}$ ..... (6)

Pembobot dalam Penduga-M bergantung pada residual dan koefisien. Prosedur untuk mendapatkan estimasi parameter yaitu iterasi yang disebut dengan *Iteratively Reweighted Least Squares* (IRLS), tahapannya yaitu:

1. Menaksir parameter regresi dan didapatkan residual $e_{i,0}$.
2. Menentukan $\hat{\sigma}^{(0)}$ dan fungsi pembobot $w_{i,0}$, $w_{i,0} = \psi(e_{i,0}^*)/(e_{i,0}^*)$
3. Mencari estimasi pada iterasi *l* (*l* = 1, 2, …) dengan *weighted least square*.
4. $\boldsymbol{b_1 = \left(X^T W_{l-1} X\right)^{-1} X^T W_{l-1} y}$, dengan w_{l-1} merupakan matriks diagonal dengan elemen diagonalnya adalah $w_{i,l-1}$.
5. Mengulang tahap 2 dan 3 hingga didapatkan penaksiran parameter yang konvergen.