BAB II

KAJIAN PUSTAKA

Beberapa teori yang diperlukan untuk mendukung pembahasan diantaranya adalah variabel random, regresi linear berganda, metode kuadrat terkecil (MKT), pengujian asumsi analisis regresi, pencilan *(outlier)*, regresi *robust*, koefisien determinasi, dan *breakdown point*.

A. Variabel Random

Pada sub bab ini akan dijelaskan mengenai variabel random. Tujuannya adalah untuk mengembangkan model matematika untuk menggambarkan peluang hasil dari peristiwa yang terjadi dalam ruang sample. Hasil- hasil percobaan dapat digambarkan dengan nilai-nilai numerik sederhana, dan variabel random dapat didefinisikan sebagai deskripsi numerik dari hasil percobaan.

Definisi 2.1. (Bain & Engelhardt, 1992:53) *Variabel random X merupakan fungsi yang memetakan setiap hasil yang mungkin e pada ruang sampel S dengan suatu bilangan riil x, sedemikian sehingga $X(e) = x$.*

Dengan simbol huruf besar X menotasikan suatu variabel random, sedangkan simbol huruf kecil x sebagai bilangan riil yang merupakan hasil nilai-nilai mungkin dari variabel random.

Contoh 2.1. Ruang sampel pelambungan 3 keping uang logam diperoleh $S = \{AAA, AAG, AGA, GAA, AGG, GAG, GGA, GGG\}$. Kemudian X merupakan banyaknya angka yang mungkin muncul dalam kejadian tersebut, sehingga variabel random X-nya adalah

$X(AAA) = 3$ $\quad$ $X(AAG) = 2$

$X(AGA) = 2$ $\quad$ $X(GAA) = 2$

$X(AGG) = 1$ $\quad$ $X(GAG) = 1$

$X(GGA) = 1$ $\quad$ $X(GGG) = 0$

Setiap variabel random memiliki peluang yang dapat diberikan oleh suatu fungsi yang dinamakan fungsi kepadatan peluang. Variabel random dikelompokkan menjadi dua jenis, yaitu variabel random diskrit dan variabel random kontinu. Berikut merupakan penjelasan singkat dari kedua jenis variabel random:

Variabel random diskrit

Variabel random diskrit adalah variabel random yang tidak mengambil seluruh nilai yang ada dalam interval atau variabel yang hanya memiliki nilai tertentu. Nilainya berupa bilangan bulat dan bilangan asli, tidak berbentuk pecahan.

Definisi 2.2. (Bain & Engelhardt ,1992:56) *Variabel random X disebut variabel random diskrit apabila himpunan semua nilai yang memenuhi variabel random X adalah himpunan terhitung $x_1, \ldots, x_n$ atau $x_1, x_2, \ldots$.*

Contoh 2.2. Sebuah koin dilempar sebanyak 20 kali, dan variabel random X merupakan banyaknya sisi gambar yang muncul. Dengan demikian, X hanya dapat bernilai dari $\{0, 1, 2, \ldots, 20\}$, X disebut variabel random diskrit.

Dalam variabel random diskrit terdapat fungsi kepadatan peluang diskrit dan fungsi distribusi kumulatifnya. Sedangkan pada variabel random kontinu mempunyai fungsi kepadatan peluang yang merupakan turunan dari fungsi distribusi kumulatifnya.

Dari pengertian variabel random diskrit, dapat didefinisikan fungsi kepadatatan peluang diskritnya, yaitu:

Definisi 2.3. (Bain & Engelhardt, 1992:56) *Fungsi* $f(x) = P(X = x)$, $x = x_1, x_2 \ldots$ *merupakan peluang untuk setiap nilai* x *yang mungkin disebut fungsi kepadatan peluang diskrit (*pdf*).*

Contoh 2.3. Jika sebuah dadu dilempar satu kali, maka terdapat 6 kemungkinan nilai yang akan terjadi. Peluang masing-masing kemungkinan tersebut adalah sama, dan dapat dituliskan sebagai berikut:

$$P(X = 1) = \frac{1}{6} \qquad P(X = 4) = \frac{1}{6}$$
$$P(X = 2) = \frac{1}{6} \qquad P(X = 5) = \frac{1}{6}$$
$$P(X = 3) = \frac{1}{6} \qquad P(X = 6) = \frac{1}{6}$$

Definisi 2.4. (Bain & Engelhardt, 1992:58) *Fungsi distribusi kumulatif (cumulative distribution function/cdf) dari variabel random* X *mendefinisikan untuk setiap bilangan real* x*, dengan* $F(x) = P(X \le x)$*.*

Hal itu berarti bahwa fungsi distribusi kumulatif adalah jumlahan nilai-nilai fungsi peluang untuk kemungkinan nilai *X* lebih kecil atau sama dengan *x*.

Fungsi $F(x)$ disebut fungsi distribusi kumulatif diskrit jika dan hanya jika memenuhi:

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i) \tag{2.1}$$

Fungsi tersebut mempunyai sifat-sifat:

(1) $0 \le F(x) \le 1$

(2) jika $x \le y$, maka $F(x) \le F(y)$ (2.2)

Contoh 2.4. Pelambungan sebuah mata dadu sebanyak satu kali, dengan ruang sample $S = \{1, 2, 3, 4, 5, 6\}$. Peluang munculnya mata dadu kurang dari atau sama dengan 3 dapat ditulis:

$$F(3) = P(X \leq 3) = \sum_{x_i \leq 3} f(x_i)$$

$$F(3) = P(X \leq 3) = f(1) + f(2) + f(3)$$

$$F(3) = P(X \leq 3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Jadi, peluang munculnya mata dadu kurang dari atau sama dengan 3 adalah $\frac{1}{2}$.

Variabel random kontinu

Pada umumnya, jika $F(X)$ merupakan cdf dari variabel random X, maka turunannya (asumsikan ada) dapat dinotasikan dengan $f(x)$ yang merupakan fungsi pdf.

Definisi 2.5. (Bain & Engelhardt, 1992:64) *Variabel random X disebut variabel random kontinu jika terdapat fungsi yang merupakan fungsi kepadatan peluang (pdf) dari X, sehingga fungsi distribusi kumulatifnya dapat ditunjukkan sebagai:*

$$F(x) = \int_{-\sim}^{x} f(t)\, dt$$

Contoh 2.5. Cdf dari $F(x) = P[X \leq x] = cx$. Misalkan $c = \frac{1}{5}$ dan $F(x) = \frac{x}{5}$ jika $0 \leq x \leq 5$. Pada umumnya, jika $F(x)$ adalah fungsi cdf variabel random X, maka dapat dinotasikan dengan turunannya yaitu $f(x)$, pada kondisi tertentu terdapat $f(x)$ sebagai pdf dari X. Sebagai contoh $F(X)$ dapat direpresentasikan untuk nilai dari x pada interval $[0,5]$ sebagai integral dari turunannya:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{0}^{x} \frac{1}{5}\, dt = \frac{x}{5}$$

B. Regresi Linier

Secara umum, analisis regresi pada dasarnya adalah studi mengenai variabel dependen (terikat) yang bergantung dengan satu atau lebih variabel independen (bebas), dengan tujuan untuk mengestimasi dan memprediksi rata-rata populasi atau nilai rata-rata variabel dependen berdasarkan nilai variabel independen yang diketahui.

Regresi linier merupakan suatu metode analisis statistik yang mempelajari pola hubungan antara dua variabel atau lebih menggunakan model persamaan linier, sehingga salah satu variabel pada model regresi dapat diduga dari variabel lainnya.

Regresi Linier Sederhana

Model regresi linier sederhana ini merupakan suatu model regresi dasar yang melibatkan satu variabel independen saja. Bentuk umum regresi linier sederhana dapat dituliskan sebagai berikut:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{2.5}$$

(Draper & Smith, 1998:22) dengan y_i merupakan variabel dependen pada observasi ke-i, x_i adalah konstanta yang diketahui yaitu nilai variabel independen yang diketahui, β_0 dan β_1 adalah parameter koefisien regresi, sedangkan e_i merupakan suatu *error*. Model tersebut dikatakan sederhana sebab hanya ada satu variabel independen, serta dapat dikatakan linier dalam parameter dan dalam variabel independen dikarenakan tidak ada parameter maupun variabel yang muncul sebagai suatu argumen dari fungsi transenden, hasil kali dan hasil bagi dengan parameter lain dan variabel ini berpangkat satu.

Regresi Linier Berganda

Pada suatu penelitian yang memanfaatkan analisis regresi, seringkali seorang peneliti ingin menyelidiki sejumlah variabel independen secara bersama yang berpengaruh terhadap variabel dependen. Hal ini dilakukan mengingat model dengan hanya satu variabel independen seringkali akan memberikan prediksi yang jauh lebih teliti. Suatu model yang lebih kompleks yang memuat beberapa variabel independen penting, biasanya lebih berguna karena dapat memberikan prediksi yang lebih teliti terhadap variabel dependen. Menurut Montgomery & Peck (1992:53), Model regresi linier berganda dengan k variabel independen adalah sebagai berikut:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i$$

atau dapat ditulis

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \quad i = 1,2,\dots,n \tag{2.6}$$

dengan:

y_i = nilai variabel dependen pada observasi ke-i

$\beta_0, \beta_1, \dots, \beta_k$ = parameter koefisien regresi

x_{ij} = nilai variabel independen yang ke-j pada observasi ke-i

e_i = *random error*

Parameter β_1 dan β_2 dalam model regresi linier berganda dikenal dengan nama koefisien regresi parsial, yang mempunyai makna sebagai berikut:

1. Parameter β_1 menunjukkan perubahan rata-rata variabel dependen untuk setiap kenaikan x_1 satu satuan bila x_2 dipertahankan konstan.

2. Parameter β_2 menunjukkan perubahan rata-rata variabel dependen untuk setiap kenaikan x_2 satu satuan bila x_1 dipertahankan konstan.

Bila pengaruh x_1 terhadap rata-rata variabel dependen tidak bergantung pada taraf x_2, dan sebagai akibat pengaruh x_2 terhadap rata-rata variabel dependen juga tidak bergantung pada taraf x_1, maka kedua variabel independen x_1 dan x_2 tidak bergantung atau saling mempengaruhi satu sama lain.

C. Metode Kuadrat Terkecil

Parameter $\beta_0, \beta_1, \dots, \beta_k$ tidak diketahui dan perlu ditentukan nilai estimasinya. Menurut Montgomery & Peck (1992:112), Metode Kuadrat Terkecil (MKT) digunakan untuk mengestimasi koefisien $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ yaitu dengan meminimumkan jumlah kuadrat galat. Fungsi yang meminimumkan adalah:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} {e_i}^2$$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \qquad (2.7)$$

Fungsi S akan diminimalkan dengan menentukan turunannya terhadap $\beta_0, \beta_1, \dots, \beta_k$, harus memenuhi

$$\left. \frac{\partial S}{\partial \beta} \right|_{\beta_0, \beta_1, \dots, \beta_k} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) = 0$$

Selanjutnya nilai $\beta_0, \beta_1, \dots, \beta_k$ diestimasi menjadi $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, sehingga menjadi

$$\left. \frac{\partial S}{\partial \beta} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \qquad (2.8)$$

dan

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \tag{2.9}$$

Selanjutnya dari Persamaan (2.8) dan Persamaan (2.9) di atas, menghasilkan persamaan normal kuadrat terkecil sebagai berikut:

$$\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{k} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{k} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^{\ 2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^{\ 2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i2} x_{ik} = \sum_{i=1}^{n} x_{i2} y_i \\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{ik} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} x_{ik} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^{\ 2} = \sum_{i=1}^{n} x_{ik} y_i
\end{aligned} \tag{2.10}$$

Diketahui bahwa ada $p = k + 1$ persamaan normal, satu untuk masing-masing koefisien regresi yang tidak diketahui. Solusi dari persamaan normal tersebut akan menjadi estimator kuadrat terkecil $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Akan lebih mudah apabila model regresi dinyatakan dalam matriks. Notasi matriks yang diberikan pada Persamaan (2.6) adalah

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

dengan $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$; $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$; $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$; $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$

Pada umumnya $\boldsymbol{Y}$ adalah matriks berukuran $(n \times 1)$, sedangkan $\boldsymbol{X}$ adalah matriks berukuran $(n \times k)$, $\boldsymbol{\beta}$ berukuran $(k \times 1)$, dan $\boldsymbol{e}$ adalah matriks berukuran $(n \times 1)$. *Error* dapat diturunkan dari persamaan di atas, sehingga diperoleh:

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X\beta}$$

Menurut Montgomery & Peck (1992:121), untuk menentukan estimator-estimator kuadrat terkecil, $\widehat{\boldsymbol{\beta}}$ yang meminimumkan $S(\beta_j)$ adalah:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{e_i}^2 = \boldsymbol{e^T e} \\ &= (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y^T Y} - \boldsymbol{Y^T X \beta} - \boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X \beta} \\ &= \boldsymbol{Y^T Y} - \boldsymbol{2\beta^T X^T Y} + \boldsymbol{\beta^T X^T X \beta} \qquad (2.11) \end{aligned}$$

Matriks $\boldsymbol{\beta^T X^T Y}$ adalah matriks berukuran (1×1), atau sebuah skalar, dan transpose $\boldsymbol{\beta^T X^T Y} = \boldsymbol{Y^T X \beta}$ yang merupakan skalar.

Kemudian akan ditentukan turunan parsial fungsi $S(\boldsymbol{\beta})$ terhadap $\boldsymbol{\beta}$ untuk menentukan estimator kuadrat terkecil,

$$\begin{aligned} \left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\beta} &= \frac{\partial(\boldsymbol{Y^T Y} - \boldsymbol{2\beta^T X^T Y} + \boldsymbol{\beta^T X^T X \beta})}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial(\boldsymbol{Y^T Y})}{\partial \boldsymbol{\beta}} - 2\frac{\partial(\boldsymbol{\beta^T X^T Y})}{\partial \boldsymbol{\beta}} + \frac{\partial(\boldsymbol{\beta^T X^T X \beta})}{\partial \boldsymbol{\beta}} \\ &= 0 - \boldsymbol{2X^T Y} + \boldsymbol{2X^T X \beta} \\ &= \boldsymbol{-2X^T Y} + \boldsymbol{2X^T X \beta} \end{aligned}$$

$$\begin{aligned} \left.\frac{\partial S}{\partial \widehat{\boldsymbol{\beta}}}\right|_{\widehat{\beta}} &= \frac{\partial\left(\boldsymbol{Y^T Y} - \boldsymbol{2\widehat{\beta}^T X^T Y} + \boldsymbol{\widehat{\beta}^T X^T X \widehat{\beta}}\right)}{\partial \widehat{\boldsymbol{\beta}}} \\ &= \boldsymbol{-2X^T Y} + \boldsymbol{2X^T X \widehat{\beta}}. \qquad (2.12) \end{aligned}$$

Agar diperoleh estimator-estimator kuadrat terkecil, maka harus meminimalkan turunan parsial fungsi $S(\boldsymbol{\beta})$ terhadap $\widehat{\boldsymbol{\beta}}$ dan memenuhi

$$\frac{\partial S}{\partial \widehat{\boldsymbol{\beta}}} = 0.$$

Dengan menyelesaikan Persamaan (2.12) di atas, akan diperoleh estimator untuk $\boldsymbol{\beta}$, yaitu:

$$\frac{\partial S}{\partial \widehat{\boldsymbol{\beta}}} = 0$$

$$-2\boldsymbol{X^T Y} + 2\boldsymbol{X^T X}\widehat{\boldsymbol{\beta}} = 0$$

$$2\boldsymbol{X^T X}\widehat{\boldsymbol{\beta}} = 2\boldsymbol{X^T Y}$$

$$\boldsymbol{X^T X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X^T Y}. \qquad (2.13)$$

Apabila kedua ruas dikalikan invers dari matriks $(\boldsymbol{X}^T\boldsymbol{X})$, maka estimasi kuadrat terkecil dari $\boldsymbol{\beta}$, yaitu

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \qquad (2.14)$$

Diasumsikan bahwa invers matriks $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ ada. Diperoleh matriks dari Persamaan normal (2.13) yang identik dengan bentuk skalar pada Persamaan (2.10). Dari Persamaan (2.13), diperoleh

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \dots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} {x_{i1}}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & & \sum_{i=1}^{n} x_{i1}\, x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & & \sum_{i=1}^{n} {x_{ik}}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

Matriks $\boldsymbol{X^T X}$ adalah matrik persegi berukuran $k \times k$ dan $\boldsymbol{X^T Y}$ adalah vektor $k \times 1$. Diagonal elemen matriks $\boldsymbol{X^T X}$ merupakan jumlah kuadrat dari kolom-kolom $\boldsymbol{X}$, dan elemen-elemen selain diagonalnya merupakan perkalian elemen dalam kolom $\boldsymbol{X}$. Sedangkan elemen-elemen matriks $\boldsymbol{X^TY}$ adalah jumlah perkalian antara kolom $\boldsymbol{X}$ dan observasi $\boldsymbol{y_i}$.

Model regresi dengan variabel independen $x^T = [1, x_1, x_2, \dots, x_k]$, diperoleh

$$\hat{y} = \boldsymbol{x^T\hat{\beta}} = [1, x_1, x_2, \dots, x_k] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

sehingga

$$\hat{y} = \boldsymbol{x^T\hat{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

dengan penjabaran $x^T = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = X$, maka dapat dituliskan

$$\boldsymbol{\hat{y}} = \boldsymbol{X\hat{\beta}} = \boldsymbol{X}(\boldsymbol{X^TX})^{-\boldsymbol{1}}\boldsymbol{X^TY} = \boldsymbol{HY}$$

dengan matriks persegi yang disebut matriks *hat*

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^TX})^{-\boldsymbol{1}}\boldsymbol{X^T} \tag{2.15}$$

D. Pengujian Asumsi Analisis Regresi

Pada model regresi, perlu dilakukan uji asumsi analisis regresi untuk mengetahui apakah model memenuhi asumsi atau tidak. Apabila ada uji asumsi yang tidak terpenuhi, dapat dipastikan data observasi mengandung *outlier*. Apabila semua uji asumsi terpenuhi, belum tentu data observasi tidak mengandurng *outlier*. Sehingga uji *outlier* harus dilakukan. Asumsi yang memenuhi analisis regresi dengan MKT antara lain: residu berdistribusi normal, homoskedastisitas, non autokorelasi, dan non multikolinieritas.

Uji Normalitas

Analisis regresi linier mengasumsikan bahwa sisaan (e_i) berdistribusi normal. Pada regresi linier klasik diasumsikan bahwa tiap sisaan (e_i) berdistribusi normal dengan $e_i \sim N(0, \sigma^2)$ (Gujarati, 2004:109).

Salah satu cara untuk menguji asumsi kenormalan adalah dengan uji *Kolmogorov-Smirnov*. Uji ini didasarkan pada nilai D, yaitu:

$$D = \max|F_0(X_i) - S_n(X_i)|, i = 1,2, \ldots, n.$$

dengan $F_0(X_i)$ adalah fungsi distribusi frekuensi kumulatif relatif dari distribusi teoritis di bawah H_0. Kemudian $S_n(X_i)$ adalah distribusi frekuensi kumulatif pengamatan sebanyak sampel. Hipotesis nol (H_0) adalah sisaan berdistribusi normal. Selanjutnya nilai D ini dibandingkan dengan nilai D kritis dengan signifikansi α (tabel *Kolmogorov-Smirnov).* Apabila nilai $D > D_{tabel}$ atau nilai kesalahan yang didapat peneliti dari perhitungan statistik yaitu $p - value$ pada *output* program, akan kurang dari nilai taraf nyata (α), maka asumsi normalitas dipenuhi.

Uji Homoskedastisitas

Salah satu asumsi klasik adalah homoskedastisitas atau non heteroskedastisitas yaitu asumsi yang menyatakan bahwa varian setiap sisaan (e_i)

masih tetap sama baik untuk nilai-nilai pada variabel independen yang kecil maupun besar. Asumsi ini dapat ditulis sebagai berikut

$$Var(e_i) = \sigma^2, \qquad i = 1,2,\dots,n$$

notasi n menunjukkan jumlah observasi. Salah satu cara menguji kesamaan variansi yaitu dengan melihat pola tebaran sisaan (e_i) terhadap nilai estimasi Y. Jika tebaran sisaan bersifat acak (tidak membentuk pola tertentu), maka dikatakan bahwa variansi sisaan homogen (Draper & Smith, 1998:65).

Untuk lebih tepatnya, menurut Gujarati (2004:406) salah satu cara untuk mendeteksi homoskedastisitas adalah menggunakan uji korelasi rank Spearman yang didefinisikan sebagai berikut

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

dengan d_i adalah *rank* variabel dependen dikurangi *rank* variabel independen yang ditempatkan pada dua karakteristik yang berbeda dari individual atau fenomena ke-i, dan n adalah banyaknya individual yang diranking. Koefisien *rank* korelasi tersebut dapat digunakan untuk mendeteksi non heterokedastisitas dengan mengasumsikan $Y_i = X_i + e_i$. Adapun tahapannya dalah sebagai berikut

1) Mencocokkan regresi terhadap data mengenai Y dan X serta mendapatkan sisaan e_i.
2) Dengan mengabaikan tanda dari e_i, yaitu dengan mengambil nilai mutlaknya $|e_i|$, meranking baik harga mutlak $|e_i|$ dan X_i sesuai dengan

urutan yang meningkat atau menurun dan menghitung koefisien rank korelasi Spearman yang telah diberikan sebelumnya.

3) Dengan mengasumsikan bahwa koefisien rank korelasi populasi ρ_s adalah nol dan $n > 8$, signifikan dari r_s yang disampel dapat diuji dengan pengujian *t* sebagai berikut :

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

Jika nilai t yang dihitung melebihi nilai t kritis dengan derajat bebas $n-2$ maka H_0 ditolak, artinya asumsi homoskedastitas tidak dipenuhi. Jika model regresi meliputi lebih dari satu variabel X, r_s dapat dihitung antara $|e_i|$ dan tiap-tiap variabel X secara terpisah dan dapat di uji untuk tingkat penting secara statistik dengan pengujian t yang diberikan di atas. Atau apabila menggunakan SPSS, nilai sig > α maka tidak terjadi heteroskedastisitas.

Uji Non Autokorelasi

Salah satu asumsi penting dari regresi linear adalah bawa tidak ada autokrelasi antara serangkaian pegamatan yang diurutkan menurut waktu. Adanya kebebasan antar sisaan dapat dideteksi secara grafis dan empiris. Pendeteksian autokorelasi secara grafis yaitu denan melihat pola tebaran sisaan terhadap urutan waktu. Jika tebaran sisaan terhadap urutan waktu tidak membentuk suatu pola tertentu atau bersifat acak maka dapat disimpulkan tidak ada autokorelasi antar sisaan (Draper & Smith, 1998:68).

Menurut Gujarati (2004:467), pengujian secara empiris dilakukan dengan menggunakan statistik uji Durbin-Watson. Hipotesis yang diuji adalah:

H_0: Tidak terdapat autokorelasi antar sisaan

H_1: Terdapat autokorelasi antar sisaan

Adapun rumusan matematis uji Durbin-Watson adalah:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Kaidah keputusan dalam uji Durbin-Watson adalah:

1. Jika $d < d_L$ atau $d > 4 - d_L$, maka H_0 ditolak berarti bahwa terdapat autokorelasi antar sisaan.
2. Jika $d_U < d < 4 - d_U$, maka H_0 tidak ditolak yang berarti bahwa asumsi non autokorelasi terpenuhi.
3. Jika $d_L \leq d \leq d_U$ atau $4 - d_U \leq d \leq 4 - d_L$ maka tidak dapat diputuskan apakah H_0 diterima atau ditolak, sehingga tidak dapat disimpulan ada atau tidak adanya autokorelasi.
4. Statistik d yaitu d_U dan d_L dari Durbin-Watson dapat dilihat pada tabel.

Uji Non Multikolinieritas

Menurut Montgomery, Peck, & Vining (2006:111), kolinearitas terjadi karena terdapat korelasi yang cukup tinggi di antara variabel independen. VIF (Variance Inflation Factor) merupakan salah satu cara untuk mengukur besar kolineritas dan didefinisikan sebagai berikut

$$VIF = \frac{1}{1 - R_j^2}$$

dengan $j = 1,2, \dots, p$

dan p adalah banyaknya variabel independen, sedangkan R_j^2 adalah koefisien determinasi yang dihasilkan dari regresi variabel independen X_j dengan variabel independen lain. Nilai VIF menjadi semakin besar jika terdapat korelasi yang semakin besar diantara variabel independen. Jika VIF lebih dari 10, multikolinearitas memberikan pengaruh yang serius pada pendugaan metode kuadrat kecil, sehingga dapat dikatakan terjadi multikolinieritas.

E. Pencilan (*Outlier*)

Pencilan (*Outlier*) merupakan pengamatan yang jauh dari pusat data observasi dari data yang lainnya dan mungkin berpengaruh besar terhadap koefisien regresi (Pardoe, 2012:189). Pencilan dalam data yang telah diperoleh akan mengganggu proses analisis data sehingga banyak dihindari pada beberapa hal. Menurut Soemartini (2007:7) dalam kaitannya dengan analisis regresi, pencilan disebabkan oleh hal-hal berikut:

1. Residual yang besar dari model yang berbentuk.
2. Varians pada data tersebut menjadi lebih besar.
3. Taksiran interval memiliki jarak yang lebar.

Menurut Soemartini (2007:14), pada analisis regresi, terdapat 3 tipe *outlier* yang mempengaruhi hasil estimasi kuadrat terkecil yaitu sebagai berikut:

1. *Vertical outlier*

   Merupakan suatu titik yang menjadi *outlier* karena memiliki koordinat Y yang ekstrim. Dengan kata lain, data yang terpencil pada sumbu Y tetapi tidak pada sumbu X.

2. *Good leverage point*

Merupakan suatu titik yang menjadi *outlier* pada variabel independen tetapi terletak dekat dengan garis linear, yang berarti bahwa observasi (x_i, y_i) apabila x_i menjauh tetapi y_i cocok dengan garis linear. *Good leverage* ini tidak berpengaruh terhadap estimasi kuadrat terkecil, tetapi berpengaruh terhadap inferensi statistik karena dapat meningkatkan estimasi standar *error*.

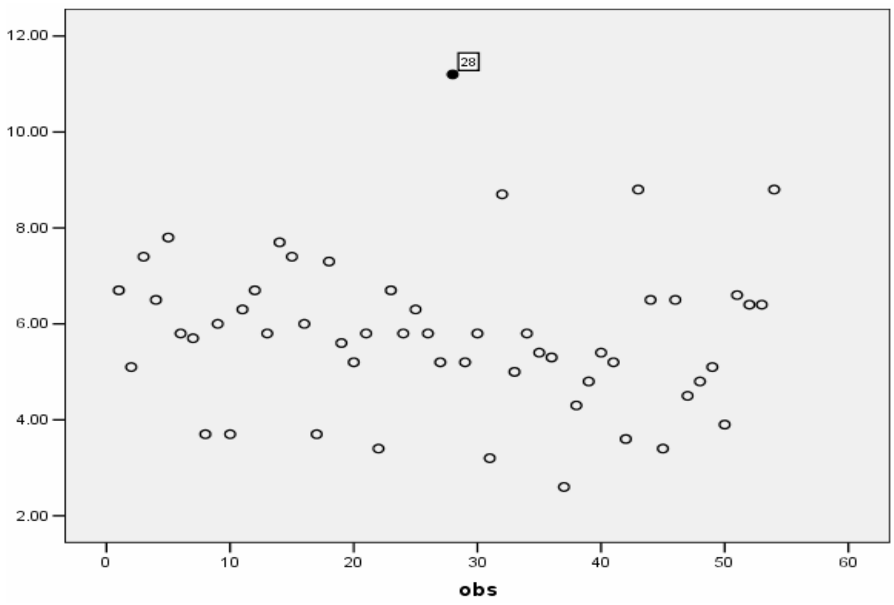
3. *Bad leverage point*

Merupakan suatu titik yang menjadi *outlier* pada variabel independen tetapi terletak jauh dengan garis linear. *Bad laverage* ini berpengaruh signifikan terhadap estimasi kuadrat terkecil.

Terdapat beberapa metode untuk menentukan batasan pencilan dalam sebuah analisis, diantaranya:

1. Metode Grafis

Metode grafis merupakan salah satu metode yang dapat digunakan untuk memecahkan masalah *linear progamming* yang menitikberatkan pada sumbu X dan Y. Dalam hal ini X dan Y merupakan variabel-variabel yang ingin dikombinasikan dan ingin dicari kombinasi yang optimal. Untuk melihat ada tidaknya pencilan pada data, dapat dilakukan dengan membuat plot sederhana antara data dengan observasi ke-i $(i = 1, 2, 3, \dots, n)$ seperti Gambar 2.1.

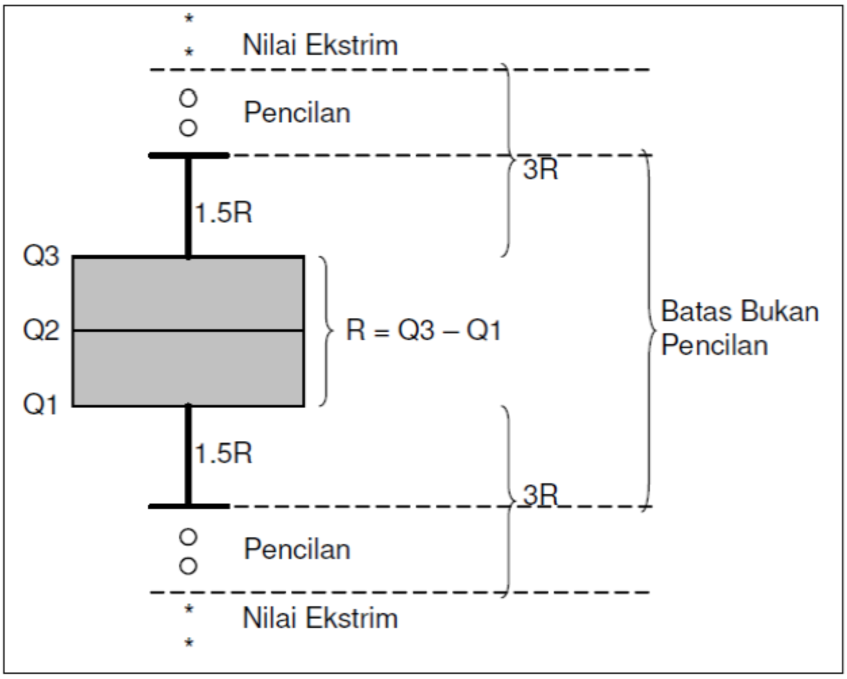


Gambar 2.1 Contoh *scatter-plot* dari data pada observasi ke-i

Dari contoh di atas terdapat salah satu data, yakni observasi ke-28 yang mengindikasikan merupakan pencilan. Selain melalui *scatter-plot* di atas, jika sudah didapatkan model regresi maka dapat dilakukan dengan cara memplot antara residual (e) dengan nilai prediksi Y $(\hat{Y})$. Jika terdapat satu atau beberapa data yang terletak jauh dari pola kumpulan data keseluruhan maka hal ini mengindikasikan adanya pencilan. Kelemahan dari metode ini adalah keputusan bahwa suatu data yang merupakan pencilan sangat bergantung pada subjektivitas peneliti karena hanya mengandalkan visualisasi grafis. Demi meminimumkan kesalahan teknis, maka pendeteksian pencilan dilakukan melalui perhitungan statistik.

2. *Boxplot*

Metode *boxplot* merupakan metode yang sering digunakan peneliti untuk mendeteksi keberadaan pencilan dengan menggunakan nilai kuartil dan jangkauan. Kuartil 1, 2, dan 3 akan membagi urutan data menjadi empat bagian. Jangkauan IQR (*Interquartile Range*) didefinisikan sebagai selisih kuartil 1 terhadap kuartil 3, yaitu IQR = Q3 – Q1. Data-data pencilan dapat ditentukan yaitu nilai yang kurang dari 1.5*IQR terhadap kuartil 1 dan nilai yang lebih dari 1.5*IQR terhadap kuartil 3, (Soemartini, 2007:9).



Gambar 2.2 Skema identifikasi pencilan menggunakan IQR atau *Boxplot*

3. Residu *Jackknife* (*R-Student*)

Salah satu metode yang digunakan untuk mengidentifikasi adanya *outlier* yang berpengaruh dalam koefisien regresi adalah residu *Jackknife*. Menurut Chatterjee & Hadi (1986:380), definisi *Jackknife* atau biasa juga disebut sebagai *R-student*, yang dilambangkan dengan t_i adalah:

$$t_i = e_i(\hat{\sigma}_i) = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \tag{2.16}$$

dengan $p = variabel + 1$ dan t_i berdistribusi t_{n-p-1} jika model asumsi terpenuhi dan $e_i \sim N(0, \sigma^2 I)$. Notasi h_{ii} merupakan elemen diagonal ke-i dari matriks *hat* (Persamaan 2.15) dan e_i merupakan residu ke-i.

Menurut Chatterjee & Hadi (1986:380), ${\hat{\sigma}_{(i)}}^2$ adalah:

$$ {\hat{\sigma}_{(i)}}^2 = \frac{\boldsymbol{x}_{(i)}^{\boldsymbol{T}}\left(\boldsymbol{I} - \boldsymbol{H}_{(ii)}\right)\boldsymbol{X}_{(i)}}{n-p-1} = \frac{(n-p)\hat{\sigma}^2}{n-p-1} - \frac{{e_i}^2}{(n-p-1)(1-h_{ii})} \qquad (2.17) $$

Matriks $\boldsymbol{X}_{(i)}$ merupakan matriks tanpa baris ke-i dan $\boldsymbol{x}_{(i)}$ merupakan matriks baris ke-i sementara $\hat{\sigma}^2$ mempunyai derajat kebebasannya $(n-p)$. Sedangkan ${\hat{\sigma}_{(i)}}^2$ mempunyai derajat kebebasannya $[(n-p)-1]$ karena observasi ke-i dihapus, dengan nilai p yaitu banyaknya variabel ditambah 1.

Nilai residu *Jackknife* yang diidentifikasi sebagai *outlier* adalah data dengan nilai *Jaccknife* atau nilai $|(\boldsymbol{t_i})|$-nya melebihi nilai kritis $t_{\frac{\alpha}{2};(n-p-1)}$, dengan *p* merupakan parameter dan *n* banyaknya observasi.

F. Regresi *Robust*

Regresi *robust* merupakan metode regresi yang digunakan ketika distribusi dari sisaan tidak normal dan/atau adanya beberapa pencilan yang berpengaruh pada model (Ryan, 1997:150). Data yang memiliki distribusi sisaan yang tidak normal pasti mengandung pencilan, akan tetapi, tidak semua data yang mengandung pencilan berdistribusi normal. Metode ini merupakan alat penting untuk menganalisa data yang dipengaruhi oleh pencilan sehingga dihasilkan model yang dapat mengatasi data yang mengandung pencilan.

Suatu estimator yang *robust* adalah relatif tidak berpengaruh oleh adanya perubahan besar pada bagian kecil data atau perubahan kecil pada sebagian besar data (Huber, 2009:8). Menurut Chen (2002:1), metode-metode estimasi dalam regresi *robust* diantaranya:

a. Estimasi-M (*Maximum likelihood type*) yang diperkenalkan oleh Huber (1973) merupakan metode yang sederhana, baik dalam perhitungan maupun secara teoritis. Metode ini memiliki nilai *breakdown point* sebesar $\frac{1}{n}$.

b. Estimasi-LMS (*Least Median Squares*) merupakan metode yang diperkenalkan oleh Hampel (1975). Metode ini memiliki nilai *breakdown point* hingga 50%.

c. Estimasi-LTS (*Least Trimmed Squares*) merupakan metode yang memiliki nilai *breakdown point* tinggi yang diperkenalkan oleh Rousseeuw (1984).

d. Estimasi-S (*Scale*) juga merupakan metode dengan memiliki nilai *breakdown point* tinggi yaitu 50% yang diperkenalkan oleh Rousseeuw dan Yohai (1984). Meski memiliki nilai *breakdown point* yang sama dengan estimasi-LTS.

e. Estimasi-MM (*Method of Moment*) merupakan metode yang diperkenalkan oleh Yohai (1987). Metode ini merupakan metode yang menggabungkan estimasi-S (estimasi yang memiliki nilai *breakdown point* tinggi) dan estimasi-M

G. Koefisien Determinasi

Koefisien determinasi atau biasa dilambangkan dengan R^2 merupakan salah satu ukuran yang sederhana dan sering digunakan untuk menguji kualitas suatu persamaan garis regresi (Gujarati, 2004:81). Nilai koefisien determinasi memberikan gambaran tentang kesesuaian variabel independen dalam memprediksi variabel dependen.

Sifat dari koefisien determinasi adalah:

a. R^2 merupakan besaran yang non-negatif

b. Batasnya adalah $0 \leq R^2 \leq 1$

Untuk mengetahui metode estimasi yang memberikan hasil yang lebih baik, maka kriteria yang digunakan adalah dengan membandingkan nilai *R-Square* (R^2) yang menunjukkan seberapa besar proporsi variasi variabel dependen yang dijelaskan oleh variabel independen (Harmi, 2012:66). Menurut Imam (2011:97), nilai R^2 yang kecil berarti kemampuan variabel-variabel independen dalam menjelaskan variasi variabel dependen sangat terbatas. Nilai yang mendekati satu berarti variabel-variabel independen memberikan hampir semua informasi yang dibutuhkan untuk memprediksi variasi variabel dependen.

Apabila nilai koefisien determinasi semakin besar, maka semakin besar kemampuan semua variabel independen dalam menjelaskan varians dari variabel dependennya. Secara sederhana koefisien determinasi dihitung dengan mengkuadratkan koefisien korelasi (R).

H. *Breakdown Point*

Breakdown point yaitu bagian terkecil data yang menyimpang yang menyebabkan nilai estimator menjadi tidak berguna (Montgomery, Peck & Vining, 2006:385). *Breakdown point* merupakan ukuran umum proporsi dari *outlier* yang dapat ditangani sebelum observasi tersebut mempengaruhi model prediksi. Menurut Sahari (2012), semakin besar nilai persentase dari *breakdown* *point* pada suatu estimator, maka estimator tersebut semakin *robust*, karena semakin besar nilai persentase *breakdown point*, maka semakin kuat juga suatu metode estimasi tersebut dalam menangani banyaknya pencilan.

Regresi *robust* yang mempunyai *breakdown point* adalah regresi *robust* dengan metode estimasi-S, LTS, LMS, dan MM. Estimasi-S dan estimasi-MM dapat digunakan untuk mengatasi masalah *outlier* dengan proporsi hingga 50%, sedangkan estimasi-M memiliki proporsi *breakdown point* sebesar $\frac{1}{n}$.