BAB II
LANDASAN TEORI

Sebagai pendukung dalam pembahasan selanjutnya, diperlukan beberapa teori dan definisi mengenai variabel random, regresi linier, metode kuadrat terkecil, pengujian asumsi analisis regresi, *outlier*, dan regresi *robust*.

A. Variabel Random

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 real x, sedemikian sehingga $X(e) = x$. Huruf besar X digunakan untuk menotasikan variabel random, sedangkan huruf kecil seperti *x* digunakan untuk menotasikan bilangan riil yang merupakan hasil nilai-nilai yang mungkin dari variabel random.

Dilihat dari segi tipe nilainya, variabel random dibedakan menjadi 2, yaitu variabel random diskrit dan variabel random kontinu.

1. Variabel Random Diskrit

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

Dalam variabel random diskrit terdapat fungsi kepadatan peluang diskrit dan 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.

Sedangkan untuk fungsi distribusi kumulatif variabel random diskrit:

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

Hal itu berarti bahwa fungsi distribusi kumulatif adalah jumlahan nilai-nilai fungsi peluang untuk nilai X lebih kecil atau sama dengan *x*. Fungsi $F(x)$ disebut fungsi distribusi kumulatif diskrit jika dan hanya jika memenuhi:

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

Fungsi tersebut mempunyai sifat-sifat:

(1) $\lim_{x \to -\infty} F(x) = 0$

(2) $\lim_{x \to \infty} F(x) = 1$

(3) $\lim_{h \to -0^+} F(x + h) = F(x)$

(4) $a < b$, maka $F(a) \leq F(b)$ (2.2)

2. Variabel Random Kontinu

Jika nilai yang mungkin variabel random X adalah sebuah interval atau kumpulan interval-interval, maka X disebut variabel random kontinu. Pada variabel

random kontinu mempunyai fungsi kepadatan peluang yang merupakan turunan dari fungsi distribusi kumulatifnya.

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:

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

Sebuah fungsi $f(x)$ disebut fungsi kepadatan peluang dari variabel random kontinu X jika memenuhi:

(1) $f(x_i) \geq 0 \quad , \forall x_i$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

B. Regresi Linier

Pengertian regresi secara umum adalah sebuah metode dalam statistik yang memberikan penjelasan tentang pola hubungan antara dua variabel atau lebih. Dalam analisis regresi dikenal 2 jenis variabel, yaitu:

(1) variabel respon atau variabel dependen yaitu variabel yang keberadaaanya dipengaruhi oleh variabel lainnya dan dinotasikan dengan variabel Y; dan

(2) variabel prediktor atau variabel dependen yaitu variabel yang tidak dipengaruhi oleh variabel lainnya dan dinotasikan dengan X.

1. Model Regresi Linier Sederhana

Regresi linier sederhana digunakan untuk mendapatkan hubungan matematis dalam bentuk satu persamaan antara satu variabel independen dengan satu variabel dependen. Menurut Sembiring (1995: 32), model regresi adalah

model yang memberikan gambaran mengenai hubungan antara variabel bebas dengan variabel terikat. Jika analisis dilakukan untuk satu variabel bebas dengan variabel terikat, maka regresi ini disebut regresi linier sederhana. Menurut Draper & Smith (1998: 22) bentuk umum dari regresi linier sederhana adalah sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (2.3)$$

dengan,

Y_i : nilai variabel dependen pada observasi ke-i

X_i : nilai variabel independen pada observasi ke-i

β_0, β_1 : parameter koefisien regresi

ε_i : *error* yang bersifat random

2. Model Regresi Linier Berganda

Regresi linier berganda adalah suatu analisis yang digunakan untuk mempelajari hubungan sebuah variabel dependen dengan dua atau lebih variabel independen. Menurut Montgomery & Peck (1992: 53), model regresi linier berganda dari variabel dependen Y dengan variabel independen $X_1, X_2, \ldots, X_k$ dapat ditulis sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

atau dapat ditulis

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \quad i = 1,2, \ldots, n \quad (2.4)$$

dengan:

y_i : nilai variabel dependen pada observasi ke-i

$\beta_0, \beta_1, \ldots, \beta_k$: parameter koefisien regresi

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

e_i : *random error*

3. Uji Asumsi dalam Analisis Regresi

Menurut Imam Ghozali (2011: 160), uji asumsi klasik terhadap model regresi yang digunakan dilakukan agar dapat diketahui apakah model regresi baik atau tidak. Tujuan pengujian asumsi klasik adalah untuk memberikan kepastian bahwa persamaan regresi yang diperoleh memiliki ketepatan dalam estimasi, tidak bias, dan konsisten. Sebelum melakukan analisis regresi, terlebih dahulu dilakukan pengujian asumsi. Asumsi-asumsi yang harus dipenuhi dalam analisis regresi, antara lain: normalitas, homoskedastisitas, non autokorelasi, dan non multikolinieritas.

a. Uji Normalitas

Analisis regresi linier mengasumsikan bahwa sisaan (ε_i) berdistribusi normal. Pada regresi linier diasumsikan bahwa tiap sisaan (ε_i) berdistribusi normal dengan $\varepsilon_i \sim N(0, \sigma^2)$ (Gujarati, 2004: 109). Uji normalitas bertujuan untuk mengetahui apakah dalam persamaan regresi tersebut residual berdistribusi normal. Uji normalitas dapat dilakukan dengan normal P-P Plot dan uji *Kolmogorov-Smirnov*. Normal P-P plot, uji normalitasnya dapat dilihat dari penyebaran data (titik) pada sumbu diagonal grafik atau dengan melihat histogram dari residunya. Dasar pengambilan keputusannya, jika data menyebar di sekitar garis diagonal dan mengikuti arah garis diagonal atau grafik histogramnya menunjukkan pola distribusi normal, maka model regresi memenuhi asumsi normalitas.

Cara lain untuk menguji asumsi kenormalan adalah dengan uji *Kolmogorov-Smirnov*. Menurut sidney Siegel (1986: 59), uji *Kolmogorov-Smirnov* didasarkan pada nilai D atau deviasi maksimum, yaitu:

$$D = \max|F_0(X_i) - S_n(X_i)|, i = 1,2, \dots, n \quad (2.5)$$

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. Kriteria keputusan uji *Kolmogorov-Smirnov* adalah jika nilai $D < D_{tabel}$ atau $p - value$ pada *output* SPSS lebih dari nilai taraf nyata (α) maka asumsi normalitas dipenuhi. Tabel uji *Kolmogorov-Smirnov* dapat dilihat pada lampiran 6 (halaman: 78).

b. Uji Homoskedastisitas

Salah satu asumsi klasik adalah homoskedastisitas atau non heteroskedastisitas yaitu asumsi yang menyatakan bahwa varian setiap sisaan (ε_i) masih tetap sama baik untuk nilai-nilai pada variabel independen yang kecil maupun besar. Asumsi ini dapat ditulis sebagai berikut :

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

notasi n menunjukkan jumlah observasi. Salah satu cara menguji kesamaan variansi yaitu dengan melihat pola tebaran sisaan (ε_i) terhadap nilai estimasi Y. Hal ini dapat dilihat dari plot data, jika tebaran sisaan bersifat acak (tidak membentuk pola tertentu), maka dikatakan bahwa variansi sisaan homogen. Model regresi yang baik adalah tidak terjadi heteroskedastisitas. Meskipun demikian, untuk meyakinkan plot data tersebut bersifat homoskedastisitas perlu dilakukan

pengujian statistik lain. Salah satu pengujian untuk menentukan ada tidaknya masalah heteroskedastisitas adalah uji *Glejser*.

Uji *Glejser* dapat dilakukan dengan meregresikan nilai absolut residual terhadap variabel independen. Jika varians residual dari satu pengamatan ke pengamatan lain tetap maka disebut homoskedastisitas (Imam Ghozali, 2011: 125). Langkah-langkah pengujian:

(1) Mencari nilai residual ε_i menggunakan persamaan

$$\varepsilon_i = y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}.$$

(2) Mencari nilai absolut residual $|\varepsilon_i|$.

(3) Melakukan analisis regresi dengan variabel $|\varepsilon_i|$ sebagai variabel dependen dan X_{ik} sebagai variabel independen.

(4) Penilaian berdasarkan uji t dengan hipotesis sebagai berikut:

H_0: tidak terjadi heteroskedastisitas

H_1: terjadi heteroskedastisitas

kriteria keputusan untuk uji t, jika nilai signifikansi untuk masing-masing variabel independen pada persamaan model regresi terhadap nilai absolut residualnya lebih dari 0,05 atau nilai $-t_{(\alpha,db)} < t_{hitung} < t_{(\alpha,db)}$ dengan derajat bebas $(db) = n - k$, n: banyaknya data, dan k: banyaknya variabel bebas maka H_0 diterima, artinya tidak terjadi heteroskedastisitas.

c. Uji Non Autokorelasi

Salah satu asumsi penting dari regresi linear adalah bawa tidak ada autokorelasi antara serangkaian pegamatan yang diurutkan menurut waktu. Adanya kebebasan antar sisaan dapat dideteksi secara grafis dan empiris. Pendeteksian

autokorelasi secara grafis yaitu dengan 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

Mekanisme uji *Durbin watson* adalah:

(1) Mengestimasi model regresi dengan metode kuadrat terkecil untuk memperoleh nilai ε_i .

(2) Mencari nilai *d* yang diperoleh dengan rumus

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}{\varepsilon_i}^2} \tag{2.6}$$

(3) Untuk ukuran sampel dan banyaknya variabel tertentu dapat dilihat pada tabel *Durbin-Watson* mengenai pasangan nilai kritis d_L dan d_U (lampiran 6).

(4) Kriteria keputusan dalam uji *Durbin-Watson* adalah:

1. Jika $d < d_L$ atau $d > 4 - d_L$, maka H_0 ditolak artinya terjadi autokorelasi.
2. Jika $d_U < d < 4 - d_U$, maka H_0 diterima artinya tidak terjadi autokorelasi.
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 disimpulkan ada tidaknya autokorelasi.

d. Uji Non Multikolinieritas

Menurut Montgomery, Peck, & Vining (1992: 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

$$VIF = \frac{1}{1-R_j^2} \tag{2.7}$$

dengan $j = 1,2, \dots, k$ dan k adalah banyaknya variabel independen, sedangkan R_j^2 adalah koefisien determinasi yang dihasilkan dari regresi variabel independen X_j dengan variabel independen lain. Hipotesis nol (H_0) pengujian multikolinieritas adalah tidak terdapat multikolinieritas, dengan kriteria keputusan jika nilai $VIF < 10$ maka H_0 diterima artinya tidak terdapat multikolinieritas.

C. Metode Kuadrat Terkecil

Salah satu metode untuk mengestimasi parameter dalam model regresi adalah 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 digunakan untuk mengestimasi koefisien $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ yaitu dengan meminimumkan jumlah kuadrat galat. Fungsi yang meminimumkan adalah:

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

$$= \textstyle\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2 \tag{2.8}$$

Fungsi S akan diminimumkan 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

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

dan

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_k} = -2 \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) x_{ij} = 0 \qquad (2.10)$$

Selanjutnya dari persamaan (2.9) dan (2.10), menghasilkan persamaan normal kuadrat terkecil sebagai berikut:

$$\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{k}x_{i1} + \hat{\beta}_1\sum_{i=1}^{k}x_{i2} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik} = \sum_{i=1}^{n}y_i \\
&\hat{\beta}_0\sum_{i=1}^{n}x_{i1} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^{\,2} + \hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} = \sum_{i=1}^{n}x_{i1}y_i \\
&\hat{\beta}_0\sum_{i=1}^{n}x_{i2} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}x_{i2} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2}^{\,2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{i2}x_{ik} = \sum_{i=1}^{n}x_{i2}y_i \\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
&\hat{\beta}_0\sum_{i=1}^{n}x_{ik} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}x_{ik} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2}x_{ik} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^{\,2} = \sum_{i=1}^{n}x_{ik}y_i \qquad (2.11)
\end{aligned}$$

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 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. Akan lebih mudah apabila model regresi dinyatakan dalam matriks. Notasi matriks yang diberikan pada persamaan (2.8) adalah

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

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

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:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{e_i}^2 = \boldsymbol{e^T e} \\ &= (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y^T Y} - \boldsymbol{Y^T X\beta} - \boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X\beta} \\ &= \boldsymbol{Y^T Y} - \boldsymbol{2\beta^T X^T Y} + \boldsymbol{\beta^T X^T X\beta} \end{aligned} \quad (2.12)$$

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,

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

$$= \frac{\partial(\boldsymbol{Y^T Y})}{\partial \boldsymbol{\beta}} - 2\frac{\partial(\boldsymbol{\beta^T X^T Y})}{\partial \boldsymbol{\beta}} + \frac{\partial(\boldsymbol{\beta^T X^T X\beta})}{\partial \boldsymbol{\beta}}$$

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

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

sehingga,

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

$$= -\boldsymbol{2X^T Y} + \boldsymbol{2X^T X\widehat{\beta}}. \qquad (2.13)$$

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 (2.13), akan diperoleh estimator untuk $\boldsymbol{\beta}$, yaitu:

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

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

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

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

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

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

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.15}$$

Diasumsikan bahwa invers matriks $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ ada. Diperoleh matriks dari persamaan normal (2.14) yang identik dengan bentuk skalar pada persamaan (2.11). Dari persamaan (2.14) 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 $\boldsymbol{X^TX}$ adalah matrik persegi berukuran $k \times k$ dan $\boldsymbol{X^TY}$ adalah vektor $k \times 1$. Diagonal elemen matriks $\boldsymbol{X^TX}$ 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

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

sehingga

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

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

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

dengan matriks persegi yang disebut matriks *hat*

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T} \tag{2.16}$$

D. Outlier

Outlier adalah kasus atau data yang memiliki karakteristik unik yang penyebaran datanya terlihat jauh dari observasi-observasi lainnya dan muncul dalam bentuk nilai ekstrim, baik untuk sebuah variabel tunggal maupun variabel kombinasi (Imam Ghozali, 2011: 40).

Menurut Ghozali (2011: 40), terdapat empat penyebab timbulnya data *outlier* antara lain: (1) kesalahan dalam memasukan data; (2) gagal dalam menspesifikasi adanya *missing value* dalam program komputer; (3) *outlier* bukan merupakan anggota populasi yang di ambil sebagai sampel; dan (4) *outlier* berasal dari populasi yang di ambil sebagai sampel, tetapi distribusi dari variabel dalam populasi tersebut memiliki nilai ekstrim serta tidak berdistribusi secara normal.

Pada analisis regresi, terdapat 3 tipe *outlier* yang berpengaruh terhadap estimasi kuadrat terkecil. Menurut Roesseuw dan Leroy (1987), mengenalkan 3 jenis *outlier* tersebut sebagai *vertical outlier*, *good leverage* dan *bad leverage*.

a. *Vertical outlier*

Merupakan semua pengamatan yang terpencil pada variabel respon, tetapi tidak terpencil pada variabel prediktor. Keberadaan *vertical outlier* berpengaruh terhadap estimasi kuadrat terkecil.

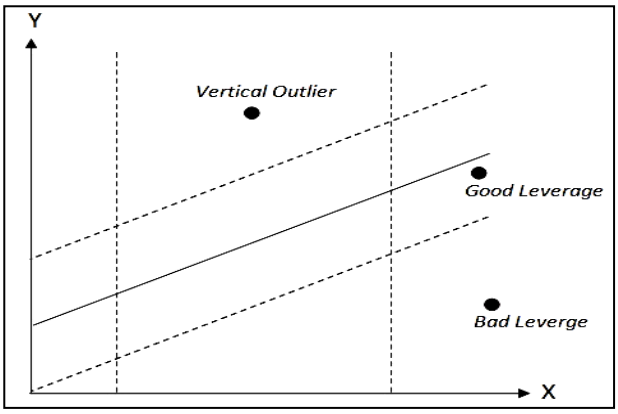
b. *Good leverage point*

Merupakan pengamatan yang terpencil pada variabel prediktor tetapi terletak dekat dengan garis regresi. Hal ini berarti pengamatan x_i menjauh tetapi y_i cocok dengan garis regresi. Keberadaan *good leverage points* tidak berpengaruh terhadap estimasi kuadrat terkecil, tetapi berpengaruh terhadap inferensi statistik karena dapat meningkatkan estimasi standar error.

c. *Bad leverage point*

Merupakan pengamatan yang terpencil pada variabel prediktor dan terletak jauh dari garis regresi. Keberadaan *bad leverage points* berpengaruh signifikan terhadap estimasi kuadrat terkecil, baik terhadap intersep maupun *slope* dari persamaan regresi.

Perbedaan antara *vertical outlier*, *good leverage* dan *bad leverage* dapat dilihat pada gambar dibawah ini.



Gambar 2.1 *Vertical Outlier*, *Good Leverage* dan *Bad Leverage*

Outlier berpengaruh terhadap proses analisis data, misalnya terhadap nilai mean dan standar deviasi. Oleh karena itu, keberadaan *outlier* dalam suatu pola data harus dihindari. *Outlier* dapat menyebabkan varians pada data menjadi lebih besar, interval dan *range* menjadi lebar, *mean* tidak dapat menunjukkan nilai yang sebenarnya (bias) dan pada beberapa analisis inferensi, *outlier* dapat menyebabkan kesalahan dalam pengambilan keputusan dan kesimpulan. Berbagai kaidah telah diajukan untuk menolak *outlier*, dengan kata lain untuk memutuskan menyisihkan *outlier* tersebut dari data, kemudian menganalisis kembali tanpa *outlier* tersebut. Penghilangan suatu *outlier* begitu saja bukanlah prosedur yang bijaksana. Adakalanya *outlier* memberikan informasi yang tidak bisa diberikan oleh data lainnya, misalnya karena *outlier* timbul dari kombinasi keadaan yang tidak biasa yang mungkin saja sangat penting dan perlu diselidiki lebih jauh. Secara filosofi *outlier* seharusnya tetap dipertahankan jika data *outlier* tersebut memang representasi dari populasi. Sebagai kaidah umum *outlier* baru dikeluarkan jika setelah ditelusuri ternyata merupakan akibat dari kesalahan ketika menyiapkan peralatan.

1. Dampak *Outlier*

Keberadaan data *outlier* akan mengganggu dalam proses menganalisis data dan harus dihindari dalam banyak hal. Dalam kaitannya dengan analisis regresi, *outlier* dapat menyebabkan hal-hal sebagai berikut (Soemartini, 2007: 7):

a. Residual yang besar dari model yang terbentuk
b. Variansi pada data tersebut menjadi lebih besar
c. Estimasi interval akan memiliki rentang yang lebih besar

2. Deteksi *Outlier*

Dalam statistik, data *outlier* harus dilihat dari posisi dan sebaran data yang lainnya sehingga akan dievaluasi apakah data *outlier* tersebut perlu dihilangkan atau tidak. Terdapat beberapa metode untuk menentukan batasan *outlier* dalam sebuah analisis, yaitu:

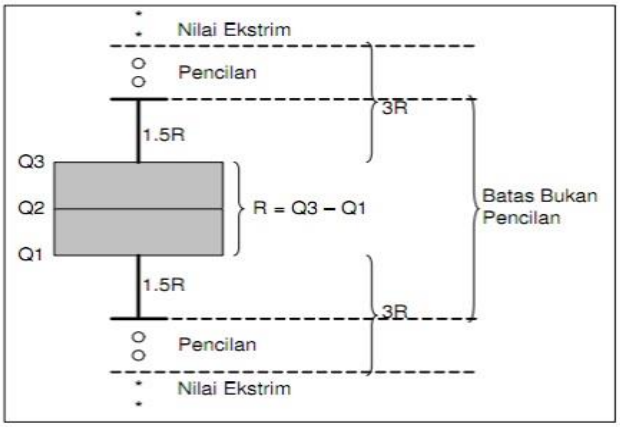
a. Scatter plot

Untuk melihat apakah terdapat *outlier* pada data, dapat dilakukan dengan membentuk diagram pencar (*scatter plot*) dari data. Jika terdapat satu atau beberapa data yang terletak jauh dari pola kumpulan data maka hal ini mengindikasikan adanya *outlier*. Kelemahan dari metode ini adalah keputusan bahwa suatu data merupakan *outlier* sangat bergantung pada *judgement* peneliti. Karena hanya mengandalkan visualisasi grafis, untuk itu dibutuhkan seseorang yang ahli dan berpengalaman dalam menginterprestasikan plot tersebut.

b. Box plot

Box plot merupakan metode grafis yang dikembangkan oleh Tukey dan sering digunakan untuk analisis data dan diintepretasikan untuk memperoleh informasi dari sebuah sampel. Metode ini merupakan metode paling umum yakni dengan mempergunakan nilai kuartil dan jangkauan. Kuartil 1, 2, dan 3 akan membagi sebuah urutan data menjadi empat bagian. Jangkauan (IQR, *Interquartile Range*) didefinisikan sebagai selisih kuartil 1 terhadap kuartil 3, atau $IQR = Q_3 - Q_1$. Data-data *outlier* dapat ditentukan yaitu nilai yang kurang dari $1{,}5 \times IQR$ terhadap kuartil 1 dan nilai yang lebih dari $1{,}5 \times IQR$ terhadap kuartil 3

(Soemartini, 2007: 9). Skema identifikasi data *outlier* dengan IQR atau *Box Plot* dapat dilihat pada gambar dibawah ini.



Gambar 2.2 Skema Identifikasi Data *Outlier* dengan IQR atau *Box Plot*

c. Standarized Residual

Pendeteksian *outlier* menggunakan metode ini yaitu dengan memeriksa residual. Rumus residual ke-i adalah sebagai berikut:

$$\varepsilon_i = y_i - \hat{y}_i \tag{2.17}$$

Sesuai dengan residual ke-i di atas, dapat didefinisikan standardized residual ke-i sebagai berikut:

$$\varepsilon_{iS} = \frac{\varepsilon_i}{\sqrt{MSE}} \tag{2.18}$$

dengan $MSE = \frac{\sum_{i=1}^{n} {\varepsilon_i}^2}{n-k}$

n : banyaknya data

k : banyaknya variabel independen

Mean Squared Error (MSE) adalah rata-rata residual kuadrat dan akar dari MSE disebut standar error. Standar error merupakan ukuran kebaikan model regresi. Standar error mengukur besarnya variansi model regresi, semakin kecil

nilainya semakin baik model regresinya. Untuk melakukan identifikasi *outlier*, diperhatikan nilai-nilai dari *standardized residual*. Jika nilai dari *standardized residual* lebih dari 3,5 atau kurang dari -3,5 maka data tersebut dikatakan sebaigai *outlier* (Yaffe, 2002: 35).

d. Cook's Distance

Metode ini diperkenalkan oleh Cook (1977), dengan rumus sebagai berikut:

$$D_i = \left(\frac{1}{k}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)\left(\frac{{\varepsilon_i}^2}{MSE(1-h_{ii})}\right) \tag{2.19}$$

k : banyaknya variabel bebas

h_{ii} : nilai $laverage$ untuk kasus ke $-$ i

Untuk kasus regresi sederhana maupun regresi berganda nilai h_{ii} dapat diuraikan sebagai berikut:

$$h_{ii} = \begin{bmatrix}1 & x_i\end{bmatrix}\begin{bmatrix}\frac{\sum_{i=1}^{n}{x_i}^2}{nS_{xx}} & \frac{-\bar{x}}{S_{xx}} \\ \frac{-\bar{x}}{S_{xx}} & \frac{1}{S_{xx}}\end{bmatrix}\begin{bmatrix}1 \\ x_i\end{bmatrix} \tag{2.20}$$

$$= \begin{bmatrix}\frac{\sum_{i=1}^{n}{x_i}^2}{nS_{xx}} - \frac{\bar{x}x_i}{S_{xx}} & \frac{-\bar{x}}{S_{xx}} + \frac{x_i}{S_{xx}}\end{bmatrix}\begin{bmatrix}1 \\ x_i\end{bmatrix}$$

$$= \frac{\sum_{i=1}^{n}{x_i}^2}{nS_{xx}} - \frac{\bar{x}x_i}{S_{xx}} + x_i\left(\frac{-\bar{x}}{S_{xx}} + \frac{x_i}{S_{xx}}\right)$$

$$= \frac{\sum_{i=1}^{n}{x_i}^2}{nS_{xx}} - \frac{\bar{x}x_i}{S_{xx}} - \frac{\bar{x}x_i}{S_{xx}} + \frac{{x_i}^2}{S_{xx}}$$

$$= \frac{\sum_{i=1}^{n}{x_i}^2}{nS_{xx}} + \frac{{x_i}^2}{S_{xx}} - 2\frac{\bar{x}x_i}{S_{xx}}$$

$$h_{ii} = \frac{\frac{\sum_{i=1}^{n} {x_i}^2}{n}}{S_{xx}} - \frac{(\bar{x})^2}{S_{xx}} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

$$= \left(\frac{\frac{\sum_{i=1}^{n} {x_i}^2}{n} - (\bar{x})^2}{S_{xx}} \right) + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

dengan h_{ii} dan n jumlah data pengamatan, suatu data disebut *outlier* apabila nilai $D_i > 4/n$ (Yaffe, 2002: 44).

E. Breakdown Point

Menurut Huber (1981: 13) mendefinisikan *breakdown point* sebagai fraksi terkecil atau persentase dari *outlier* yang menyebabkan nilai dari estimator menjadi besar. Berdasarkan definisi tersebut maka jelas bahwa dalam kasus univariat median memiliki nilai *breakdown point* sebesar 50% sedangkan mean memiliki nilai *breakdown point* sebesar 0. *Breakdown point* digunakan untuk menjelaskan ukuran kerobustan dari tekhnik *robust*. Kemungkinan tertinggi *breakdown point* untuk sebuah estimator adalah 50%. Jika *breakdown point* lebih dari 50% berarti estimasi model regresi tidak dapat menggambarkan informasi dari mayoritas data. Beberapa contoh nilai-nilai *breakdown point* sebagai berikut:

1. Nilai *breakdown point* untuk mean sampel

   Dinyatakan suatu n sampel random $x_1, x_2, \dots, x_n$ dengan mean sampel dinyatakan dengan $\frac{x_1, x_2, \dots, x_n}{n}$. Jika $x_1, x_2, \dots, x_{n-1}$ tetap dan x_n diubah menjadi tak berhingga maka mean sampel juga menjadi tak berhingga, dengan kata lain *outlier* mempengaruhi nilai mean. Sampel berhingga mempunyai *breakdown*

point sebesar $\frac{1}{n}$, sedangkan asimtotik *breakdown point* memiliki nilai sebesar 0.

2. Nilai *breakdown point* untuk median

   Dinyatakan n sampel random , kemudian $\left[\frac{(n-1)}{2}\right]$ diubah menjadi tak berhingga. Maka nilai median akan berubah tapi tidak terlalu buruk. Median pada sampel berhingga memiliki *breakdown point* sebesar $\left[\frac{(n-1)}{2n}\right]$ dan asimtotik *breakdown point* sebesar $\frac{1}{2}$.

F. Regresi *Robust*

Regresi *robust* diperkenalkan oleh Andrews (1972). Regresi *robust* merupakan metode regresi yang digunakan ketika distribusi dari error tidak normal dan atau adanya beberapa *outlier* yang berpengaruh pada model (Olive, 2005: 3). Regresi *robust* digunakan untuk mendeteksi *outlier* dan memberikan hasil yang resisten terhadap adanya *outlier*. Efisiensi dan *breakdown point* digunakan untuk menjelaskan ukuran kerobustan dari tekhnik *robust*. Efisiensi menjelaskan seberapa baiknya suatu tekhnik *robust* sebanding dengan metode kuadrat terkecil tanpa *outlier*. Semakin tinggi efisiensi dan *breakdown point* dari suatu estimator maka semakin *robust* (resisten) terhadap *outlier*. Ukuran statistik yang bersifat *robust* ini ditunjukkan untuk mengakomodasi keberadaan data ekstrim dan sekaligus meniadakan pengaruhnya terhadap nilai analisis tanpa terlebih dahulu mengadakan identifikasi terhadapnya.

Metode regresi *robust* merupakan metode yang mempunyai sifat: (1) sama baiknya dengan metode kuadrat terkecil jika semua asumsi klasik regresi terpenuhi

dan tidak terdapat data *outlier*; (2) dapat menghasilkan model regresi yang lebih baik daripada metode kuadrat terkecil jika asumsi tidak terpenuhi dan terdapat data outlier; dan (3) perhitungannya cukup sederhana dan mudah dimengerti, tetapi dilakukan secara iteratif sampai diperoleh estimasi terbaik yang mempunyai standar error parameter yang paling kecil.

Menurut Chen (2002: 1), terdapat 3 kelas masalah yang dapat menggunakan tekhnik regresi *robust* yaitu:

(1) masalah dengan *outlier* yang terdapat pada variabel Y;

(2) masalah dengan *outlier* yang terdapat pada variabel X (*leverage points*); dan

(3) masalah dengan *outlier* yang terdapat pada keduanya yaitu variabel Y dan variabel X.

Banyak metode yang dikembangkan dalam regresi untuk mengatasi masalah *outlier*. Dalam regresi *robust* terdapat beberapa metode estimasi yaitu:

1. Estimasi-M

Wilcox (2005: 51) menjelaskan estimasi-M pertama kali diperkenalkan oleh Huber pada tahun 1973 dan merupakan penggambaran dari suatu percobaan yang menggabungkan metode kuadrat terkecil dan ketahanan estimasi yang meminimumkan jumlah nilai mutlak dari residual. Estimasi ini merupakan estimasi paling sederhana baik secara perhitungan maupun teoritis. Meskipun estimasi ini tidak cukup kekar dengan *leverage point*, estimasi ini tetap digunakan secara luas dalam menganalisis data dengan mengasumsikan bahwa sebagian besar data yang terkontaminasi *outlier* merupakan data pada variabel respon.

2. Estimasi *Least Trimmed Squares* (LTS)

   Estimasi LTS diperkenalkan oleh Rousseeuw pada tahun 1984 ini adalah metode estimasi dengan nilai *breakdown point* tinggi. Metode ini kemudian dikembangkan oleh Rousseuw dan Van Driessen pada tahun 1998 dengan algoritma cepat LTS.

3. Estimasi-S

   Metode regresi *robust* estimasi-S merupakan metode *high breakdown value* yang diperkenalkan pertama kali oleh Rousseuw dan Yohai pada tahun 1984. Menurut Wilcox (2005: 55), estimasi-S merupakan solusi dengan kemungkinan terkecil dari penyebaran residual. Estimasi-S mempunyai nilai *breakdown point* tinggi sebesar 50%, estimasi ini memiliki efisiensi statistik yang lebih tinggi dibanding estimasi-LTS.

4. Estimasi-MM

   Wilcox (2005: 56) menjelaskan metode estimasi-MM diperkenalkan oleh Yohai pada tahun 1987 merupakan kombinasi dari estimasi-S dan estimasi-M. Estimasi ini memiliki nilai *breakdown point* yang tinggi dan memiliki efisiensi statistik yang lebih besar dibanding estimasi-S.

G. *R-Square* dan *Adjusted R-Square*

R-Square atau koefisien determinasi merupakan salah satu ukuran yang sederhana dan sering digunakan untuk menguji kualitas suatu persamaan garis regresi (Gujarati, 2004: 81). Nilai *R-Square* memberikan gambaran tentang kesesuaian variabel independen dalam memprediksi variabel dependen. Adapun perhitungan nilai *R-Square* adalah sebagai berikut:

$$R^2 = \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} \quad (2.21)$$

Sifat dari *R-Square* 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. Menurut Imam Ghozali (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.

Masalah yang terjadi jika melakukan pengujian dengan menggunakan *R-Square* adalah jika variabel bebasnya lebih dari satu maka nilai *R-Square* akan bertambah besar. Pengujian dengan *adjusted R-Square* ($\bar{R}$) secara obyektif melihat pengaruh penambahan variabel bebas, apakah variabel tersebut mampu memperkuat variasi penjelasan variabel terikat. Adapun perhitungan nilai *adjusted R-Square* adalah sebagai berikut:

$$\bar{R} = 1 - (1 - R^2) \times \frac{n-1}{n-k} \quad (2.22)$$

dengan n: banyaknya data observasi dan k: banyaknya variabel independen.