

BAB II LANDASAN TEORI

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

A. Regresi Linear Berganda

Analisis regresi secara konseptual merupakan metode sederhana untuk memeriksa hubungan antara variabel (Chatterjee & Hadi, 1986). Hubungan antara variabel yang dimaksudkan tersebut digambarkan dalam bentuk persamaan atau model yang menghubungkan antara variabel dependen (Y) dan satu atau lebih variabel independen (X).

Variabel dependen dinotasikan dengan Y dan himpunan dari variabel independen dinotasikan dengan $X_1, X_2, \dots, X_k$, dimana k merupakan jumlah variabel independen. Model regresi linear yang terdiri dari satu variabel dependen dan satu variabel independen disebut dengan regresi linear sederhana, sedangkan model regresi linear yang terdiri dari beberapa variabel independen dan satu variabel dependen merupakan model regresi linear berganda. Model regresi linear berganda (Faraway, 2002):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + e_i \quad (2.1)$$

dengan Y_i merupakan nilai variabel dependen dalam observasi ke- i , $X_{i1}, X_{i2}, \dots, X_{ik}$ merupakan variabel independen pada observasi ke- i dan parameter ke- k , dan $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ merupakan parameter regresi yang tidak diketahui

nilainya dan akan dicari nilai estimasinya, e_i merupakan galat yang berdistribusi normal dengan *mean*-nya nol dan variansinya σ^2 atau $e_i \sim N(0, \sigma^2)$.

Selain menggunakan notasi pada persamaan (2.1), penggunaan matriks terhadap regresi linear mempunyai banyak keuntungan yaitu menyajikan bentuk ringkas untuk menangani model regresi yang memuat banyak variabel. Persamaan (2.1) merupakan penjabaran dari himpunan n persamaan berikut (Faraway, 2002):

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_k X_{1k} + e_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_k X_{2k} + e_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_k X_{nk} + e_n \end{aligned} \quad (2.2)$$

Dalam bentuk matriks persamaan (2.2) menjadi

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad (2.3)$$

Persamaan (2.3) dapat ditulis secara sederhana sebagai berikut

$$Y = X\beta + e \quad (2.4)$$

Keterangan:

Y merupakan vektor observasi variabel dependen yang berukuran $n \times 1$

X merupakan variabel independen yang berukuran $n \times (k + 1)$

β merupakan vektor koefisien variabel independen yang berukuran $k \times 1$
dari parameter $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ yang tidak diketahui

e merupakan vektor galat yang berukuran $n \times 1$

B. Metode Kuadrat Terkecil (MKT)

Metode kuadrat terkecil merupakan suatu metode yang digunakan untuk mengestimasi $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ dengan cara meminimumkan jumlah kuadrat galat. Parameter $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ tidak diketahui dan perlu dicari nilai estimasinya (Montgomery, Peck, & Vining, 2006). Dari persamaan (2.1) dapat ditulis (Eye & Schuster, 1998):

$$Q(\beta_j) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik})^2 \quad (2.5)$$

Pada notasi matriks jumlah kuadrat galat e_i^2 dapat ditulis sebagai berikut

$$e_i^T e_i = \begin{bmatrix} 1 & e_2 & \dots & e_i \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \end{bmatrix} = e_1^2 + e_2^2 + \dots + e_i^2 = e_i^2 \quad (2.6)$$

Berdasarkan persamaan (2.3) diperoleh

$$e = Y - X\beta \quad (2.7)$$

Oleh karena itu, perkalian matriks galat menjadi (Hocking, 2003):

$$e_i^T e_i = Y - X\beta^T (Y - X\beta)$$

$$e_i^T e_i = Y^T Y - Y^T \beta - X^T \beta^T + X^T \beta^T X$$

$$(\text{karena } X^T \beta^T Y = Y^T X \beta)$$

$$e_i^T e_i = Y^T Y - 2Y^T \beta^T Y + X^T \beta^T X \beta \quad (2.8)$$

Untuk mencari nilai-nilai β yaitu dengan meminimumkan jumlah kuadrat galat, kemudian dicari turunan dari Q β_j secara parsial terhadap $\beta_j, j = 0, 1, 2, \dots, k$ dan disama dengan nol, sehingga diperoleh persamaan normal

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^n y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik} = 0, \\
\frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^n x_{i1} y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik} = 0, \\
\frac{\partial Q}{\partial \beta_2} &= -2 \sum_{i=1}^n x_{i2} y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik} = 0, \\
&\vdots \\
\frac{\partial Q}{\partial \beta_k} &= -2 \sum_{i=1}^n x_{ik} y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik} = 0
\end{aligned} \tag{2.9}$$

Setelah disusun kembali dan mengganti semua parameter dengan estimatornya, maka sistem persamaan (2.9) dapat ditulis sebagai

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

Persamaan (2.10) disebut dengan persamaan normal. Jika ditulis dalam bentuk matriks maka bentuknya menjadi

$$\begin{aligned}
&\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \dots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i2} x_{i1} & \dots & \sum_{i=1}^n x_{ik} x_{i1} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1} x_{i2} & \sum_{i=1}^n x_{i2}^2 & \dots & \sum_{i=1}^n x_{ik} x_{i2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1} x_{ik} & \sum_{i=1}^n x_{i2} x_{ik} & \dots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \\
&\quad \quad \quad (X^T X) \quad \quad \quad \beta \quad \quad \quad X^T \quad \quad \quad Y
\end{aligned}$$

atau secara lengkap jika ditulis kedalam bentuk matriks menjadi

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y} \quad (2.11)$$

Pada persamaan (2.11) kedua ruasnya dikalikan invers dari matriks $\mathbf{X}^T \mathbf{X}$, sehingga diperoleh:

$$\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\mathbf{I} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\boldsymbol{\beta} = \mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \mathbf{Y}$$

Sehingga diperoleh estimator untuk MKT adalah

$$\boldsymbol{\beta} = \mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \mathbf{Y} \quad (2.12)$$

C. Pengujian Asumsi Analisis Regresi

Pengujian asumsi analisis regresi merupakan pengujian asumsi-asumsi statistik yang harus dipenuhi pada analisis regresi linear berganda yang berbasis metode kuadrat terkecil. Uji asumsi yang dilakukan pada model regresi adalah

1. Uji Asumsi Normalitas

Analisis regresi linear mengasumsikan bahwa residual e_i berdistribusi normal. Pada regresi linear klasik diasumsikan bahwa setiap e_i didistribusikan secara random dengan $e_i \sim N(0, \sigma^2)$ (Gujarati, 2004).

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

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

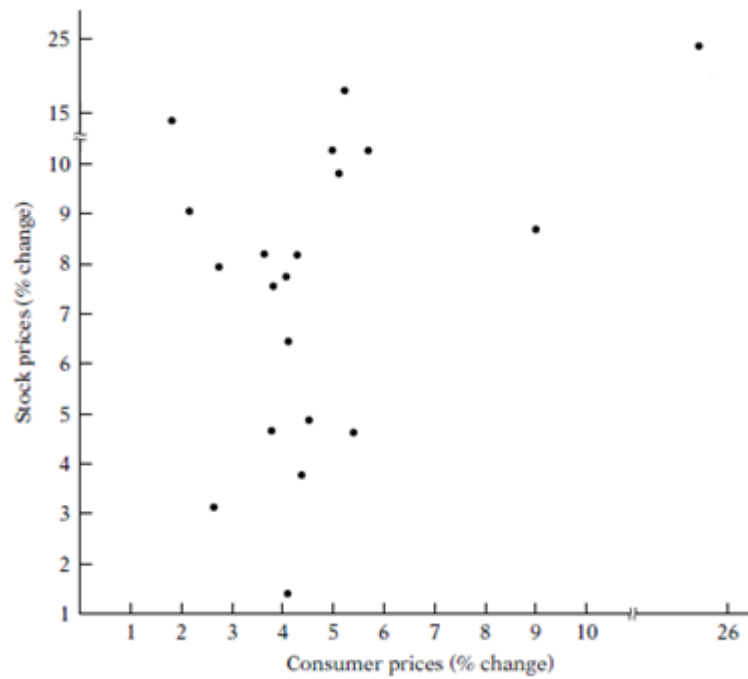
dengan $F_0 X_i$ merupakan fungsi distribusi kumulatif dari distribusi teoritis dibawah H_0 . $S_n X_i$ merupakan distribusi frekuensi kumulatif dari observasi sebanyak n . H_0 merupakan residu yang berdistribusi normal. Selanjutnya nilai D ini dibandingkan dengan nilai D kritis dengan signifikansi α pada tabel *Kolmogorov-Smirnov*. Apabila nilai $D > D_{tabel}$, maka asumsi normalitas dipenuhi.

2. Uji Asumsi Homoskedastisitas

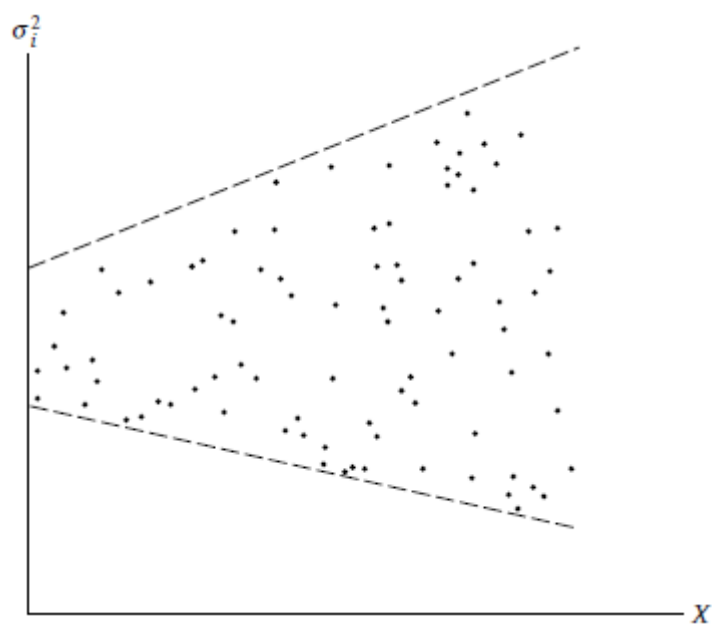
Salah satu asumsi penting dalam analisis regresi adalah variansi residu (e_i) pada setiap variabel adalah homoskedastisitas (Gujarati, 2004). Tujuan dari pengujian ini adalah untuk mengetahui apakah variansi pada tiap residu e_i konstan. Jika variansi pada tiap residu e_i berbeda disebut heteroskedastisitas. Asumsi ini dapat ditulis sebagai berikut

$$Var\ e_i = \sigma^2, \quad i = 1, 2, \dots, n$$

Salah satu cara untuk menguji kesamaan variansi yaitu dengan melihat pola sebaran residu (e_i) terhadap nilai estimasi y . Jika sebaran residu bersifat acak (tidak membentuk pola tertentu), maka dikatakan bahwa variansi sisaan homogen (Draper & Smith, 1981). Penjelasan tersebut dapat terlihat pada gambar 2.1 berikut:



(a.) Homoskedastisitas



(b.) Heterokedastisitas

Gambar 2.1 *scatter-plot* Uji Homoskedastisitas dan Heterokedastisitas
 Sumber: (Gujarati, 2004)

Dari gambar 2.1(a.) terlihat bahwa pola sebaran data menyebar secara merata dan tidak membentuk suatu pola tertentu sehingga data bersifat

homoskedastisitas. Sedangkan pada gambar 2.1 (b.) terlihat bahwa pola sebaran data membentuk pola tertentu sehingga data bersifat heterokedastisitas.

Salah satu cara untuk mendeteksi heteroskedasitas adalah dengan pengujian korelasi *rank spearman* yang didefinisikan sebagai berikut (Gujarati, 2004):

$$r_s = 1 - 6 \frac{d_i^2}{n(n^2 - 1)}$$

dengan d_i merupakan selisih antara masing-masing *rank* variabel independen dengan variabel dependen dan n merupakan banyaknya data yang di *rank*.

Tahapan-tahapan dalam mendeteksi heteroskedasitas adalah sebagai berikut (Gujarati, 2004):

- a. Melakukan analisis regresi dengan menggunakan metode kuadrat terkecil (MKT) untuk menghitung e_i melalui nilai $y_i - \hat{y}_i$.
- b. Mengabsolutkan nilai e_i , kemudian merangking nilai absolut e_i dan X_i sesuai dengan urutan yang meningkat atau menurun dan menghitung koefisien *rank* korelasi *spearman* yang telah diberikan sebelumnya.
- c. Mengasumsikan bahwa koefisien *rank* korelasi populasi ρ_s adalah nol dan $n > 8$, signifikan dari r_s dapat diuji dengan pengujian t sebagai berikut:

$$\text{i. } H_0 : r_s = 0$$

$$H_1 : r_s \neq 0$$

- ii. Pilih $\alpha = 0,05$
- iii. Daerah kritis: H_0 ditolak jika $t_{hitung} > \alpha = 0,05$
- iv. Statistik uji

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

dengan derajat kebebasannya adalah $n - 2$, dengan r_s merupakan nilai koefisien korelasi *spearman* dan n merupakan banyaknya sampel. Uji signifikansi t di atas berfungsi apabila peneliti ingin mencari makna hubungan variabel independen terhadap variabel dependen.

- v. Kriteria keputusan

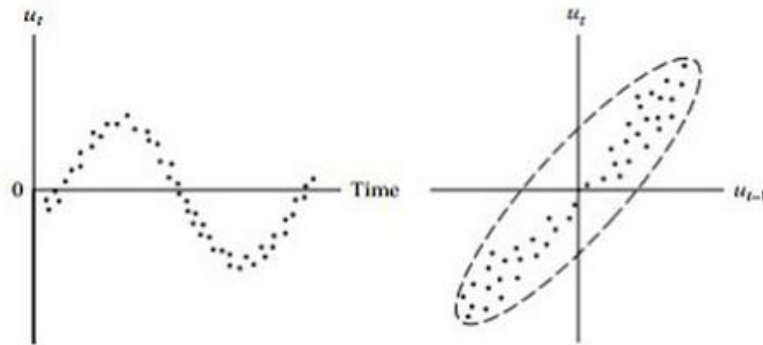
Jika H_0 maka asumsi heteroskedastisitas dipenuhi.

3. Uji Asumsi Non Autokorelasi

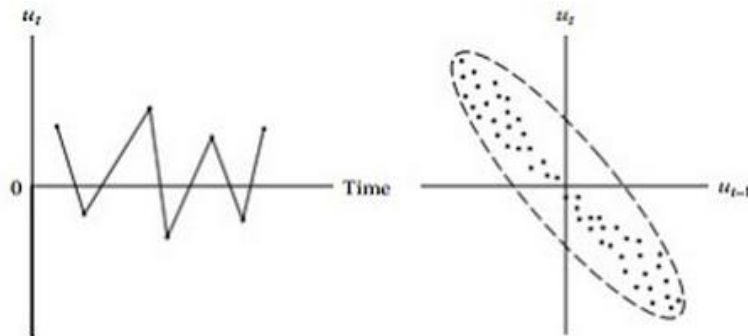
Salah satu asumsi dari regresi linear adalah bahwa tidak adanya autokorelasi antara serangkaian pengamatan yang diurutkan menurut waktu. Pendeteksian autokorelasi dapat dideteksi secara grafis yaitu dengan melihat *scatter-plot* residu terhadap urutan waktu. Jika sebaran residu terhadap urutan waktu tidak membentuk pola tertentu atau bersifat acak maka dapat disimpulkan tidak ada autokorelasi antar residu.

Untuk mendeteksi autokorelasi dapat menggunakan metode grafik, karena metode ini paling sederhana sekaligus merupakan langkah awal untuk mendeteksi adanya autokorelasi. Sesuai dengan definisinya, metode ini membandingkan antara residu dengan variabel X . Selain itu, dengan membandingkan antara residu ke- t dengan residu ke- $(t-1)$. Suatu grafik

mengindikasikan adanya autokorelasi dapat dilihat dari polanya. Suatu grafik dikatakan mengandung autokorelasi ketika terdapat pola antara residu dengan waktu atau antara residu ke- t sampai ke- $(t-1)$.



(a) Autokorelasi Positif



(b) Autokorelasi Negatif

Gambar 2.2 Grafik Autokorelasi Positif dan Autokorelasi Negatif

Pada gambar 2.2 dapat dilihat bahwa grafik membentuk pola siklus sehingga diindikasikan terdapat autokorelasi. Hal ini juga didukung dengan grafik antara residu ke- t dengan residu ke- $(t-1)$ yang menunjukkan adanya hubungan linear. Serta pada gambar 2.2 di atas terdapatnya autokorelasi positif dan negatif, dimana autokorelasi positif terlihat pada gambar 2.2 (a), sedangkan autokorelasi negatif terlihat pada gambar 2.2 (b) (Gujarati, 2004).

Pengujian autokorelasi secara empiris dilakukan dengan menggunakan uji *Durbin-Watson* (Gujarati, 2004). Adapun langkah-langkahnya adalah

- a. Melakukan perhitungan MKT untuk memperoleh nilai e_i ,
- b. Mencari besarnya nilai d

$$d = \frac{\sum_{t=2}^n e_t - e_{t-1}^2}{\sum_{t=2}^n e_t^2}$$

- c. Untuk ukuran sampel n dan $k = p - 1$, dengan p merupakan banyaknya parameter sehingga diperoleh nilai kritis d_L dan d_U ,
- d. Untuk statistik d dari *Durbin-Watson* dapat dilihat pada tabel di lampiran 19.

Pengambilan keputusan ada tidaknya autokorelasi adalah (Gujarati, 2004):

- a. Bila nilai DW terletak antara batas atas atau *upper bound* d_U dan $(4 - d_U)$, maka koefisien autokorelasi sama dengan nol, berarti tidak ada autokorelasi.
- b. Bila nilai DW lebih rendah daripada batas bawah atau *lower bound* (d_L), maka koefisien autokorelasi lebih besar daripada nol, berarti ada autokorelasi positif.
- c. Bila nilai DW lebih besar daripada $(4 - d_L)$, maka koefisien autokorelasi lebih kecil daripada nol, berarti ada autokorelasi negatif.
- d. Bila nilai DW terletak antara batas atas (d_U) dan batas bawah (d_L) dan jika nilai DW terletak antara $(4 - d_U)$ dan $(4 - d_L)$, maka hasilnya tidak dapat disimpulkan.

4. Uji Asumsi Non Multikolinearitas

Kolinearitas terjadi karena terdapat korelasi yang cukup tinggi di antara variabel independen. VIF (*Variance Inflation Factor*) merupakan

salah satu cara untuk mengukur besarnya kolinearitas dan didefinisikan sebagai berikut (Montgomery, Peck, & Vining, 2006):

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

dengan R_j^2 merupakan koefisien determinasi yang dihasilkan dari regresi variabel independen. Nilai VIF menjadi semakin besar jika terdapat korelasi yang semakin besar diantara variabel independen. Jika nilai VIF lebih dari 10, maka multikolinearitas memberikan pengaruh yang serius pada estimasi metode kuadrat terkecil (Hair, Black, Babin, & Anderson, 2010).

D. *Outlier*

Menurut (Sheskin, 2004) *outlier* merupakan suatu observasi pada sekumpulan data yang tidak konsisten terhadap kumpulan data keseluruhan. Menurut (Kleinbum, Kupper, Nizam, & Keith, 2008), definisi dari *outlier* adalah sesuatu yang langka atau observasi yang tidak biasa yang muncul pada salah satu titik esktrim dari sebagian besar data. Demikian pula menurut (Hampel, Ronchetto, Rousseeuw, & Stahel, 1986), definisi *outlier* adalah data yang tidak mengikuti pola umum dalam model regresi yang dihasilkan, atau tidak mengikuti pola data secara keseluruhan. Munculnya *outlier* pada data disebabkan oleh beberapa kemungkinan, yaitu terdapatnya kesalahan prosedur dalam memasukkan data; kesalahan dalam pengukuran atau analisis; dan dikarenakan adanya keadaan yang benar-benar khusus, seperti pandangan responden terhadap sesuatu yang menyimpang dikarenakan adanya suatu alasan yang tidak diketahui oleh peneliti sendiri.

Outlier sangat berpengaruh dalam proses analisis data, yaitu salah satunya terhadap nilai *mean* dan standar deviasi. Oleh karena itu, keberadaan *outlier* dalam suatu data harus diatasi. Dalam kaitannya dengan analisis regresi, *outlier* dapat menyebabkan hal-hal berikut (Soemartini, 2007):

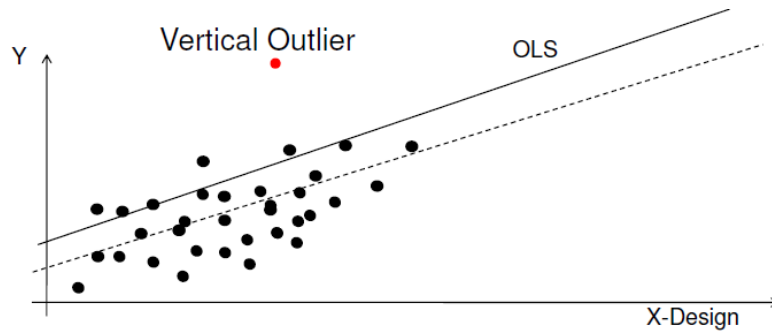
1. Residu yang besar dari model yang terbentuk atau $E(e_i) \neq 0$
2. Varians pada data menjadi lebih besar
3. Interval data memiliki rentang yang lebar

Penolakan begitu saja terhadap suatu *outlier* pada data bukanlah prosedur yang bijaksana, karena adakalanya data *outlier* timbul dari kombinasi keadaan yang tidak biasa yang mungkin saja sangat penting dan perlu diselidiki lebih lanjut. Data *outlier* dapat merupakan suatu pengamatan yang berpengaruh, artinya pengamatan yang dapat mempengaruhi hasil estimasi koefisien regresi. Oleh karena itu tindakan membuang pengamatan berpengaruh akan mengubah secara berarti persamaan regresi serta kesimpulannya (Draper & Smith, 1981).

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

a. *Vertical outlier*

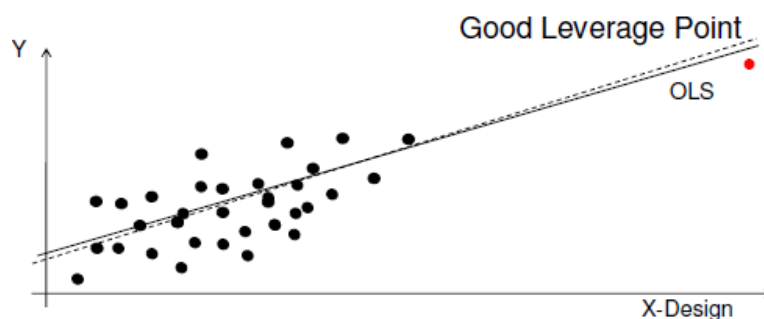
Merupakan suatu titik yang menjadi *outlier* karena memiliki koordinat *y* yang ekstrim. *Vertical outlier* dapat dijelaskan berdasarkan pada gambar 2.2 berikut



Gambar 2.3 Vertical Outlier
 Sumber: (Verardi, 2008)

b. *Good leverage point*

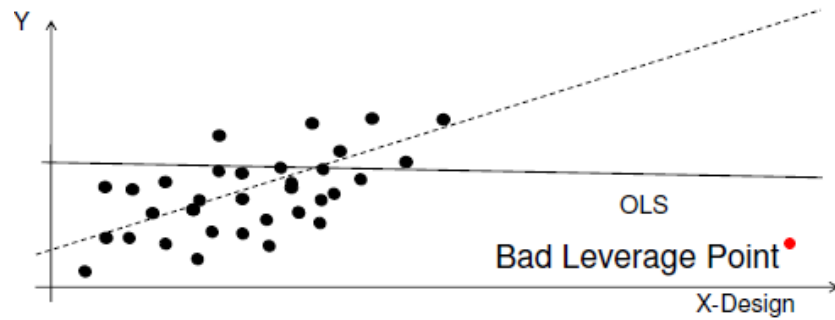
Merupakan suatu titik yang menjadi *outlier* pada variabel independen tetapi terletak dekat dengan garis linear, yang berarti bahwa observasi x_i, y_i apabila x_i menjauh tetapi y_i cocok dengan garis linear. *Good leverage* ini tidak berpengaruh terhadap estimasi kuadrat terkecil, tetapi berpengaruh terhadap inferensi statistik karena dapat meningkatkan estimasi standar *error*. *Good leverage point* dapat dijelaskan berdasarkan pada gambar 2.3 berikut



Gambar 2.4 Good Leverage Point
 Sumber: (Verardi, 2008)

c. *Bad leverage point*

Merupakan suatu titik yang menjadi *outlier* pada variabel independen tetapi terletak jauh dengan garis linear. *Bad leverage* ini berpengaruh signifikan terhadap estimasi kuadrat terkecil. *Bad leverage point* dapat dijelaskan berdasarkan pada gambar 2.4 berikut



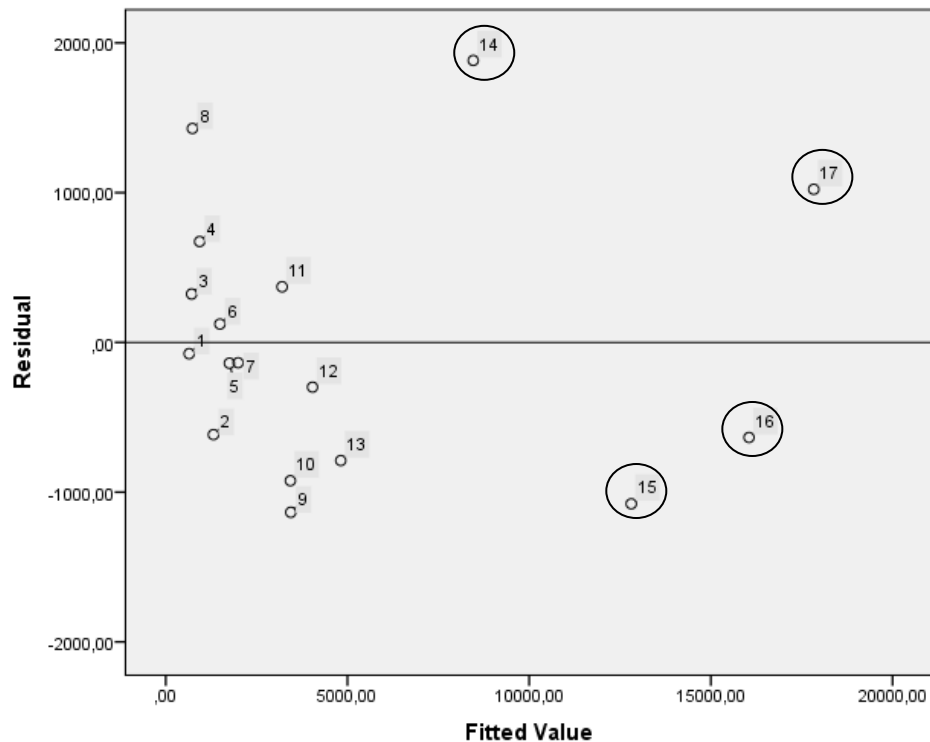
Gambar 2.5 Bad Leverage Point
 Sumber: (Verardi, 2008)

Metode yang digunakan untuk mengidentifikasi adanya *outlier* yang berpengaruh dalam koefisien regresi adalah sebagai berikut:

1. Diagram Pencar (*Scatter Plot*)

Keuntungan dari metode ini adalah mudah untuk dipahami karena menampilkan data secara grafis dan tanpa melibatkan perhitungan yang rumit. Sedangkan kelemahan pada metode ini adalah keputusan yang memperlihatkan data yang merupakan *outlier* atau bukan hanya tergantung pada kebijakan peneliti, karena hanya mengandalkan visualisasi melalui gambar.

Untuk melihat apakah terdapat *outlier* pada data observasi dapat dilakukan dengan memplotkan antara nilai residu (e_i) dengan nilai prediksi $y(y_i)$ seperti pada gambar 2.6 berikut



Gambar 2.6 Contoh *scatter-plot* antara residu (e_i) dengan nilai prediksi $y(y)$

Dari contoh di atas terlihat bahwa observasi ke-14, ke-15, ke-16 dan ke-17 merupakan data observasi yang mengindikasikan adanya *outlier*, karena keempat titik tersebut berada jauh sekumpulan data yang lainnya.

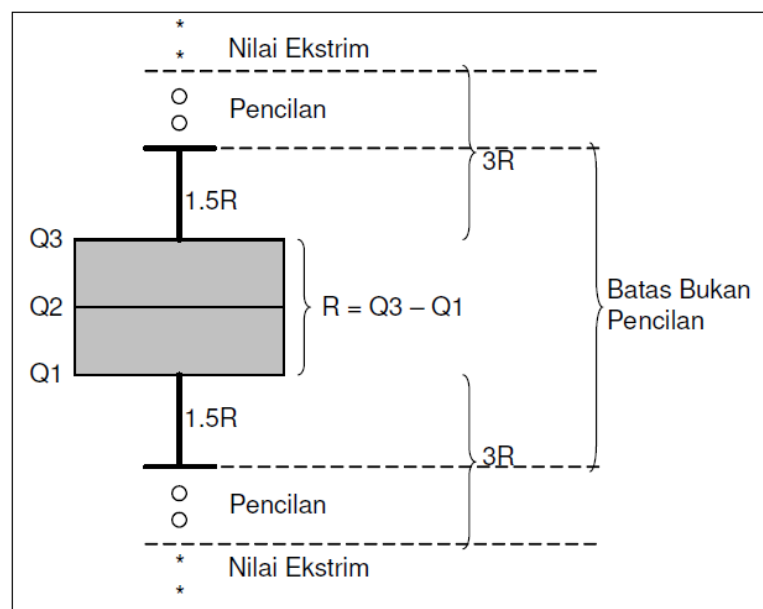
Selain menggunakan metode grafis *scatter-plot* di atas, jika model regresi telah didapatkan kemudian memplotkan antara residu (e_i) dengan nilai prediksi $y(y)$. Jika terdapat satu atau beberapa data yang terletak jauh dari pola kumpulan data keseluruhan, maka hal ini mengindikasikan adanya *outlier* pada data.

Kelemahan dari metode ini adalah keputusan suatu data merupakan *outlier* sangat bergantung pada kebijakan peneliti, karena hanya mengandalkan visualisasi grafis. Untuk meminimumkan kesalahan teknis, maka pendeteksian adanya *outlier* pada data perlu dilakukan melalui perhitungan statistis yang akan dijelaskan pada bagian selanjutnya.

2. *Boxplot*

Metode ini merupakan yang paling umum yakni dengan mempergunakan nilai kuartil dari 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 dengan kuartil yang kurang dari $1,5 \times IQR$ terhadap kuartil 1 dan nilai dengan kuartil yang lebih dari $1,5 \times IQR$ terhadap kuartil 3.



Gambar 2.7 Skema Identifikasi *Outlier* Menggunakan IQR atau *boxplot*

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

Metode yang digunakan untuk mengidentifikasi adanya *outlier* yang berpengaruh dalam koefisien regresi adalah residu *Jackknife*. Residu *Jackknife* merupakan residu yang digunakan untuk mendeteksi keberadaan *outlier*. Definisi *Jackknife* (atau disebut juga dengan *externally studentized* atau *crossvalidated*)

residual atau biasa juga disebut sebagai *R-student*, yang dilambangkan dengan t_i adalah (Faraway, 2002):

$$t_i = \frac{e_i}{\sigma_i \sqrt{1-h_{ii}}} = \frac{y_i - \hat{y}_i}{\sigma_i \sqrt{1 + x_i^T X_i^{-1} x_i}} \quad (2.13)$$

dengan t_i berdistribusi t_{n-p-1} jika model asumsi terpenuhi dan $e_i \sim N(0, \sigma^2 I)$.

X_i merupakan matriks tanpa baris ke- i dan x_i merupakan matriks baris ke- i . h_{ii}

merupakan elemen diagonal ke- i dari matriks *hat* dengan $h_{ii} = x_i^T X_i^{-1} x_i$

dan e_i merupakan residu ke- i .

Persamaan (2.13) di atas ekuivalen dengan:

$$t_i = s_i \frac{\sigma^2}{\sigma_i^2} = s_i \frac{n-p-1}{n-p-s_i^2}^{\frac{1}{2}} \quad (2.14)$$

bahwa t_i merupakan fungsi monoton dari s_i , s_i merupakan *standardized residual*.

Dengan $\sigma_{(i)}^2$ merupakan estimasi dari σ^2 berdasarkan pada $(n-1)$ observasi yang tersisa setelah observasi ke- i dihapus. Menurut (Chatterjee & Hadi, 1986) $\sigma_{(i)}^2$ adalah

$$\sigma_{(i)}^2 = \frac{x_i^T (I - H_{ii}) x_i}{n-p-1} = \frac{n-p}{n-p-1} \sigma^2 - \frac{e_i^2}{(n-p-1)(1-h_{ii})}, \quad (2.15)$$

Sementara σ^2 mempunyai derajat kebebasannya $(n-p)$. Sedangkan $\sigma_{(i)}^2$ mempunyai derajat kebebasannya $[(n-p)-1]$ karena observasi ke- i dihapus.

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

E. Regresi *Robust*

Regresi *robust* diperkenalkan oleh Andrews (1972) merupakan metode regresi yang sering digunakan ketika terdapat beberapa *outlier* yang berpengaruh pada model dan tanpa menghapus data yang teridentifikasi adanya *outlier* tersebut. Metode ini merupakan alat penting untuk menganalisis data yang dipengaruhi oleh *outlier* sehingga dihasilkan model yang *robust* atau *resistance* terhadap *outlier*. Suatu estimator yang *robust* adalah relatif tidak berpengaruh oleh adanya perubahan besar pada bagian kecil data atau perubahan kecil pada bagian besar data (Widodo, Guritno, & Haryatmi, 2013).

Menurut (Chen, 2002) metode-metode estimasi dalam regresi *robust* diantaranya adalah:

1. Estimasi-M (*Maximum likelihood type*) yang diperkenalkan oleh Huber (1973) merupakan metode yang sederhana baik dalam perhitungan maupun secara teoritis.
2. Estimasi-LMS (*Least Median Squares*) merupakan metode yang diperkenalkan oleh Hampel (1975). Metode ini memiliki nilai *breakdown point* hingga 50%, namun memiliki efisiensi yang sangat rendah. *Breakdown point* merupakan ukuran umum dari data *outlier* yang dapat ditangani sebelum observasi mempengaruhi model prediksi (Rousseeuw, 1984).
3. Estimasi-LTS (*Least Trimmed Squares*) merupakan metode yang memiliki nilai *breakdown point* tinggi yang diperkenalkan oleh Rousseeuw (1984).

4. Estimasi-S (*Scale*) juga merupakan metode dengan memiliki nilai *breakdown point* tinggi yang diperkenalkan oleh Rousseeuw dan Yohai (1984). Meski memiliki nilai *breakdown point* yang sama dengan estimasi-LTS, namun estimasi-S memiliki efisiensi yang lebih tinggi dibanding estimasi-LTS.
5. 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.

F. Koefisien Determinasi (R^2)

Koefisien determinasi atau biasa disebut dengan R^2 merupakan salah satu ukuran yang sederhana dan sering digunakan untuk menguji kualitas suatu persamaan garis regresi (Gujarati, 2004). Nilai koefisien determinasi memberikan gambaran tentang kesesuaian variabel independen dalam memprediksi variabel dependen. Semakin besar nilai R^2 , maka semakin besar variasi variabel dependen (Y) yang dijelaskan oleh variasi variabel-variabel independen (X). Sebaliknya, semakin kecil nilai R^2 , maka semakin kecil variasi variabel dependen yang dapat dijelaskan oleh variasi variabel independen.

Sifat dari koefisien determinasi adalah (Gujarati, 2004):

- a. R^2 merupakan besaran yang non-negatif
- b. Batasnya adalah $0 \leq R^2 \leq 1$

Apabila nilai koefisien determinasi semakin besar atau mendekati 1, menunjukkan adanya hubungan yang sempurna. Sedangkan apabila nilai koefisien

determinasinya sebesar 0 menunjukkan bahwa tidak terdapat hubungan antara variabel independen dengan variabel dependen.

G. Breakdown point

Breakdown point merupakan fraksi terkecil dari data yang terkontaminasi *outlier* yang dapat menyebabkan estimator tidak berfungsi (Montgomery, Peck, & Vining, 2006). *Breakdown point* adalah jumlah observasi minimal yang dapat menggantikan sejumlah observasi awal yang berakibat pada nilai estimator yang dihasilkan sangat berbeda dari estimator sebenarnya. Dengan kata lain, *breakdown point* sebagai suatu ukuran kerobustan dari suatu estimator. *Breakdown point* merupakan ukuran umum proporsi dari *outlier* yang dapat ditangani sebelum observasi tersebut mempengaruhi model prediksi. Semakin besar nilai persentase dari *breakdown point* pada suatu estimator, maka estimator tersebut semakin *robust* (Sahari, 2012). Regresi *robust* yang mempunyai *breakdown point* adalah regresi *robust* dengan metode estimasi-S, LTS, LMS, dan MM. Estimasi-S dapat digunakan untuk mengatasi masalah *outlier* dengan proporsi hingga 50% serta digunakan ketika variabel dependen dan variabel independen terdapat *outlier*.

H. Residu Robust dan Jarak Robust (Robust Distance)

Residu *robust* dan jarak *robust* memiliki banyak keuntungan. Pertama, *robust residual* (RD) menunjukkan adanya *outlier* pada regresi lebih baik dibandingkan dengan metode kuadrat terkecil. Jarak *Mahalanobis* didefinisikan sebagai (SAS Institute, 2004):

$$MD X_i = \sqrt{(X_i - X)^T C(X)^{-1} (X_i - X)}$$

dimana $X = \frac{1}{n} \sum_{i=1}^n X_i$ dan $C = \frac{1}{n-1} \sum_{i=1}^n (X_i - X)^T (X_i - X)$.

$X_i = (X_{i1}, \dots, X_{i(p-1)})^T$ disini bukan merupakan variabel konstan. Hubungan antara jarak *Mahalanobis* $MD X_i$ dan matriks *hat* $H = h_{ii} = X(X^T X)^{-1} X^T$ adalah

$$h_{ii} = \frac{1}{n-1} MD_i^2 + \frac{1}{n}$$

Jarak *robust* didefinisikan sebagai

$$MD X_i = \sqrt{(X_i - T(X))^T C(X)^{-1} (X_i - T(X))}$$

dimana $T(X)$ dan $C(X)$ merupakan vektor rata-rata *robust* dan matriks kovarians *robust*. Jarak *Mahalanobis* dan jarak *robust* digunakan untuk mendiagnosa titik *leverage*. Akan tetapi jarak *robust* lebih dapat dipercaya untuk mendiagnosa titik *leverage* daripada jarak *Mahalanobis* atau matriks *hat*.

Leverage adalah observasi dengan nilai ekstrim pada variabel independen atau ukuran jauhnya variabel independen menyimpang dari rata-ratanya. Titik *leverage* didefinisikan sebagai

$$Leverage = \begin{cases} 0 & \text{jika } RD X_i \leq C_p \\ 1 & \text{untuk lainnya} \end{cases}$$

dengan nilai *cutoff* $C_p = \sqrt{\chi_{p;1-\alpha}^2}$.

Pada regresi linear, *outlier* merupakan observasi dengan nilai residu yang besar, artinya pada observasi tersebut nilai variabel independen tidak sesuai dengan nilai yang diberikan oleh variabel dependen. Titik *outlier* dapat dideteksi dengan menggunakan nilai residunya, dan didefinisikan sebagai

$$Outlier = \begin{cases} 0 & e_i \leq 3 \\ 1 & \text{untuk lainnya} \end{cases}$$

dimana $e_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})$, $i = 1, \dots, k$ berdasarkan pada estimasi regresi *robust* digunakan untuk mendeteksi *vertical outlier*.