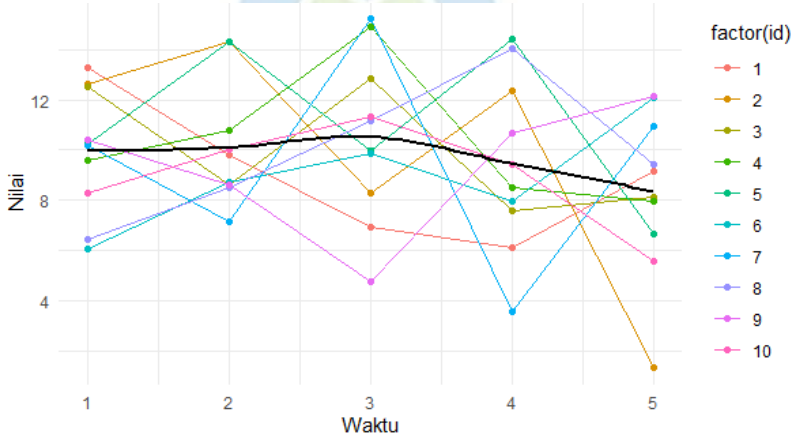
BAB II

LANDASAN TEORI

2.1 Data Longitudinal

Data longitudinal adalah kombinasi data *time series* dan data *cross section*. Data *time series* merupakan data pada selang waktu tertentu dari sebuah subjek, sementara data *cross section* merupakan data dari satu atau lebih subjek penelitian pada periode waktu yang sama.



Gambar 2.1 Plot Data longitudinal dengan Tren Umum

Data longitudinal sering disebut sebagai data grup (*pooled* data), kombinasi siklik, data microlongitudinal, dll. Keuntungan dari data longitudinal meliputi [10] :

1) Data longitudinal dapat memberikan lebih banyak data dan informasi yang lebih komprehensif. Ini memberi lebih banyak derajat kebebasan (df), sehingga hasil perkiraannya lebih baik.
2) Penggabungan informasi data deret waktu dan data *cross section* mampu memecahkan masalah yang disebabkan oleh penghilangan variabel.
3) Data longitudinal dapat menyebabkan kolinearitas yang buruk antar variabel.
4) Data longitudinal cocok untuk pendeteksian dan pengukuran efek yang tidak mungkin dilakukan pada data deret waktu murni.

5) Kemampuan untuk pengujian dan dibangunnya suatu model perilaku yang lebih rumit.
6) Data longitudinal mampu membuat bias lebih minimal karena ditimbulkan oleh agregasi individual dan ada banyaknya observasi.

2.2 Model Analisis Regresi

A. Analisis Regresi Linear

Analisis regresi linear merupakan sebuah bidang yang berkaitan dengan keterikatan antar variabel terikat (*response variable*) Y dengan variabel bebas (*predictor variable*) $X_1, X_2, \ldots, X_k$ yang saling berhubungan dan dimodelkan dalam persamaan matematika pada suatu data. Adapun nilai dari $k = 1$ maka regresinya adalah regresi sederhana (*simple regression*), namun jika $k > 1$ maka regresinya adalah regresi berganda (*multiple regression*). Proses pengujian persamaan analisis regresi ada empat langkah yaitu :

1. Penentuan estimasi parameter yang ada di persamaan regresi,
2. Pengujian normalitas pada data,
3. Pengujian asumsi homoskedatisitas,
4. Pengujian asumsi multikolineritas.

Koefisien a dan b adalah koefisien regresi yang bisa dicari menggunakan persamaan berikut :

$$a = \sum_{i=1}^{n} Y_i/n \; - b \sum_{i=1}^{n} X_i/n \tag{2.1}$$

$$b = \frac{\sum (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \tag{2.2}$$

Model regresi linear sederhana yang memuat satu k variabel prediktor didefinisikan dengan (*simple regression*) [1] :

$$Y = a + bX \tag{2.3}$$

Di mana a dan b didefinisikan pada persamaan (2.1) dan (2.2).

B. Analisis Regresi Linear Berganda

Model regresi yang memuat lebih dari satu k variabel prediktor ditulis sebagai berikut (*multiple regression*) [1] :

$$Y = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \; ; \; i = 1, 2, \ldots, n \tag{2.4}$$

Oleh karena itu, i menunjukkan observasi sehingga akan ada n perbedaan :

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + \varepsilon_1$$
$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + \varepsilon_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + \varepsilon_i$$

2.3 Regresi Data Longitudinal

Analisis regresi bisa diterapkan dalam data longitudinal, yang kemudian dikembangkan model persamaannya menjadi [10] :

$$Y_{it} = \alpha + \beta' X_{it} + \varepsilon_{it} \quad ; \; i = 1, \ldots, N \; ; \; t = 1, \ldots, T \tag{2.5}$$

atau dapat didefinisikan juga sebagai berikut :

$$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \beta_n X_{nit} + e_{it}$$

Estimasi atau pendugaan model regresi data longitudinal bisa digunakan pada berbagai pendekatan. Dua di antaranya adalah model efek umum dan model efek acak. Keuntungan dari penggunaan data longitudinal adalah untuk mengontrol variabel-variabel pada observasi yang dihilangkan atau variabel yang berubah dari waktu ke waktu. Terdapat beberapa model regresi data longitudinal yang berkaitan dengan *slope* konstan dan *intercept* yang bervariasi.

Model regresi data longitudinal pada satu unit yang mempengaruhi disebut dengan model komponen satu arah, sedangkan oleh kedua unit disebut dengan model komponen dua arah. Sehingga ada dua pendekatan yang bisa digunakan untuk menaksir model data longitudinal yaitu model tanpa adanya pengaruh individu (*commont effect model*) dan model dengan adanya pengaruh individu (*fixed effect model* dan *random effect model*).

Berikut beberapa alternatif model yang bisa digunakan untuk menyelesaikan model regresi data longitudinal yaitu [11]

Model 1 : Semua koefisien *intercept* dan *slope* koefisien konstan

$$Y_{it} = \beta_1 + \sum_{k=2}^{K} \beta_k X_{kit} + \varepsilon_{it} \tag{2.6}$$

Model 2 : *Slope* koefisien konstan, namun *intercept* berbeda sebab ada perbedaan unit *cross section*

$$Y_{it} = \beta_{1i} + \sum_{k=2}^{K} \beta_k X_{kit} + \varepsilon_{it} \tag{2.7}$$

Model 3 : *Slope* koefisien konstan, namun *intercept* berbeda sebab ada unit *cross section* dan adanya perubahan waktu

$$Y_{it} = \beta_{1it} + \sum_{k=2}^{K} \beta_k X_{kit} + \varepsilon_{it} \tag{2.8}$$

Model 4 : *Intercept* dan *slope* koefisien berbeda sebab adanya perbedaan unit *cross section*

$$Y_{it} = \beta_{1i} + \sum_{k=2}^{K} \beta_{ki} X_{kit} + \varepsilon_{it} \tag{2.9}$$

Model 5 : *Intercept* dan *slope* koefisien berbeda sebab ada perbedaan unit *cross section* dan adanya perubahan waktu

$$Y_{it} = \beta_{1it} + \sum_{k=2}^{K} \beta_{kit} X_{kit} + \varepsilon_{it} \tag{2.10}$$

A. *Commont Effect Model* (CEM)

Model ini merupakan model tanpa pengaruh individu (*commont effect*) lalu estimasi semua data *time series* dan *cross section* digabungkan (*pooled*) untuk estimasi parameternya dengan estimasi OLS (*Ordinary Least Square*). Metode OLS adalah metode pendugaan nilai parameter dengan menggunakan persamaan regresi linear. Persamaan *Commont Effect Model* didefinisikan [10] :

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it} \quad ; \; i = 1, \dots, N \; ; \; t = 1, \dots, T$$

Jika nilai $cov(\varepsilon_{it}, \varepsilon_{jt}) = 0; cov(\varepsilon_{it}, \varepsilon_{it-1}) = 0; E(\varepsilon_{it}) = 0; dan\ var(\varepsilon_{it}) = \sigma^2$, dapat diestimasikan dengan model yang waktunya dipisahkan sehingga ada T regresi dan N observasi. Dapat didefinisikan sebagai berikut :

$$\begin{matrix} Y_{i1} = \alpha + \beta X_{i1} + \varepsilon_{i1} \\ Y_{i2} = \alpha + \beta X_{i2} + \varepsilon_{i2} \\ \vdots \\ Y_{iT} = \alpha + \beta X_{iT} + \varepsilon_{iT} \end{matrix} \quad ; i = 1,2, \dots, N$$

Model yang digunakan juga dapat dipisahkan dengan cara mengestimasi cross section-nya yang nantinya diperoleh N regresi pada masing-masing T observasi. Bisa didefinisikan sebagai berikut :

$$\begin{matrix} i = 1; \; Y_{1t} = \alpha + \beta X_{1t} + \varepsilon_{1t} \\ i = 2; \; Y_{2t} = \alpha + \beta X_{2t} + \varepsilon_{2t} \\ \vdots \\ i = N; \; Y_{Nt} = \alpha + \beta X_{Nt} + \varepsilon_{Nt} \end{matrix} \quad ; t = 1,2, \dots, T$$

B. *Fixed Effect Model* (FEM)

Fixed Effect Model merupakan estimasi parameter regresi dengan cara menambahkan variabel *dummy* maka metode ini disebut sebagai *Least Square Dummy Variable* model. Persamaan regresi untuk *Fixed Effect Model* didefinisikan sebagai [10] :

$$Y_{it} = \beta_{0i} + \sum_{k=1}^{p} \beta_k X_{kit} + \varepsilon_{it} \qquad (2.11)$$

Pada model *Fixed Effect* memiliki asumsi koefisien *slope* konstan nilainya namun *intercept*-nya bervariasi sepanjang individunya dan bernilai tidak konstan. *Fixed effect* awalnya dibentuk dari kenyataan bahwa *intercept* β_{0i} pada setiap individu berbeda akan tetapi nilai *intercept* antar waktunya sama (*time invarian*), sementara itu nilai *slope* β_k konstan baik antar individu maupun antar waktunya. Untuk mengestimasi parameter regresinya menggunakan variabel *dummy* yang bisa

mendefinisikan adanya *intercept* antar individu yang berbeda. Maka persamaannya dapat ditulis [10]:

$$Y_{it} = \beta_{0i} D_{it} + \sum_{k=1}^{p} \beta_k X_{kit} + \varepsilon_{it} \tag{2.12}$$

di mana $D_{it} = 1$ untuk objek yang pertama sedangkan untuk $D_{it} = 0$ untuk objek berikutnya.

Parameter yang digunakan untuk estimasi pada *Fixed Effect Model* (FEM) pada data longitudinal digunakan metode *Ordinary Least Square* (OLS). Karena adanya penggunakan variabel dummy maka pendekatannya disebut dengan *Least Square Dummy Variable* (LSDV). Pemasukkan unsur *time effect* juga bisa digunakan untuk melihat bahwa *intercept* individunya tidak konstan lagi. *Time effect* ini mempengaruhi penambahan variabel *dummy* untuk waktu yang dihitung.

C. *Random Effect Model* (REM)

Random Effect Model (REM) mempunyai perilaku berbeda antar individunya dan waktu yang digunakan pada model serta nilai kesalahannya (*error*). Karena adanya dua komponen yang digunakan pada error yang dibentuk, yaitu waktu dan individu (observasi), maka kesalahan acak (*random error*) perlu dijelaskan lebih rinci menjadi kesalahan (*error*) untuk komponen waktu dan kesalahan gabungan. Oleh karena itu, persamaan *Random Effect Model* (REM) dimodelkan sebagai berikut [10] :

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \varepsilon_{it} \tag{2.13}$$

2.4 Estimasi Parameter *Ordinary Least Square*

Draper dan Smith (1992) mengklaim bahwa metode kuadrat terkecil pertama kali ditemukan secara terpisah oleh Carl Friedrich Gauss dari Jerman (1777-1855) dan Adrien Marie Legendre dari Prancis (1752-1833). Metode kuadrat terkecil juga merupakan metode untuk mengestimasi kesalahan (*error*) kuadrat total dari model regresi yang dibentuk dengan meminimalkan parameter regresi. Misalkan (x_i, y_i), i = 1, 2, …., n, koefisien regresi α dan β ditentukan sehingga [1] :

$$\mathrm{Y_i} = \alpha + \beta \mathrm{X_i} + \varepsilon_\mathrm{i}$$

$$\varepsilon_i = y_i - \alpha - \beta X_i \quad (2.14)$$

$$(\varepsilon_i)^2 = (y_i - \alpha - \beta X_i)^2$$

Dari persamaan yang dijelaskan diatas dapat dihitung koefisien regresinya

$$J = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

Minimumkan fungsi

$$\sum Y_i - n\alpha - \beta \sum X_i = 0$$

$$\sum_{i=1}^{n} Y_i X_i - \alpha \sum_{i=1}^{n} X_i - \beta \sum_{i=1}^{n} {X_i}^2 = 0$$

Taksiran dari α dan β adalah a dan b

Bila didefinisikan $\bar{X} = \sum_{i=1}^{n} X_i/n$ dan $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ maka, diperoleh persamaan a yaitu:

$$\sum Y_i - na - b \sum X_i = 0$$

$$na = \sum_{i=1}^{n} Y_i - b \sum_{i=1}^{n} X_i$$

$$a = \sum_{i=1}^{n} Y_i/n - b \sum_{i=1}^{n} X_i/n$$

$$a = \bar{Y} - b\bar{X}$$

Akibat

$$\sum_{i=1}^{n} Y_i X_i - a \sum_{i=1}^{n} X_i - b \sum_{i=1}^{n} {X_i}^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - (\bar{Y} - b\bar{X}) \sum_{i=1}^{n} X_i - b \sum_{i=1}^{n} {X_i}^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i + b\bar{X} \sum_{i=1}^{n} X_i - b \sum_{i=1}^{n} {X_i}^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i + b \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i - b \sum_{i=1}^{n} {X_i}^2 = 0$$

$$b = \left(\sum_{i=1}^{n} {X_i}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right)^2 \right) = \sum_{i=1}^{n} Y_i X_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i$$

Jadi

$$b = \frac{\sum Y_i X_i - (\sum Y_i)(\sum X_i)/n}{\sum {X_i}^2 - (\sum X_i)^2/n}$$

2.5 Uji Asumsi Model Regresi Data Longitudinal

Yudiatmaja (2013) menyatakan bahwa model regresi data panel dapat dianggap baik jika memenuhi kriteria *BLUE* (*Best, Linear, Unbiased*, dan Estimator). *BLUE* dapat dicapai dengan memenuhi asumsi klasik. Persamaan tidak dapat menghasilkan nilai prediksi yang akurat jika tidak memenuhi kaidah *BLUE*. Meskipun demikian, ini tidak menunjukkan bahwa persamaan tersebut tidak dapat digunakan untuk memprediksi. Agar persamaan tersebut dikategorikan telah memenuhi kaidah *BLUE*, data yang digunakan harus memenuhi beberapa asumsi yang biasa digunakan dalam uji asumsi klasik.

Ada empat uji asumsi klasik yang digunakan yaitu uji normalitas, uji uji heteroskedastisitas, uji autokorelasi, dan uji multikolinearitas. Estimator yang tidak bias dapat dihasilkan hanya dengan menggabungkan keempat uji ini.

A. Uji Normalitas

Uji normalitas digunakan untuk menentukan distribusi residual atau variabel pengganggu dalam model regresi normal. Hasil uji t dan F sebelumnya menunjukkan bahwa nilai residual mengikuti distribusi normal. Jika asumsi ini dilanggar, uji statistik tidak valid [12].

Ada berbagai jenis uji statistik normalitas yang dapat digunakan, seperti *Kolmogorov-Smirnov, Lilliefors, Chi-Square*, dan *Shapiro-Wilk*. Selain itu, pengguna juga bisa memanfaatkan *software* komputer seperti SPSS, *Minitab*, R, dan lainnya. Pada dasarnya, *software* tersebut menjalankan uji statistik *Kolmogorov-Smirnov, Lilliefors, Chi-Square,* dan *Shapiro-Wilk* yang telah diprogram di dalamnya. Pengguna

bisa memilih uji statistik normalitas yang paling sesuai dengan kebutuhannya karena setiap uji memiliki kelebihan dan kelemahan masing-masing.

B. Uji Multikolinearitas

Cara pendeteksian ada atau tidaknya multikolinearitas pada variabel prediktor atau prediktor yaitu dengan menggunakan cara perhitungan nilai koefisien korelasi pada antar variabel prediktor yang ada. Untuk ukuran korelasi yang digunakan adalah [13] :

Tabel 2.1 Interpretasi Koefisien Korelasi

Nilai	Interpretasi
0,00 – 0,199	Sangat Rendah
0,20 – 0,399	Rendah
0,40 – 0,599	Sedang
0,69 – 0,799	Kuat
0,80 – 1,000	Sangat Kuat

C. Uji Homoskedastisitas

Uji homoskedastisitas adalah langkah penting dalam analisis regresi untuk memastikan bahwa asumsi varians residual konstan di seluruh rentang nilai prediktor terpenuhi. Asumsi homoskedastisitas diperlukan agar estimasi parameter regresi tetap tidak bias dan efisien. Homoskedastisitas berarti bahwa varians dari residual atau kesalahan model regresi tidak berubah saat nilai prediktor berubah. Jika varians residual berubah, fenomena ini disebut heteroskedastisitas, yang dapat mengganggu validitas model dan hasil analisis [7]. Metode uji yang digunakan meliputi [14]:

1. Grafik Residual vs. Fitted Values: Melihat pola pada grafik ini dapat mengidentifikasi heteroskedastisitas. Pola yang teratur menunjukkan heteroskedastisitas.
2. Uji Breusch-Pagan : Menguji heteroskedastisitas dengan memperhitungkan residual yang diperkirakan dan variabel independen.

3. Uji White : Uji heteroskedastisitas yang tidak memerlukan asumsi spesifik tentang bentuk heteroskedastisitas.
4. Uji Goldfeld-Quandt : Menguji adanya heteroskedastisitas dengan membandingkan varians residual antara dua subkelompok data.

D. Uji Autokorelasi

Dalam model regresi linear, uji autokorelasi adalah mengukur ada tidaknya kolerasi antara kesalahan pengganggu pada periode t dan kesalahan pengganggu pada periode sebelumnya $t-1$. Tujuan dari uji autokolerasi adalah untuk mengetahuinya. Autokorelasi sering terjadi pada data *time series* karena muncul dari observasi berurutan selama waktu yang sama. Oleh karena itu, model regresi yang tidak mengandung autokorelasi dapat dianggap baik. Uji run dan uji Durbin-Watson dapat digunakan untuk menemukan autokorelasi [12].

Estimator-estrimator ini tidak lagi efektif karena memiliki varian terkecil disebabkan oleh autokorelasi pada metode kuadrat terkecil. Estimasi kuadrat terkecil tidak bias dan linier. Akibatnya, baik interval estimasi maupun uji hipotesis yang didasarkan pada distribusi t dan F tidak dapat digunakan untuk mengevaluasi hasil regresi [12].

2.6 Koefisien Determinasi

Sugiyono (2016) mengatakan bahwa koefisien determinasi (R^2) berkaitan dengan penggunaan kontribusi variabel prediktor (X) terhadap variabel respon (Y) yang diprediksi. Kisaran nilai pada koefisien determinasi (R^2) adalah $0 < R^2 < 1$. Apabila nilai R^2 makin kecil maka semakin kecil pula variabel prediktor dengan variabel responnya saling mempengaruhi. Semakin dekat nilai R^2 dengan 1, maka semakin kuat variabel prediktor (X) terhadap variabel respon (Y) saling mempengaruhi.

Nilai *R-squared* digunakan dalam kasus regresi sederhana, tetapi *fitting R-squared* digunakan dalam kasus regresi berganda [12]. Rumus untuk menentukan R^2 :

$$(Y_i - \bar{Y}) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right) \tag{2.15}$$

variasi akibat sisa

$$JKT = \sum (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 \quad (2.16)$$

$$JKR = \sum \left(\widehat{Y}_i - \overline{Y}\right)^2 = b^2 \sum (X_i - \overline{X})^2 \quad (2.17)$$

$$JKE = \sum \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \quad (2.18)$$

$$R^2 = \frac{\sum \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum (Y_i - \overline{Y})^2} \quad (2.19)$$

2.7 Deteksi *Outlier*

Outlier atau pencilan terjadi karena adanya data yang tidak sesuai dengan pola model umum atau biasanya *residual* (*error*-nya) tiga kali dari nilai standar deviasi-nya atau lebih jauh dari nilai rata-rata residual [16]. *Outlier* juga mampu memberi pengaruh terhadap hasil pengestimasian parameter regresinya, sehingga bisa timbul pelanggaran yang berkaitan dengan asumsi data yang normal. *Outlier* dianalisis regresi menimbulkan sebab di mana sisaan yang dihasilkan bernilai besar berdasarkan dari pembentukan modelnya, data yang beragam pun akan menjadi lebih besar sehingga, akan menimbulkan adanya data yang tidak homogen [17].

Kehadiran data *outlier* dapat membingungkan proses analisis data dan harus dihindari dengan berbagai cara. Dari perspektif analisis regresi, *outlier* dapat menyebabkan :

1. Besar sisa model yang terbentuk, atau $E(\varepsilon) \neq 0$
2. Penyebaran data meningkat.
3. Ada berbagai estimasi interval parameter

Sementara untuk metode deteksi *outlier* dapat dikategorikan menjadi tiga jenis yaitu deteksi *outlier* terawasi, deteksi *outlier* semi terawasi, dan deteksi *outlier* tanpa pengawasan. *Outlier* juga dapat dikategorikan ke dalam empat kelompok berdasarkan penyebabnya yaitu Pengamatan Umum, Poin *Leverage* Baik, *Outlier* Vertikal, dan Poin *Leverage* Buruk [18]. *Outlier* dapat menyebabkan estimasi yang tidak akurat dari

parameter (*slopes* dan *intercept*) dan terkadang menyebabkan efek yang merugikan pada analisis statistik dalam beberapa hal yaitu :

a) Pertama, proses memperbesar estimasi varians residual dan minimalkan nilai pangkat uji statistik untuk menolak hipotesis nol.
b) Kedua, hal tersebut dapat mempengaruhi distribusi normal 1% yang didasarkan pada lokasi *outlier* dalam data. Sebagai contoh, jika ada 1% kemungkinan akan mendapatkan data *outlier* dari distribusi normal; itu artinya 1% dari observasi tersebut harus memiliki 3 standar deviasi dari nilai rata-ratanya.
c) Ketiga, *outlier* dapat memberikan estimasi yang mungkin menjadi masalah substantif.

Kriteria adanya *outlier* dapat digambarkan pada grafik, plot residual, dan beberapa nilai kriteria *outlier* yang ditunjukkan pada tabel di bawah ini [19] :

Tabel 2.2 Metode deteksi *outlier*

Metode Deteksi Outlier	**Cutoff**
Nilai Leverage	$> \frac{2p}{n}$
DfFit	$> 2\sqrt{\frac{p}{n}}$
Studentized Residuals	$> -2 < t_i < 2$
DfBeta	$> \frac{2}{\sqrt{n}}$

2.7.1 Titik *Leverage*

Titik *leverage* terbagi menjadi dua yaitu titik *leverage* baik *(good leverage point)* dan titik *leverage* buruk *(bad leverage point). Leverage point* yang baik adalah *outlier* yang disebabkan oleh variabel prediktornya saja ($X\ outlier$), sedangkan *leverage point* yang buruk adalah *outlier* yang disebabkan oleh variabel respon dan variabel prediktor ($X - Y\ outlier$).

Secara umum, nilai *leverage* mewakili kasus yang diregresi pada *scatterplot* x atau variabel prediktor dan menunjukkan seberapa jauh pengamatannya dari rata-rata

set data variabel prediktor-nya. Titik *leverage* juga diperoleh berdasarkan matriks hat (H) yaitu matriks yang digunakan dalam regresi linier untuk meproyeksikan vector respons Y ke vector prediksi $\hat{Y}$. Matriks ini disebut "hat matriks" karena "menaruh topi" di atas Y untuk menghasilkan $\hat{Y}$ dan ditentukan ukurannya $(n \times n)$[3]

$$H = X(X^T X)^{-1} X^T \tag{2.20}$$

di mana X adalah matriks $n(k+1)$, n adalah jumlah data dan k adalah jumlah variabel prediktor (X). Diagonal H memuat titik leverage. Diagonal untuk kasus ke-i (h_{ii}) adalah nilai pada baris ke-i, kolom ke-i dari H:

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \tag{2.21}$$

di mana x_i adalah vektor baris dari pengamatan ke-i dan X adalah matriks desain dari semua prediktor. Nilai yang paling banyak digunakan ditentukan berdasarkan batasan $h_{ii} > \frac{2(k+1)}{n}$, jika nilai h_{ii} yang melampaui nilai batasnya maka data tersebut terdeteksi sebagai *outlier*. Berdasarkan aturan *thumb* kisaran *leverage* adalah dari 0 sampai 1, nilai yang semakin dekat dengan 1 menunjukkan pengaruh *leverage* yang semakin besar [17].Adapun hubungan antara matriks H dan h_{ii} adalah :

1. Matriks H mengandung *leverage* semua pengamatan dalam model. Setiap diagonal h_{ii} dari H memberikan *leverage* spesifik untuk pengamatan ke-i.
2. Matriks H digunakan untuk menghitung prediksi $\hat{Y}$ dari vektor respon Y yaitu $\hat{Y} = HY$.
3. *Leverage* h_{ii} adalah ukuran individual yang menunjukkan pengaruh relatif dari pengamatan ke-i dalam ruang prediktor.

2.7.2 *Internally Studentized Residuals*

Discrepancy atau dikenal dengan jarak antara nilai estimasi dengan nilai pengamatan dari variabel respon (Y) yaitu $Y_i - \hat{Y}_i$, adalah nilai dari sisaan, e_i. Dalam perhitungannya digunakan metode *Internally Studentized Residuals*. Nilai *discrepancy* didasarkan atas nilai t_{hitung} yang bertujuan untuk mengevaluasi signifikansi statistik dari residual dalam model regresi dan mengidentifikasi *outlier*. Residual yang di-*studentized* harus berada pada kisaran $+2$ sampai -2, sehingga nilai *discrepancy* yang

melebihi kisaran tersebut dianggap sebagai *outlier*. *Internally studentized residuals* adalah jumlah banyaknya nilai pengamatan ke-i dengan standar deviasi dari sisaan pengamatan ke-i [20].

Residual untuk pengamatan ke-i dihitung sebagai :

$$e_i = Y_i - \hat{Y}_i \tag{2.22}$$

Tentukan estimasi standar deviasi dari residuals yang dihitung sebagai :

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} e_{it}^2}{n-p}} \tag{2.23}$$

Internally Studentized Residuals dihitung dengan rumus :

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{2.24}$$

Ukuran residu yang diperiksa secara internal bervariasi dari 0 sampai $\sqrt{n-k-1}$. Dengan demikian, residu yang diperiksa secara internal tidak dapat diinterpretasikan dengan kurva standar atau tabel-t [17]. Jika model regresi sesuai dengan data, maka residual yang diperiksa secara eksternal mengikuti distribusi t dengan nilai $df = n - k - 1$ [20]. Nilai batasnya dihitung dengan distribusi t, jika nilai $t_i > t_{tabel}$ dengan derajat kebebasan α, maka data tersebut memiliki nilai *discperancy* yang besar dan diidentifikasi sebagai *outlier*.

2.7.3 Titik *Influence*

Dalam analisis regresi, titik pengaruh (*influential point*) adalah pengamatan yang memiliki dampak signifikan pada hasil model regresi. Pengamatan ini tidak hanya mungkin menjadi *outlier* dalam hal nilai residu (*discrepancy*), tetapi juga memiliki *leverage* tinggi, yang berarti metode tersebut berada jauh dari pusat distribusi variabel prediktor dan secara signifikan dapat mempengaruhi parameter model ketika dihilangkan atau dimasukkan dalam analisis. Metode yang digunakan untuk mengevaluasi pola efek adalah metode *global first effect* (*DfFit* dan *Cook'sD*) yang memberikan tentang bagaimana pengamatan ke-i memberi pengaruh pada karakteristik global dari model regresi. Sementara untuk metode yang

kedua adalah *direct effect* (*DfBeta*) yang menggambarkan bagaimana kasus ke-i mempengaruhi masing-masing koefisien regresinya [19].

A. *Difference in Fits* (*DfFit*)

DfFit (*Difference in Fit*) diperkenalkan oleh Belsley, Kuh, dan Welsch dalam buku mereka *"Regression Diagnostics: Identifying Influential Data and Sources of Collinearity"* yang diterbitkan pada tahun 1980. Buku ini menjadi salah satu referensi utama dalam bidang diagnostik regresi dan menguraikan berbagai metode untuk mengidentifikasi pengamatan yang memiliki pengaruh besar terhadap model regresi. *DfFit* juga merupakan salah satu metode diagnostik dalam analisis regresi yang digunakan untuk mengidentifikasi pengamatan yang memiliki pengaruh signifikan terhadap model regresi. *DfFit* mengukur seberapa besar perubahan prediksi yang terjadi ketika suatu pengamatan dikeluarkan dari analisis. Nilai *DfFit* yang tinggi menunjukkan bahwa pengamatan tersebut memiliki pengaruh besar dan mungkin merupakan *outlier* atau titik pengaruh yang signifikan. Dikatakan juga bahwa nilai *influence* diperoleh dari hasil perkalian antara nilai dari titik *leverage* dan nilai *discrepancy*-nya [21]. Rumus metode *DfFit* :

$$DfFit_i = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \tag{2.25}$$

Jika hasilnya menunjukkan bahwa $|DfFit| > 2\sqrt{\frac{p}{n}}$, maka pengamatan tersebut dianggap sebagai *outlier* yang memiliki pengaruh besar terhadap model regresi, dengan p adalah jumlah parameter dalam model dan n adalah jumlah total pengamatan. Maka pengamatan yang sesuai dipertimbangkan sebagai berpengaruh dalam dimensi-X. Dengan t_i adalah *internally stundentized residuals* sedangkan h_{ii} adalah nilai dari *leverage*. Jika nilai t_i dan h_{ii} meningkat, maka nilai *DfFit* akan meningkat. Ini menunjukkan bahwa hasil analisis regresi dipengaruhi secara signifikan oleh kasus ke-i. Sementara apabila nilai $DfFit = 0$ maka kasus ke-i berada tepat pada garis regresi, sehingga jika kasus ke-i tidak dimasukkan atau dihilangkan, nilai $\hat{Y}_i$ tidak berubah. Tanda untuk nilai *DfFit* $\hat{Y}_i > \hat{Y}_{i(i)}$ bernilai positif ini menunjukkan bahwa nilai prediksi

$\hat{Y}_i$ dari pengamatan ke-i meningkat ketika pengamatan ke-i dihilangkan dari model regresi dan juga sebaliknya jika nilai *DfFit* $\hat{Y}_i < \hat{Y}_{i(i)}$ maka akan bernilai negatif menunjukkan bahwa nilai prediksi $\hat{Y}_i$ dari pengamatan ke-i menurun ketika pengamatan ke-i dihilangkan dari model regresi. Jika *DfFit* besar secara absolut (baik positif maupun negatif), ini menunjukkan bahwa pengamatan tersebut memiliki pengaruh signifikan pada model regresi, baik dengan meningkatkan atau mengurangi nilai prediksi.

B. *Difference in Beta* (*DfBeta*)

DfBeta mengukur perbedaan antara koefisien suatu prediktor tertentu ketika suatu variabel tertentu dikeluarkan dan dimasukkan ke dalam regresi. *DfBeta* dihitung untuk setiap koefisien dan setiap variabel dalam model secara terpisah. Rumus *DfBeta* didefinisikan sebagai berikut :

$$DfBeta_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{var(\hat{\beta}_{j(i)})}} \tag{2.26}$$

di mana $\hat{\beta}_j$ adalah estimasi koefisien untuk data keseluruhan ketika *outlier*-nya belum dihilangkan, tapi $\hat{\beta}_{j(i)}$ adalah estimasi koefisien setelah observasi ke-i dikeluarkan dari data. Jika hasil $|DfBeta| > \frac{2}{\sqrt{n}}$ maka observasi tersebut terdapat pengaruh yang signifikan terhadap koefisien dan dianggap memiliki pengaruh yang tinggi [22].

2.8 R Studio

R disebut sebagai bahasa yang dipakai pada perangkat lunak statistik, pertama kali dikemukakan oleh Ross Ihaka dan Robert Gentlement di Universitas Auckland di Selandia Baru. Sebelum R diketahui banyak orang, ada S, keterampilan komputasi statistik yang sama yang dikemukakan oleh John Chambers dan rekan-rekannya di Bell Laboratories. Perbedaan keduanya adalah R merupakan perangkat lunak bebas dan tidak berbayar. Prinsip ini berlaku untuk semua bahasa pemrograman. Setiap bahasa memiliki seperangkat peraturan dan konvensi penulisan yang berbeda. Pemanfaatan

prinsip sederhana di R memungkinkan untuk menyusun algoritma ritme dengan mudah, baik untuk pemahaman penggunanya sendiri maupun untuk yang akan membaca kodenya. Berikut merupakan kelebihan dari *R Studio* [23]:

1. *R studio* memiliki opsi untuk memilih versi gratis *R Studio*, yang memungkinkan untuk mempertahankan semua fitur penting R tanpa batasan apa pun.
2. *R Studio Cloud* menyediakan layanan cloud yang memungkinkan pengguna mengakses dan memanfaatkannya melalui *browser web* di perangkat apa pun.
3. Dengan pemanfaatan *Shiny Apps*, aplikasi berbasis *web* dapat dibuat langsung dari R. Aplikasi ini memiliki berbagai tujuan, mulai dari berfungsi sebagai dasbor interaktif hingga berfungsi sebagai mesin perhitungan otomatis.
4. *R Markdown* menawarkan berbagai opsi file keluaran, termasuk docx, pptx, pdf, html, md, dan banyak lagi. Selain itu, *R Studio* juga memiliki kemampuan untuk menghasilkan *e-book* melalui penggunaan perpustakaan *bookdown*.

Library atau *packages* terdiri dari fungsi standar yang telah dikembangkan dan dapat ditemukan di halaman web CRAN atau GitHub. *Library* ini dapat diinstal dan dimanfaatkan dengan mudah. Beberapa contoh *library* dan *packages* yang sering digunakan:

1) `dplyr`: data carpentry menggunakan tidy principle.
2) `ggplot2`: *data visualization*.
3) `rvest`: *web scraping*.
4) `tidytext`: *text analysis*.
5) `reshape2`: data manipulation.
6) `readxl` atau `openxlsx`: *export* dan *import excel files*.
7) `officer`: membuat Ms. Office files seperti excel, docx, dan powerpoint.
8) `expss` : SPSS di R.
9) `xaringan` : membuat file presentasi berformat html.
10) `plm()` : mengestimasi model regresi panel.
11) `pdata.frame()` : mengkonversi data *frame* menjadi format data panel.

12) `car` (*Companion to Applied Regression*) : menyediakan berbagai alat diagnostik regresi dan tes statistik yang sering digunakan dalam analisis regresi linier.
13) `robustbase` : menyediakan alat untuk regresi *robust* dan analisis statistik yang tahan terhadap *outlier* dan data yang tidak normal.
14) `MASS` (Modern Applied Statistics with S) : berisi berbagai fungsi statistik dan alat analisis data yang luas, termasuk fungsi untuk regresi linier, analisis diskriminan, analisis faktor, dan banyak lagi.