II. TINJAUAN PUSTAKA

2.1 Analisis Regresi

Analisis regresi adalah suatu metode analisis data yang menggambarkan hubungan fungsional antara variabel respon dengan satu atau beberapa variabel prediktor. Misalkan X adalah variabel prediktor dan Y adalah variabel respon untuk n data pengamatan berpasangan $\{(x_i, yi)\}_{i=1}^{n}$, maka hubungan antara variabel prediktor dan variabel respon tersebut dapat dinyatakan sebagai berikut:

$$y_i = f(x_i) + \varepsilon_i \quad ; i = 1,2,3,\ldots,n$$

Dengan ε_i adalah galat yang diasumsikan independen, menyebar normal, dan variansi σ^2 (konstan). $f(x_i)$ disebut sebagai fungsi regresi atau kurva regresi (Hardle,1994).

2.1.1 Analisis Regresi Linear Sederhana

Dalam analisis regresi linear sederhana akan ditentukan persamaan yang menghubungkan dua variabel yang dapat dinyatakan sebagai bentuk persamaaan

pangkat satu (persamaan linier / persamaan garis lurus). Dimana variabel terikat dijelaskan oleh satu variabel bebas. Persamaan umum garis regresi untuk regresi linear sederhana adalah:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

dengan,

Y_i = variabel tak bebas pengamatan ke-i

X_i = variabel bebas pengamatan ke-i

α = konstanta (parameter)

β = koefisien regresi atau slope (parameter)

ε_i = sisaan (galat) pengamatan ke-i

Dalam regresi linier sederhana yang akan diduga adalah α dan β. Persamaan linier untuk pendugaan garis regresi linier ditulis dalam bentuk:

$$\hat{y}_i = a + bx_i$$

dengan,

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

x_i = nilai variabel bebas pengamatan ke-i

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

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

2.1.2 Asumsi Analisis Regresi Linear

Agar mampu memiliki kesimpulan yang benar tentang parameter α dan β, pemenuhan asumsi-asumsi model regresi yang harus terpenuhi (Drapper dan Smith, 1992):

1. Nilai ε_i adalah bebas satu dengan yang lainnya atau korelasi (ε_i , ε_j) = 0. Untuk asumsi pertama yang menyatakan *independent*, artinya ε_i merupakan variabel acak dengan nilai tengah nol dan ragam σ^2 yang tidak diketahui. Jadi, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$. ε_i dan ε_j tidak berkorelasi, $i \neq j$, sehingga $covv(\varepsilon_i\ \varepsilon_j) = 0$. Jadi, $E(Y_i) = \alpha + \beta X_i$, Y_i dan Y_j , $i \neq j$, tidak berkorelasi. ε_i merupakan variabel acak normal, dengan nilai tengah nol dan ragam σ^2_ε dengan kata lain $\varepsilon_i \sim N\ (0, \sigma^2_\varepsilon)$.
2. Nilai tengah dari Y adalah fungsi linier dari X, yaitu jika dihubungkan titik-titik dari nilai tengah yang berbeda, maka akan diperoleh garis lurus $\mu_{(y/x)} = \alpha + \beta X$. Untuk asumsi kedua yang disebut garis linier, artinya X mempunyai hubungan linier dengan Y. Nilai tengah Y untuk kombinasi tertentu dari nilai X adalah fungsi linier dari X, yaitu $\mu_{Y|X}$. ε adalah variabel acak dengan $\mu = 0$ untuk nilai X yang tetap, sehingga $\mu_{\varepsilon|X} = 0$ untuk sembarang X, dengan nilai X yang tetap maka nilai $E(Y) = E(\beta X) = \beta X$. ε menggambarkan seberapa jauh setiap Y menyimpang dari regresi populasinya. Yang dimaksud dengan kelinieran adalah linier dengan koefisien. Jika hubungan titik-titik dari nilai tengah $\mu_{Y|X}$ yang berbeda akan diperoleh garis lurus. Asumsi ini diperlukan agar uji-uji statistik seperti uji F dan uji t menjadi signifikan.

3. Ragam galat homogen (homoskedastik) yaitu galat memiliki nilai ragam yang sama antara galat ke-*i* dan galat ke-*j*. Secara matematis ditulis Var (ε_i) = σ^2 ; i = 1,2,. . . , n dan *n* = banyaknya pengamatan. Untuk asumsi ketiga yang menyatakan varian Y adalah sama untuk setiap kombinasi tetap X; yaitu $\sigma^2_{Y|X}$ = *var* (Y|X) = σ^2 untuk semua X. Asumsi ini sering dikenal dengan sebutan *homoscedasticity*, dengan *homo* berarti sama *scedastic* berarti sebaran. Model regresi menganggap galat menyebar secara normal disekitar nilai tengah nol, dan mempunyai ragam yang sama. Banyak metode yang dapat dipergunakan untuk memeriksa apakah asumsi ini terpenuhi atau tidak, salah satunya adalah dengan metode Glejser. Kehomogenan diperlukan agar uji-uji statistik seperti uji F dan uji t menjadi signifikan.
4. Ragam galat menyebar normal dengan rata-rata nol dan suatu ragam tertentu. Asumsi keempat menyatakan untuk sembarang kombinasi tetap dari variabel bebas X, variabel tak bebas Y berdistribusi normal atau yang biasa disebut asumsi kenormalan. Dengan kata lain $Y \sim N$ ($\mu_{Y|X}$, σ^2). $E(Y)$ = $E(X\beta)$ + $E(\varepsilon)$ dengan $E(\varepsilon)$ = 0 sehingga $E(Y)$ = ($X\beta$). Dan *Var* (Y) = *Var* (X + β) = *Var* (ε) = σ^2. ε_i merupakan variabel acak dengan nilai tengah nol dan ragam σ^2, sehingga $\varepsilon \sim N$ (0, σ^2).

   Sebaran normal diperlukan agar uji t maupun uji F dapat dilakukan. Kenormalan bisa dilihat secara eksploratif melalui plot sisaan sedangkan untuk uji formalnya dapat digunakan uji Kolmogorov-Smirnov.

2.2 Uji heteroskedasitas

Heteroskedasitas adalah variansi dari galat model regresi tidak konstan atau variansi antar galat yang satu dengan galat yang lain berbeda. Dampak adanya heteroskedasitas dalam model regresi adalah walaupun estimator MKT masih linier dan tidak bias, tetapi tidak lagi mempunyai variansi yang minimum dan menyebabkan perhitungan *standard error* metode MKT tidak bisa dipercaya kebenarannya. Selain itu interval estimasi maupun pengujian hipotesis yang didasarkan pada distribusi t maupun F tidak bisa lagi dipercaya untuk evaluasi hasil regresi.

Selanjutnya dilakukan deteksi masalah heteroskedasitas dalam model regresi. Untuk pengujian hetroskedasitas pada penulisan ini dilakukan dengan metode Glejser. Glejser merupakan seorang ahli elektronometrika dan mengatakan bahwa nilai variansi variabel galat model regresi tergantung dari variabel bebas. Selanjutnya untuk mengetahui apakah pola variabel galat mengandung heteroskedasitas, Glejser menyarankan untuk melakukan regresi nilai mutlak residual dengan variabel bebas. Jika hasil uji F dari model regresi yang diperoleh tidak signifikan, maka tidak ada heteroskedasitas dalam model regresi (Widarjono, 2007).

2.3 Uji Normalitas

Uji normalitas berguna pada tahap awal dalam metode pemilihan analisis data. Jika data normal, maka digunakan statistik parametrik dan jika data tidak normal digunakan statistik nonparametrik. Tujuan uji normalitas data ini adalah untuk mengetahui apakah dalam model regresi variabel pengganggu atau residual memiliki distribusi normal. Pengujian ini diperlukan karena untuk melakukan uji t dan uji F mengasumsikan bahwa nilai residual mengikuti distribusi normal (Drapper dan Smith, 1992). Untuk pengujian normalitas, pada pengujian ini dilakukan dengan uji normalitas Kolmogorov Smirnov.

Uji Kolmogorov-Smirnov menggunakan hipotesis:

Ho : Data residual berdistribusi normal

H_1 : Data residual tidak berdistribusi normal

Konsep dasar dari uji normalitas Kolmogorov Smirnov adalah dengan membandingkan distribusi data (yang akan diuji normalitasnya) dengan distribusi normal baku. Distribusi normal baku adalah data yang telah ditransformasikan ke dalam bentuk Z-Score dan diasumsikan normal. Jadi sebenarnya uji Kolmogorov Smirnov adalah uji beda antara data yang diuji normalitasnya dengan data normal baku. Seperti pada uji beda biasa, jika signifikansi di bawah 0,05 berarti terdapat perbedaan yang signifikan dan jika signifikansi di atas 0,05 maka tidak terjadi perbedaan yang signifikan. Penerapan pada uji Kolmogorov Smirnov adalah bahwa jika signifikansi di bawah 0,05 dengan $\alpha = 5\%$ berarti data yang akan diuji mempunyai perbedaan yang signifikan dengan data normal baku, berarti data

tersebut tidak normal (H_1 diterima). Lebih lanjut, jika signifikansi di atas 0,05 dengan α = 5% maka berarti tidak terdapat perbedaan yang signifikan antara data yang akan diuji dengan data normal baku, artinya data yang kita uji normal (H_0 diterima).

2.4 Metode Kuadrat Terkecil

Metode Kuadrat Terkecil (MKT) merupakan salah satu metode penduga parameter yang terbaik karena bersifat tak bias dan efisien. Metode kuadrat terkecil akan menghasilkan ragam minimum bagi parameter regresi. Prinsip dasar metode ini adalah meminimumkan jumlah kuadrat galat.

Dengan menggunakan persamaan linier untuk pendugaan garis regresi linier, MKT dapat diuraikan dengan notasi matematika yaitu sebagai berikut:

$$\hat{y}_i = a + bx_i$$

Jarak vertikal antara titik observasi (x_i , y_i) dan titik ($\hat{x}_i$, $\hat{y}_i$) pada garis dugaan dapat ditulis:

$$|y_i - \hat{y}_i| \; atau \; | y_i - \hat{a} - \hat{b}x_i|$$

Jumlah kuadrat dari semua jarak ini ditulis:

$$\sum_{i=1}^{n}(yi - \hat{y}i)^2 = \sum_{i=1}^{n}(yi - \hat{a} - \hat{b}x_i)^2$$

Solusi dari MKT dapat dilakukan sebagai berikut:

$$S(a,b) = \sum_{i=1}^{n} (yi - \hat{a} - \hat{b}x_i)^2$$

$$\frac{dS_{(a,b)}}{da} = -2 \sum_{i=1}^{n} (yi - \hat{a} - \hat{b}x_i) = 0$$

$$\frac{dS_{(a,b)}}{db} = -2 \sum_{i=1}^{n} (yi - \hat{a} - \hat{b}x_i)x_i = 0$$

Dengan menyederhanakan kedua persamaan ini maka diperoleh:

$$a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$na + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b = \frac{\sum_{i=1}^{n} XiYi - (\sum_{i=1}^{n} Xi)(\sum_{i=1}^{n} Yi)/n}{\sum_{i=1}^{n} Xi2 - (\sum_{i=1}^{n} Xi)2/n}$$

$$b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})((Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})2} \quad \text{dan} \quad a = \bar{Y} - b\,\bar{X}$$

Persamaan garis regresi kuadrat terkecil yang didapat adalah:

$$\hat{y} = a + bX \text{ atau } \hat{y} = \hat{Y} + b\,(X - \bar{X})$$

Persamaan garis diatas dapat digunakan untuk memprediksi Y oleh nilai X yang berpadanan (Myers dan Milton,1991).

2.5 Regresi Robust

Regresi *robust* adalah salah satu metode penduga regresi yang digunakan ketika distribusi dari galat tidak normal atau adanya beberapa pencilan yang berpengaruh

pada model (Ryan, 1997). Metode ini dibutuhkan karena metode kuadrat terkecil yang dianggap penduga terbaik dalam analisis regresi ternyata peka terhadap data yang menyimpang dari asumsi.

Prosedur robust ditunjukkan untuk memberikan dugaan yang lebih tepat dan cepat terhadap data yang melanggar asumsi dengan cara meniadakan identifikasi adanya data pencilan, serta bersifat otomatis dalam menanggulangi data pencilan. Beberapa metode dalam regresi *robust* diantaranya adalah Theil-Sen, *Least Trimmed Square* (LTS), *Least Mean Square* (LMS), *MM estimator*, *S estimator*, dan M *estimator* (penduga M).

2.6 Median Data

Median segugus data yang telah diurutkan dari yang terkecil sampai terbesar atau terbesar sampai terkecil adalah pengamatan yang tepat di tengah-tengah bila banyaknya pengamatan itu ganjil atau rata-rata kedua pengamatan yang di tengah bila banyaknya pengamatan genap. Kalau nilai median sama dengan Me, maka 50% dari data paling tinggi sama dengan Me sedangkan 50% lagi paling rendah sama dengan Me.

Median adalah nilai tengah dari data-data yang terurut.
Ada dua cara menentukan median:

1. Jika jumlah data adalah ganjil maka median Me setelah data disusun menurut nilainya, merupakan data paling tengah. Nilai mediannya dapat ditentukan dengan rumus :

   Index Median = ((n-1)/2+1)

Median = data ke-(Index Median)

2. Jika jumlah data genap maka median Me setelah data disusun menurut urutan nilainya merupakan rata-rata hitung dua data tengah. Nilai mediannya dapat ditentukan dengan rumus :

   Index Median = n/2

   Median= (data ke-(Index Median) + data ke-(Index Median +1))/2.

2.7 Pendugaan Persamaan Regresi Metode Theil-Sen

Pada statistika non-parametrik terdapat metode regresi *robust* yang memilih median kemiringan dari semua garis berpasangan dari dua dimensi titik sampel yang dinamakan penduga Theil-Sen. Nama ini didapat setelah Henri Theil dan Pranab K.Sen yang memperkenalkan metode ini pada tahun 1950 dan 1968.

Misalkan ada n pasangan data pengamatan, yaitu $(X_1 , Y_1), (X_2 , Y_2) , . . . , (X_n , Y_n)$ dengan persamaan regresi linear sederhana:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Dengan α adalah koefisien intercep (titik potong), β adalah koefisien kemiringan dari garis tersebut, X_i adalah variabel bebas dan Y_i adalah variabel terikat. Metode Theil-Sen ini menaksir koefisien kemiringan garis regresi dengan median kemiringan dari seluruh pasangan garis dari titik-titik variabel X dan Y.

Untuk semua titik sampel data ke-i $p_i = (x_i , y_i)$ dan titik sampel data ke-j $p_j = (x_j , y_j)$ dengan $x_i < x_j$ kemudian misalkan $b_{i,j} = (y_j – y_i) / (x_j – x_i)$ merupakan

kemiringan dari dua titik sampel data. Kemiringan suatu garis dapat diperoleh dengan menggunakan rumus:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Dengan m adalah kemiringan suatu garis, (x_1 , y_1) dan (x_2 , y_2) adalah titik sampel data pada garis tersebut. Pada metode Theil-Sen koefisien kemiringan diduga dengan menggunakan rumus sebagai berikut:

b = median(b_{ij})

= median ($\frac{y_j - y_i}{x_j - x_i}$) ; j>i dan $x_j \neq x_i$

dengan,

y_j = nilai variabel terikat pengamatan ke-j

y_i = nilai variabel terikat pengamatan ke-i

x_j = nilai variabel bebas pengamatan ke-j

x_i = nilai variabel bebas pengamatan ke-i

Sedangkan intersep diduga dengan menggunakan rumus:

$$\hat{\alpha} = M_y - \hat{\beta} M_x$$

Dimana M_y adalah nilai median atau nilai tengah dari data-data yang terurut dari yang terkecil sampai terbesar atau terbesar sampai terkecil berdasarkan sampel Y_1,..., Y_n dan M_x adalah nilai median atau nilai tengah dari data-data yang terurut

dari yang terkecil sampai terbesar atau terbesar sampai terkecil berdasarkan sampel $X_1,..., X_n$.

Analisis regresi dengan metode Theil-Sen dilandasi pada asumsi-asumsi sebagai berikut :

a. Persamaan regresinya adalah: $Y_i = \alpha + \beta X_i + \varepsilon_i$, i = 1,. . . , n dengan X_i adalah peubah bebas ke-i, α dan β adalah parameter-parameter yang tidak diketahui.
b. Untuk masing-masing nilai X_i terdapat nilai Y_i.
c. Y_i adalah nilai yang teramati ke-i dari Y yang acak dan kontinu untuk nilai X_i.
d. Semua nilai X_i saling bebas dan kita menetapkan $X_1 < X_2 < . . . < X_n$.
e. Nilai-nilai ε_i saling bebas dan berasal dari populasi yang sama.

Pengujian koefisien kemiringan ini dengan membuat statistik tataan dan memperbandingkan semua hasil pengamatan menurut nilai-nilai *X* (Daniel, 1989).