BAB II

KONSEP DASAR

2.1 Regresi Linier

Regresi Linier adalah salah satu cara yang digunakan dalam analisis data statistik. Model regresi menggambarkan hubungan antara variabel bebas (x) dengan variabel terikat/respon (Y). Model tersebut dapat dituliskan sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1,2,3, \ldots , n. \tag{2.1.1}$$

Model regresi linier dalam persamaan (2.1.1) disebut model regresi linier berganda dengan k variabel bebas yaitu X_1. X_2, ... ,X_k. Parameter β_j, j=0,1,...,k disebut koefisien regresi. Model ini menggunakan asumsi $\varepsilon_i \sim N(0,\sigma^2)$.

2.1.1 Metode Kuadrat Terkecil

Dalam model regresi linier, untuk taksiran parameternya digunakan metode kuadrat terkecil. Diberikan fungsi kuadrat terkecil sebagai berikut:

$$L= \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right]^2$$

Fungsi L tersebut diminimumkan terhadap $\beta_o, \beta_1, \ldots, \beta_k$ Estimator kuadrat terkecil $\beta_o, \beta_1, \ldots, \beta_k$ harus memenuhi:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\beta_0,\beta_1,\ldots,\beta_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum \hat{\beta}_j x_{ij}\right) = 0 \text{ dan}$$

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\beta_0,\beta_1,\ldots,\beta_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum \hat{\beta}_j x_{ij}\right)x_{ij} = 0 \text{ dengan } i=1,2,3,\ldots n\ ;\ j=1,2,3,\ldots,k$$

Sehingga diperoleh persamaan - persamaan normal kuadrat terkecil sebagai berikut:

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} x_{i1} + \hat{\beta}_2\sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0\sum_{i=1}^{n} x_{i1} + \hat{\beta}_1\sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\vdots$$

$$\hat{\beta}_o\sum_{i=1}^{n} x_{ik} + \hat{\beta}_1\sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i \qquad (2.1.1.1)$$

Penyelesaian model ini akan lebih sederhana dengan menggunakan bentuk matriks. Jika didefinisikan dalam bentuk matrik dan vektor maka persamaan (2.1.1) dapat ditulis sebagai berikut:

$$y = X\beta + \varepsilon \qquad (2.1.1.2)$$

dengan,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ . \\ . \\ \varepsilon_k \end{bmatrix}$$

y adalah vektor pengamatan pada variabel tak bebas berukuran n x 1 x adalah matriks variabel bebas berukuran n x (k+1)

β adalah vektor koefisien regresi berukuran (k+1) x 1

ε adalah vektor random error berukuran n x 1

Untuk mendapatkan vektor penaksir kuadrat terkecil $\hat{\beta}$ dilakukan dengan cara meminimumkan jumlah kuadrat galatnya, yaitu:

$$
\begin{aligned}
L &= \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon \\
&= (y - X\beta)'(y - X\beta) \\
&= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\
&= y'y - 2\beta'X'y + \beta'X'X\beta
\end{aligned}
$$

Karena $\beta' X' y$ adalah sebuah vektor matriks skalar (1X1), dan ($\beta' X' y$) = $y' X\beta$ adalah skalar yang sama, maka penaksir kuadrat terkecil itu harus memenuhi:

$$
\left.\frac{\partial L}{\partial \beta}\right|_{\hat{\beta}} = 0
$$

$$
-2X'y + 2X'X\hat{\beta} = 0
$$

penyederhanaannya menjadi:

$$
X'X\hat{\beta} = X'y \qquad (2.1.1.3)
$$

Jika $(X'X)^{-1}$ ada maka diperoleh penyelesaian untuk $\hat{\beta}$ yaitu:

$$
\hat{\beta} = (X'X)^{-1} X'y \qquad (2.1.1.4)
$$

Jaminan bahwa nilai dari LSE adalah minimum yaitu bahwa turunan ke dua dari L terhadap β harus bernilai positif. Diberikan:

$$\frac{\partial^2 L}{\partial \beta} = 2X'X$$

dengan demikian nilai L akan minimum apabila nilai 2X'X lebih besar dari 0. Karena matrix X'X adalah definit positif dengan semua unsur diagonalnya berbentuk kuadrat, maka turunan kedua dari L terhadap β bernilai positif.

Estimasi $\hat{\beta}$ dapat dinyatakan dengan persamaan:

$$\begin{aligned} E(\hat{\beta}) &= E\left[(X'X)^{-1}X'y\right] = (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned} \tag{2.1.1.5}$$

Matriks variansi dari persamaan (2.1.1.4) adalah:

$$\begin{aligned} \operatorname{var}(\hat{\beta}) &= E\left[(\hat{\beta} - E(\hat{\beta}))\right]\left[\hat{\beta} - E(\hat{\beta})\right] \\ &= E(X'X)^{-1}X[y - E(y)][y' - E(y')]X(X'X)^{-1} \\ &= (X'X)^{-1}XE(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= (X'X)^{-1}\sigma^2 \end{aligned} \tag{2.1.1.6}$$

Persamaan (2.1.1.6) merupakan varian terkecil dari semua estimator linier tak bias, hal ini dijamin dengan teorema Gauss-Markov.

Teorema Gauss-Markov

Estimator kuadrat terkecil $\hat{\beta} = (X'X)^{-1}X'y$ mempunyai variansi terkecil dalam himpunan semua estimator linier tak bias.

Bukti:

Misal $\hat{\beta}^*$ estimator linier lain dari β yang juga tak bias. Karena $\hat{\beta}^*$ estimator linier maka dapat dimisalkan bentuknya sebagai

$$\hat{\beta}^* = [(X'X)^{-1}X' + U]Y$$

untuk suatu matrix U yang merupakan fungsi dari x, sehingga nilai estimasi untuk $\hat{\beta}^*$ dapat diberikan:

$$\begin{aligned}
E(\hat{\beta}^*) &= [(X'X)^{-1}X' + U]E(Y) \\
&= [(X'X)^{-1}X' + U](X\beta) \\
&= \beta + UX\beta \\
&= (1 + UX)\beta
\end{aligned}$$

Agar $\hat{\beta}^*$ estimasi tak bias dari β maka UX = 0, sehingga:

$$\begin{aligned}
Var(\hat{\beta}^*) &= [(X'X)^{-1}X' + U]Var(Y)[U' + X(X'X)^{-1}] \\
&= \sigma^2 [(X'X)^{-1}X'U' + UU' + (X'X)^{-1} + UX(X'X)^{-1}]
\end{aligned}$$

karena UX = X'U' = 0 maka

$$var(\hat{\beta}^*) = \sigma^2 [(X'X)^{-1} + UU']$$

Matrix **UU'** adalah definit positif, karena semua unsur diagonalnya berbentuk kuadrat. Jadi terbukti bahwa variansi dari setiap unsur dari vektor $\hat{\beta}^*$ selalu lebih besar, atau paling kecil sama dengan variansi unsur $\hat{\beta}$ yang sesuai.

2.1.2 Estimasi E(y)

Estimator $\hat{\beta}$ dapat digunakan untuk mengestimasi E(y), dengan persamaan:

$$\hat{E}(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \tag{2.1.2.1}$$

berkaitan dengan $E(y) = \mathbf{X}\hat{\beta}$, diperoleh:

$$\hat{E}(y) = \mathbf{X}\hat{\beta} \tag{2.1.2.2}$$

Variansi dari estimator di atas adalah:

$$\mathrm{var}[\hat{E}(y)] = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\sigma^2 \tag{2.1.2.3}$$

Pandang y_f sebagai nilai observasi yang akan datang, yang berkaitan dengan nilai x, yaitu x_f. Maka dengan model $y_f = \mathbf{x}'_f\beta + \mathbf{e}_f$ dengan e_f adalah error. Oleh karena itu prediksi terbaik dari y_f yaitu $\tilde{y}_f$, adalah $\tilde{y}_f = \mathbf{x}'_f\hat{\beta}$. Dengan demikian $\mathbf{x}'_f\hat{\beta}$ dapat digunakan sebagai prediksi dari observasi yang akan datang dan dapat digunakan sebagai suatu estimator dari nilai harapan $E(y_f)$ berkaitan dengan x'_f. Kemudian dapat diturunkan suatu variansi dari simpangan y_f dan $\tilde{y}_f$ sebagai berikut:

$$y_f - \tilde{y}_f = y_f - \mathbf{x}'_f\hat{\beta} = \mathbf{x}'_f(\beta - \hat{\beta}) + \mathbf{e}_f \tag{2.1.2.4}$$

Karena $\hat{\beta}$ dan e_f independen maka cov ($\hat{\beta}$,e_f) = 0. Hal ini berakibat:

$$\text{var}(\,y_f - \tilde{y}_f\,) = x_f'\text{var}(\beta - \hat{\beta})x_f + \text{var}(e_f) = [x_f'(X'X)^{-1}x_f + 1]\sigma^2 \qquad (2.1.2.5)$$

Variansi dari $\hat{E}(y) = X\hat{\beta}$ adalah $[x_f'(X'X)^{-1}x_f]\sigma^2$. Dan Var($y_f$) = σ^2.

2.1.3 Jumlah Kuadrat Galat (JKG)

Jika $\hat{y}$ merupakan nilai harapan dari y yang terestimasi berkaitan dengan observasi y , maka :

$$\hat{y} = X\hat{\beta} \qquad (2.1.3.1)$$

sehingga simpangan dari y_i yang merupakan nilai observasi dengan nilai prediksinya adalah:

$$y - \hat{y} = y - X(X'X)^{-1}X'y = [I - X(X'X)^{-1}X']y \qquad (2.1.3.2)$$

dengan $I - X(X'X)^{-1}X'$ adalah simetris dan Idempoten (2.1.3.3)

serta $[I - X(X'X)^{-1}X']X = 0$ (2.1.3.4)

Jumlah kuadrat dari simpangan y_i dan nilai prediksinya biasa disebut sebagai residual, atau Jumlah Kuadrat Galat (Sum Squares Error=SSE), dengan persamaan:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (y - \hat{y})'(y - \hat{y}) \qquad (2.1.3.5)$$

Untuk selanjutnya rumus SSE dapat diperoleh dengan mensubstitusikan persamaan (2.1.3.2) ke dalam persamaan (2.1.3.5) serta menggunakan persamaan (2.1.3.3) dan (2.1.3.4), sehingga persamaannya adalah:

$$SSE = y'\left[I - X(X'X)^{-1}X'\right]y \tag{2.1.3.6}$$

$$= y'y - y'X(X'X)^{-1}X'y$$

$$= y'y - \hat{\beta}'X'y \tag{2.1.3.7}$$

dengan diketahui bahwa $\hat{\beta}' = y'X(X'X)^{-1}$, $\hat{\beta}'X'y$ merupakan jumlah dari perkalian elemen-elemen $\hat{\beta}$ dengan $X'y$, yaitu ruas kanan dari persamaan $X'X\hat{\beta} = X'y$.

SSE pada persamaan (2.1.3.6) ditulis dalam bentuk kuadratis dalam y yaitu:

$$SSE = = y'\left[I - X(X'X)^{-1}X'\right]y$$

Menurut Teorema 1 (Searle, 1971) bahwa : jika **x** berdistribusi N(μ,V)

$$E(\mathbf{x'Ax}) = tr(\mathbf{AV}) + \mathbf{\mu'A\mu}$$

Hal ini berakibat, jika y berdistribusi $N(X\beta, I\sigma^2)$, maka diperoleh:

$$E(SSE) = tr\left[I - X(X'X)^{-1}X'\right]I\sigma^2 + \beta'X'\left[I - X(X'X)^{-1}X'\right]X\beta$$

$$= tr\left[I - X(X'X)^{-1}X'\right]\sigma^2$$

$$= [n - r(\mathbf{x})]\ \sigma^2$$

Diketahui bahwa trace dari suatu matriks idempoten sama dengan ranknya. Dengan demikian suatu estimator tak bias untuk $\hat{\sigma}^2$ adalah

$$\hat{\sigma}^2 = \frac{SSE}{n - r(X)} = \frac{SSE}{n - r}$$

dengan r = r(X), rank dari matriks X.Dalam model regresi penuh r(X) = k+1 dengan k adalah banyaknya variabel prediktor yang terlibat.

2.1.4 Jumlah Kuadrat Total (JKT)

Jumlah Kuadrat Total (Sum Squares Total = SST) diberikan dengan persamaan:

$$SST = y'y = \sum_{i=1}^{n} y_i^2 \qquad (2.1.4.1)$$

Jumlah Kuadrat Regresi (Sum Square Regression = SSR) dari y_i dengan nilai prediksinya adalah yang tertera pada persamaan (2.1.3.7) merupakan selisih antara nilai jumlah kuadrat total dengan jumlah kuadrat galat dan dapat ditulis dalam persamaan:

$$SSR = SST - SSE = \mathbf{y}'\mathbf{y} - \left(\mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}\right) = \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \qquad (2.1.4.2)$$