BAB II

TEORI PENUNJANG

Dalam analisis data statistik, harus diketahui konsep dasar yang berhubungan dengan data statistik, karena dengan mengetahui konsep dasarnya, berarti akan mempermudah dalam pembahasan lebih lanjut. Disini akan dibahas konsep dasar tentang variabel random, variabel indikator, distribusi binomial, dan model regresi linier, serta metode yang berhubungan dalam analisis regresi. Adapun penyajiannya adalah sebagai berikut :

2.1 Variabel Random

Definisi 2.1.1

Pandang sebuah percobaan random dengan ruang sampel ξ. Suatu fungsi X yang mengawankan setiap elemen $c \in \xi$ dengan satu dan hanya satu bilangan riil $X(c) = x$, disebut sebagai variabel random dengan range $\mathscr{A} = \{x : x = X(c) \ , c \in \xi \}$

Definisi 2.1.2

Jika range $\mathscr{A}$ dari variabel random X memuat titik-titik yang banyaknya berhingga atau anggota $\mathscr{A}$ dapat dipasangkan berkorespondensi satu-satu dengan bilangan integer positif, maka X disebut suatu variabel random diskret.

Definisi 2.1.3

Jika range $\mathscr{A}$ dari variabel random X berupa interval atau kumpulan dari interval-interval, maka X disebut suatu variabel random kontinu.

2.2 Variabel Indikator

Definisi 2.2.1

Diberikan kejadian A, dimana merupakan suatu himpunan bagian yang didiskripsikan dari ruang sampel ξ. Fungsi indikator I_A pada kejadian A didefinisikan dalam ξ dengan :

$I_A(s) = 1$, jika s anggota A.

$I_A(s) = 0$, jika s bukan anggota A.

2.3 Fungsi Densitas Probabilita

Definisi 2.3.1

Suatu fungsi $f(x)$ yang didefinisikan pada $\mathscr{A}$ ke himpunan bilangan riil disebut sebagai fungsi densitas probabilita diskrit, jika memenuhi :

i. $f(x) > 0, \forall x \in \mathscr{A}$ dan $f(x) = 0, \forall x \notin \mathscr{A}$

ii. $\sum_{A} f(x) = 1$

Definisi 2.3.2

Suatu fungsi $f(x)$ yang didefinisikan pada $\mathscr{A}$ ke himpunan bilangan riil disebut fungsi densitas probabilita kontinu jika memenuhi :

i. $f(x) > 0, \forall x \in \mathscr{A}$ dan $f(x) = 0, \forall x \notin \mathscr{A}$

ii. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

2.4 Ekspektasi

Definisi 2.4.1

Misalkan X adalah variabel random yang mempunyai fungsi densitas probabilita $f(x)$, maka ekspektasi dari variabel random X adalah :

$E(X) = \sum_x x f(x)$ untuk variabel random diskrit.

$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\,dx$ untuk variabel random kontinu.

Sifat-sifat Ekspektasi :

1. $E(a) = a$, bila a adalah konstanta.
2. $E(aX + b) = a\,E(x) + b$, bila a, b konstanta
3. $E(\sum k_i X_i) = \sum k_i E(X_i)$
4. $E(XY) = E(X) \cdot E(Y)$, jika X dan Y saling bebas.

2.5 Variansi

Definisi 2.5.1

Diberikan variabel random X dan nilai rata-rata dari X didefinisikan dengan $\mu = E(X)$. Ekspektasi dari kuadrat deviasi antara X dengan μ disebut variansi. Notasinya adalah sebagai berikut :

$Var(X) = E\{[X - \mu]^2\}$

Sifat-sifat variansi :

1. Var (a) = 0 , bila a. konstanta
2. Var (aX + b) = a^2 Var (X) , bila a.dan b konstanta
3. Var (X+Y) = Var (X) + Var (Y) , bila X dan Y saling bebas.

2.6 Distribusi Binomial

Percobaan Bernoulli adalah percobaan random, dimana tiap percobaan tunggal hanya mempunyai dua kemungkinan hasil. Katakan sukses atau gagal, misal : baik atau buruk, laki-laki atau perempuan, hidup atau mati.

Misal X sebagai variabel random yang menyatakan percobaan Bernoulli, dan didefinisikan dengan X(sukses) = 1 dan X(gagal) = 0. Terdapat 2 hasil percobaan yaitu sukses dan gagal yang masing-masing ditunjukkan dengan 1 dan 0. Fungsi densitas probabilita dari X ditulis sebagai berikut :

$$f(x) = p^x (1-p)^{1-x} \qquad x = 0,1 \qquad \ldots(2.6.1)$$

Variabel random X yang mempunyai fungsi $f(x)$ tersebut dikatakan berdistribusi Bernoulli. Sedangkan ekspektasi dan variansinya adalah:

$$E(X) = \sum_x x \; f(x) = p \qquad \ldots(2.6.2)$$

$$Var(X) = E(X^2) - [E(X)]^2 = p(1-p) = pq \qquad \ldots(2.6.3)$$

Dilakukan percobaan Bernoulli dengan kemungkinan hasil sebanyak n kali. Misal X menunjukkan variabel random Bernoulli dengan nilai kemungkinan dari X adalah 0,1,2,…,n yang menghasilkan x kejadian sukses, dan (n-x) kejadian gagal.

Bilangan yang menunjukkan kejadian sukses sebanyak x kali diantara n percobaan adalah :

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Probabilita dari n percobaan ditunjukkan dengan $p^x(1-p)^{n-x}$. Fungsi densitas probabilita dari X adalah :

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad , x = 0,1,2,\ldots,n$$

$$= 0 \quad , \text{lainnya} \qquad \ldots(2.6.4)$$

Variabel random X yang mempunyai fungsi $f(x)$ tersebut dikatakan berdistribusi binomial.

Ekspektasinya adalah sebagai berikut :

$$E(X) = \sum_x x.f(x) = np \qquad \ldots(2.6.5)$$

Dengan variansi :

$$Var(X) = E(X^2) - [E(X)]^2 = np(1-p) = npq \qquad \ldots(2.6.6)$$

2.7 Distribusi Chi-Kuadrat

Distribusi ini berperan penting dalam uji signifikansi dari koefisien-koefisien model regresi logistik. Distribusi Chi-Kuadrat dinotasikan dengan χ^2. Suatu variabel random X dikatakan berdistribusi Chi-Kuadrat, jika mempunyai fungsi kepadatan peluang sebagai berikut :

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}} \quad , 0 < x < \infty \qquad \ldots(2.7.1)$$

$$= 0 \quad , \text{lainnya}$$

Ekspektasi dan variansi dari distribusi Chi-Kuadrat adalah :

$$E(X) = r \qquad \ldots(2.7.2)$$

$$\text{Var}(X) = 2r \qquad \ldots(2.7.3)$$

2.8 Model Regresi Linier

Regresi berganda adalah salah satu cara yang digunakan dalam analisis data statistik. Model regresi menggambarkan hubungan antara variabel bebas (X) dengan variabel terikatnya (Y). Model tersebut adalah :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i \qquad \ldots(2.8.1)$$

Model regresi diatas disebut model regresi linier berganda dengan p variabel bebas yaitu $x_1, x_2, \ldots, x_p$. Parameter β_j, $j = 0,1,2,\ldots,p$ disebut koefisien regresi. Model ini menggunakan asumsi $\varepsilon_i \sim N(0,\sigma^2)$. Dalam model regresi linier, untuk taksiran parameternya digunakan metode kuadrat terkecil.

Diberikan fungsi kuadrat terkecil sebagai berikut :

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

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \qquad \ldots(2.8.2)$$

Fungsi L tersebut diminimumkan terhadap $\beta_0,\beta_1,\ldots,\beta p$. Estimator kuadrat terkecil dan $\beta_0,\beta_1,\ldots,\beta p$ harus memenuhi :

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

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\beta_0,\beta_1,\ldots\hat{\beta}_p} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum \hat{\beta}_j \hat{x}_{ij})\, x_{ij} = 0 \qquad i,j = 1,2,\ldots,p$$

diperoleh persamaan-persamaan normal kuadrat terkecil :

$$\begin{aligned}
n\hat{\beta}_0 &+ \hat{\beta}_1\sum_{i=1}^{n}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik} = \sum_{i=1}^{n}y_i \\
\hat{\beta}_0\sum_{i=1}^{n}x_{i1} &+ \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} = \sum_{i=1}^{n}x_{i1}y_i \qquad \ldots(2.8.3) \\
\hat{\beta}_0\sum_{i=1}^{n}x_{ik} &+ \hat{\beta}_1\sum_{i=1}^{n}x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{ik}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 = \sum_{i=1}^{n}x_{ik}y_i
\end{aligned}$$

Penyelesaian model ini lebih sederhana jika menggunakan bentuk matriks. Persamaan (2.8.3) dapat ditulis dalam bentuk matriks sebagai :

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad \ldots(2.8.4)$$

dimana :

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

y adalah vektor pengamatan pada variabel tak bebas berukuran nx1

x adalah matriks variabel bebas berukuran nx(k+1)

β adalah vektor koefisien regresi berukuran (k+1)x1

ε adalah vektor random error berukuran nx1

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

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

karena $\beta'\mathbf{X}'\mathbf{y}$ adalah sebuah vektor matriks skalar (1x1), dan $(\beta'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}\beta$ adalah skalar yang sama, maka penaksir-penaksir kuadrat terkecil itu harus memenuhi:

$$\left.\frac{\partial \mathbf{L}}{\partial \beta}\right|_{\hat{\beta}} = 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{0}$$

penyederhanaannya menjadi :

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \qquad (2.8.4)$$

2.9 Metode Maksimum Likelihood

Metode maksimum likelihood adalah metode yang baik untuk memperoleh taksiran parameter yang tidak diketahui dari populasi, yaitu dengan cara memaksimumkan fungsi likelihood. Misalkan X adalah variabel random dengan distribusi probabilita $f(x, \beta)$, dimana parameter $\beta = (\beta_0, \beta_1, \ldots, \beta p)$.tidak diketahui.

Misalkan X_1, X_2 , ... ,X_n menjadi nilai yang diobservasi di dalam suatu sampel random yang besarnya n, maka fungsi likelihood tersebut adalah :

$$l(\beta) = f(x_1,\beta).f(x_2,\beta)....f(x_n,\beta)$$

$$= \prod_{i=1}^{n} f(x_i,\beta) \qquad i = 1,2,\ldots,n \qquad \ldots(2.9.1)$$

Dalam regresi logistik, error diasumsikan berdistribusi binomial, sehingga fungsi densitasnya mengikuti distribusi binomial. Taksiran maksimum likelihood β adalah nilai β yang memaksimumkan fungsi likelihood l(β). Untuk lebih mudahnya, dalam memaksimumkan l(β), maka dapat dibentuk logaritma dari fungsi likelihood ln l(β). Menurut hitung differensial, persamaan maksimum likelihood didapat dari turunan parsial pertama dari ln l(β) terhadap β, yang kemudian disamadengankan dengan nol, yaitu :

$$L(\beta) = \ln l(\beta).$$

$$\frac{\partial L(\beta)}{\partial \beta_i} = 0 \qquad i = 0,1,2,\ldots,n \qquad \ldots(2.9.2)$$

Persamaan (2.9.2) dapat ditampilkan dalam bentuk matrik sebagai berikut :

$$\frac{\partial L(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \frac{\partial L(\beta)}{\partial \beta_2} \\ \cdot \\ \cdot \\ \frac{\partial L(\beta)}{\partial \beta_p} \end{bmatrix} \qquad \ldots(2.9.3)$$

Dalam regresi logistik, persamaan log likelihoodnya bukan merupakan fungsi linier dalam β, sehingga harga taksiran β dicari dengan menggunakan iterasi Newton-Raphson. Oleh karena itu diperlukan turunan parsial kedua L(β). Pada persamaan (2.9.3) untuk setiap u, u = 0,1,2,...,p turunan parsial kedua dari log likelihood terhadap β_j , β_u yaitu :

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = \begin{bmatrix} \frac{\partial^2 L(\beta)}{\partial \beta_0^2} & \frac{\partial^2 L(\beta)}{\partial \beta_1 \partial \beta_0} & \cdots & \frac{\partial^2 L(\beta)}{\partial \beta_p \partial \beta_0} \\ \frac{\partial^2 L(\beta)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 L(\beta)}{\partial \beta_1^2} & \cdots & \frac{\partial^2 L(\beta)}{\partial \beta_p \partial \beta_0} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 L(\beta)}{\partial \beta_0 \partial \beta_p} & \frac{\partial^2 L(\beta)}{\partial \beta_1 \partial \beta_p} & \cdots & \frac{\partial^2 L(\beta)}{\partial \beta_p^2} \end{bmatrix}$$

Prosedur Newton Raphson untuk mencari taksiran b_j , j = 0,1,2,...,p yaitu dengan langkah-langkah sebagai berikut :

1. Pilih taksiran awal b_{jm}, misalkan b_{j1} =0 ,m = 1,2,3,...
2. Pada setiap iterasi ke (m+1) hitung taksiran baru.

$$b_{j(m+1)} = b_{jm} + \left[\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u}\right]^{-1} \left[\frac{\partial L(\beta)}{\partial \beta}\right]$$

3. Iterasi berlanjut hingga diperoleh $b_{j(m+1)} \approx b_{jm}$