BAB II

TEORI PENUNJANG

Untuk mempermudah pembahasan mengenai fungsi link pada generalisasi model linier untuk data biner, diperlukan beberapa teori penunjang. Dalam bab ini akan dibahas teori penunjang mengenai probabilita, variabel random kontinu, fungsi distribusi, nilai harapan, percobaan Bernoulli, distribusi Binomial,metode maksimum likelihood (MLE), deviansi, distribusi keluarga eksponensial, dan model regresi linier berganda.

2.1 Probabilita

Definisi 1.

Jika sebuah percobaan ξ mempunyai ruang sampel S dan sebuah kejadian A didefinisikan pada S, maka P(A) adalah suatu angka riil yang disebut probabilita dari peristiwa A atau probabilita A, dan fungsi P(.) mempunyai syarat-syarat sebagai berikut :

1. $0 \leq P(A) \leq 1$ untuk setiap kejadian A dari S
2. $P(S) = 1$

Definisi 2.

Bila suatu percobaan dapat menghasilkan n(S) banyaknya hasil yang mempunyai kemungkinan yang sama dalam A dan bila n(A) menunjukkan jumlah hasil-hasil yang diperoleh dalam kejadian A, maka probabilita A adalah $P(A) = \frac{n(A)}{n(S)}$

Teorema 1.

Jika $\overline{A}$ merupakan komplemen dari A, maka $P(\overline{A}) = 1 - P(A)$

Bukti

Diketahui bahwa $S = A \cup \overline{A}$, sehingga $P(S) = P(A \cup \overline{A})$.

Oleh karena A dan $\overline{A}$ saling bebas maka $P(S) = P(A) + P(\overline{A})$, dari definisi 1 $P(S) = 1$, maka $P(A) + P(\overline{A}) = 1$ sehingga diperoleh $P(\overline{A}) = 1 - P(A)$.

(Hines and Montgomery,1990)

2.2 Variabel Random Kontinu

Definisi 3.

Jika ruang hasil R_x, variabel random X adalah sebuah interval atau kumpulan dari interval-interval, maka X disebut sebuah variabel random kontinu.

Definisi 4.

Untuk sebuah variabel random kontinu X, didefinisikan :

$$P(a \le X \le b) = \int_a^b f(x)dx$$

dengan f(x) dinamakan sebagai fungsi densitas probabilitas, memenuhi kondisi-kondisi sebagai berikut :

1. $f(x) \ge 0$ untuk seluruh $x \in R_x$

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

(Hines and Montgomery,1990)

2.3 Fungsi Distribusi

Definisi 5.

Fungsi distribusi variabel random X dinotasikan sebagai F(x) dan didefinisikan sebagai $F(x) = P(X \leq x)$ untuk seluruh x yang riil.

1. Jika X adalah diskrit, maka $F(x) = \sum_{x_i \leq x} p(x_i)$
2. Jika X adalah kontinu, maka $F(x) = \int_{-\infty}^{x} f(t)\,dt$

Fungsi distribusi mempunyai sifat-sifat sebagai berikut :

1. $\lim_{x \to \infty} F(x) = 1$
2. $\lim_{x \to -\infty} F(x) = 0$
3. Fungsi tersebut adalah tidak turun; yaitu jika $b \geq a$, maka $F(b) \geq F(a)$
4. Fungsi tersebut kontinu dari kanan, yaitu untuk seluruh x dan $\delta > 0$,

   $\lim_{\delta \to 0}[F(x + \delta) - F(x)] = 0$

(Hines and Montgomery,1990)

2.4 Nilai Harapan

Definisi 6.

Jika X adalah suatu variabel random dengan fungsi probabilitas f(x), maka nilai harapan dari X yang dinotasikan E(X) adalah :

$$E(X) = \sum_i x_i \cdot f(x_i) \quad \text{untuk X yang diskrit}$$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx \quad \text{untuk X yang kontinu} \quad \text{(Walpole and Myers, 1995)}$$

Teorema 2.

Andaikan a dan b adalah konstanta maka :

$$E(aX + b) = aE(X) + b$$

Bukti :

Andaikan X variabel random yang kontinu, maka sesuai dengan definisi

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b)f(x)dx$$

$$= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx$$

$$= aE(X) + b.1 = aE(X) + b$$

Teorema ini berlaku untuk variabel random kontinu maupun variabel random diskrit. Untuk X variabel yang diskrit, pembuktian dilakukan dengan cara yang sama sesuai dengan definisi nilai harapan untuk variabel random diskrit. Dari teorema ini dapat diturunkan sifat-sifat sebagai berikut :

a. Jika k adalah konstanta maka E(k) = k. Hal ini diperoleh jika pada teorema diambil nilai a = 0.

b. Jika k adalah konstanta dan f(x) menyatakan fungsi probabilitas maka

$E[kf(x)] = kE[f(x)]$. Hal ini diperoleh jika pada teorema diambil nilai b = 0.

c. Jika μ merupakan nilai rata-rata, maka $E(X - \mu) = 0$

(Hogg and Craig,1995)

2.4.1. Ragam

Definisi 7.

Ragam digunakan untuk mengukur variabilitas suatu distribusi kemungkinan dan dinotasikan σ^2. Untuk μ(nilai rata-rata) yang diketahui, Ragam didefinisikan sebagai:

$$\sigma^2 = \sum_{i=1}^{\infty}(x_i - \mu)^2 p(x_i) \qquad \text{untuk X yang diskrit}$$

$$\sigma^2 = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)dx \qquad \text{untuk X yang kontinu}$$

Sedangkan apabila μ tidak diketahui maka Ragam didefinisikan sebagai :

$$\sigma^2 = \sum_{i=1}^{\infty}(x_i - \hat{\mu})^2 p(x_i) \qquad \text{untuk X yang diskrit}$$

$$\sigma^2 = \int_{-\infty}^{\infty}(x - \hat{\mu})^2 f(x)dx \qquad \text{untuk X yang kontinu}$$

dengan $\hat{\mu}$ merupakan nilai taksiran dari parameter μ.

(Hines and Montgomery, 1990)

Pada umumnya ragam dari suatu variabel random X ditulis Var (X). Misalkan X variabel random dengan nilai harapan E(X), maka Var (X) didefinisikan

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2$$

Teorema 3.

Misalkan X dan Y dua variabel random yang saling bebas, maka

$$E(XY) = E(X)E(Y)$$ (Walpole and Myers, 1995)

2.5 Percobaan Bernoulli

Definisi 8.

Sebuah percobaan yang hanya menghasilkan 2 kejadian yaitu sukses dan gagal disebut percobaan Bernoulli. Misalkan E sebagai kejadian sukses dan komplemennya yaitu E' sebagai kejadian gagal, bila E terjadi maka P (E) = π dan bila E' terjadi maka P (E') = 1-π. Bila Z merupakan variabel random yang hanya mempunyai nilai 0 atau 1 dinamakan variabel Bernoulli, yaitu

$$Z = \begin{cases} 1 & \text{jika kejadian sukses} \\ 0 & \text{jika kejadian gagal} \end{cases}$$

Jika probabilita sukses π, dan probabilita gagal 1-π yang dapat ditulis f(1)=P(Z=1)= π dan f(0)=P(Z=0)= 1-π, maka fungsi probabilitanya adalah

$$f(z;\pi) = \pi^z(1-\pi)^{1-z}, \quad z = 0, 1$$

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

2.6 Distribusi Binomial

Definisi 9.

Distribusi Binomial dapat dipandang sebagai n variabel random Bernoulli yang independent, yaitu banyaknya kejadian sukses dalam n percobaan Bernoulli. Misalkan Y sebagai banyaknya kejadian sukses dalam n percobaan Bernoulli, maka fungsi probabilitanya adalah

$$f(y) = \Pr(Y = y) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y} \qquad y = 0, 1, \ldots, n$$

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

2.7 Metode Maksimum Likelihood (MLE)

Misalkan X variabel random dengan distribusi probabilita $f(x,\beta)$, di mana parameter tunggal β tidak diketahui. Misalkan X_1, X_2, ..., X_n menjadi nilai yang diobservasi di dalam suatu sampel random yang berukuran n, maka fungsi likelihood sampel tersebut adalah

$$L(\beta) = f(x_1,\beta)\cdot f(x_2,\beta)\cdot \ldots \cdot f(x_n,\beta)$$

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

Estimator maksimum likelihood β adalah nilai β yang memaksimumkan fungsi likelihood L(β). Jadi jika $L(\beta) = f(x_1, x_2, \ldots, x_n \mid \beta)$ maka nilai $\hat{\beta}$ yang harus dicari harus memenuhi $f(x_1, x_2, \ldots, x_n \mid \hat{\beta}) = \max f(x_1, x_2, \ldots, x_n \mid \beta)$

Harga-harga β yang memaksimumkan $L(\beta)$ juga akan memaksimumkan logaritma likelihood yang dinotasikan $l(\beta) = \ln L(\beta)$, tetapi karena fungsi logaritma likelihood monoton naik, maka digunakan turunan parsial pertama yaitu

$$\frac{\partial}{\partial\beta} L(\beta) = 0$$

sehingga persamaan likelihood adalah

$$\frac{\partial}{\partial\beta} l(\beta) = 0 \tag{2.7.1}$$

Selanjutnya, apabila hasil turunan parsial pertama dari persamaan likelihood tidak linier dalam parameter β, maka untuk mengatasinya digunakan metode iterasi numerik yaitu metode iterasi scoring Fisher, diperoleh estimator maksimum likelihood untuk β yaitu $\hat{\beta}$.

2.8 Deviansi (Statistik RasioLog-likelihood)

Didefinisikan statistik rasio log-likelihood sebagai

$$D = 2\left[\; l(\hat{\pi}_{max}\,;\, y) - l(\hat{\pi}\,;\, y) \;\right] \tag{2.8.1}$$

yang disebut dengan deviansi, yang mempunyai nilai pendekatan $\chi^2_{1-\alpha;\,N-p-1}$

Dengan $l(\hat{\pi}_{max}\,;\, \mathbf{y})$ merupakan fungsi log-likelihood untuk model maksimal dan $l(\hat{\pi}\,;\, \mathbf{y})$ merupakan fungsi log-likelihood untuk model yang diperhatikan dengan $\hat{\pi}_{max}$ dan $\hat{\pi}$ sebagai vektor taksiran maksimum likelihood yang berhubungan dengan model maksimal dan vektor taksiran maksimum likelihood untuk model yang diperhatikan (Dobson, 1990)

2.9 Distribusi Keluarga Eksponensial

Misalkan Y variabel random yang memiliki fungsi probabilitas jika diskrit atau fungsi densitas probabilitas jika kontinu, yang tergantung pada sebuah parameter θ dari distribusi keluarga eksponensial yang ditulis dalam bentuk :

$$f(y;\theta) = s(y)t(\theta)\exp[a(y)b(\theta)] \tag{2.9.1}$$

dengan a, b, s, dan t adalah fungsi-fungsi yang diketahui. Jika s(y) = exp (d(y)) dan t(θ) = exp (c(θ)), maka persamaan (2.9.1) dapat dituliskan sebagai berikut :

$$\begin{aligned} f(y;\theta) &= \exp[d(y)]\cdot\exp[c(\theta)]\cdot\exp[a(y)b(\theta)] \\ &= \exp[a(y)b(\theta) + c(\theta) + d(y)] \end{aligned} \tag{2.9.2}$$

Jika a(y) = y, distribusi dalam persamaan (2.9.2) dikatakan distribusi dalam bentuk kanonik dan b(θ) disebut parameter natural dari distribusi.

Sedangkan nilai harapan dan variansi dari a(y) adalah :

$$E[a(Y)] = E(Y) = \frac{-c'(\theta)}{b'(\theta)}, \text{ dan}$$

$$\mathrm{Var}[a(Y)] = \mathrm{Var}(Y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

dengan $b'(\theta)$ = derivative pertama dari $b(\theta)$ dan $b''(\theta)$ = derivative kedua dari $b(\theta)$

$c'(\theta)$ = derivative pertama dari $c(\theta)$ dan $c''(\theta)$ = derivative kedua dari $c(\theta)$

Dalam keluarga eksponensial, ada beberapa distribusi yang termasuk di dalamnya yaitu distribusi Binomial, Poisson, Normal, dan Gamma.

Secara umum, jika terdapat N variabel random $Y_1, \ldots, Y_N$ yang saling bebas, mempunyai distribusi yang sama yaitu dalam kerluarga eksponensial, maka fungsi densitas bersamanya adalah :

$$f(y_1,\ldots,y_N;\theta)=\prod_{i=1}^{N}\exp[a(y_i)b(\theta)+c(\theta)+d(y_i)]$$

$$=\exp\left[b(\theta)\sum_{i=1}^{N}a(y_i)+Nc(\theta)+\sum_{i=1}^{N}d(y_i)\right] \qquad (2.9.3)$$

(Dobson, 1990)

Contoh : Misalkan variabel random Y adalah banyaknya sukses dalam n percobaan yang saling bebas, dengan probabilitas sukses π yang sama dalam semua percobaan. Maka Y berdistribusi binomial $b(n,\pi)$ dengan fungsi probabilitasnya

$$f(y;\pi)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y} \qquad y=0,1,\ldots,n$$

Fungsi probabilitas di atas juga dapat ditulis dalam bentuk

$$f(y;\pi)=\exp\left[y\log\pi-y\log(1-\pi)+n\log(1-\pi)+\log\binom{n}{y}\right]$$

$$=\exp\left[y\log\left(\frac{\pi}{1-\pi}\right)+n\log(1-\pi)+\log\binom{n}{y}\right]$$

Persamaan di atas mempunyai bentuk yang sama dengan persamaan (2.7.2), dengan komponen-komponennya

$$b(\pi)=\log\left(\frac{\pi}{1-\pi}\right) \qquad c(\pi)=n\log(1-\pi) \qquad d(y)=\log\binom{n}{y}$$

Maka dapat diperoleh nilai harapan dan variansi dari variabel random Y, yaitu

$$E(Y)=\frac{-c'(\pi)}{b'(\pi)}=\frac{-\left(\dfrac{-n}{1-\pi}\right)}{\dfrac{1}{\pi(1-\pi)}}=n\pi$$

$$\mathrm{Var}(Y)=\frac{b''(\pi)c'(\pi)-c''(\pi)b'(\pi)}{[b'(\pi)]^3}=\frac{\frac{2\pi-1}{\pi^2(1-\pi)^2}\cdot\left(\frac{-n}{1-\pi}\right)-\left(\frac{-n}{(1-\pi)^2}\right)\cdot\frac{1}{\pi(1-\pi)}}{\left[\frac{1}{\pi(1-\pi)}\right]^3}$$

$$=\frac{\frac{n-2n\pi}{\pi^2(1-\pi)^3}-\left(\frac{-n}{\pi(1-\pi)^3}\right)}{\left[\frac{1}{\pi(1-\pi)}\right]^3}=\frac{\frac{n-2n\pi+n\pi}{\pi^2(1-\pi)^3}}{\left[\frac{1}{\pi(1-\pi)}\right]^3}$$

$$\mathrm{Var}(Y)=n\pi(1-\pi)$$

2.10 Model Regresi Linier Berganda

Dalam beberapa masalah terdapat dua atau lebih variabel yang hubungannya tidak dapat dipisahkan., dan hal tersebut biasanya diselidiki sifat hubungannya. Analisis regresi adalah sebuah teknik statistik untuk membuat model dan menyalidiki hubungan antara dua variabel atau lebih. Hubungan antara variabel-varibel ini digolongkan dengan sebuah model secara matematik yang disebut sebuah persamaan regresi atau model regresi. Sebuah model yang mencakup lebih dari satu variabel prediktor disebut model regresi linier berganda, yang mempunyai bentuk persamaan

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+\varepsilon$$

persamaan regresi di atas dapat disebut sebagai model regresi berganda dengan p variabel bebas (prediktor). Variabel y disebut sebagai variabel tak bebas (respon), parameter $\beta_j,\ j=0,1,2,\ldots,p$ disebut koefisien regresi. Parameter β_j menyatakan rata-rata perubahan y per unit akibat perubahan x_j bila seluruh sisa variabel-variabel prediktornya x_i $(i \neq j)$ konstan, dan ε merupakan galat (error) yang diasumsikan mempunyai rataan 0 (nol) dan ragam σ^2. Jika model regresi berganda ini ditulis ke dalam bentuk matrik adalah sebagai berikut :

$$y = x\,\beta + \varepsilon$$

dengan y = variabel tak bebas (respon) dalam vektor matrik ukuran n x 1

x = variabel bebas (prediktor) dalam vektor matrik ukuran n x p

β = parameter yang akan diestimasi dalam vektor matrik ukuran p x 1

ε = galat (residual) dalam vektor matrik ukuran n x 1

Untuk mengestimasi parameter salah satu metode yang dipergunakan adalah metode kuadrat terkecil. Misalkan observasi $n > p$ yang tersedia, dan misalkan x_{ij} menyatakan observasi ke-i atau tingkat variabel x_j, dengan $j = 1, \ldots, p$, maka data tersebut adalah

y	x_1	x_2	…	x_p
y_1	x_{11}	x_{12}	…	x_{1p}
y_2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
y_n	x_{n1}	x_{n2}	…	x_{np}

Dari data di atas dapat ditulis suatu model sebagai berikut :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,2,\ldots,n$$

Fungsi kuadrat terkecil adalah :

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

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

Fungsi L tersebut diminimumkan terhadap $\beta_0, \beta_1, \beta_2, ..., \beta_p$. Estimator kuadrat terkecil $\beta_0, \beta_1, \beta_2, ..., \beta_p$ harus memenuhi

$$\frac{\partial L}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) = 0$$

dan

$$\frac{\partial L}{\partial \beta_j}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) x_{ij} = 0$$

Penyederhanaan persamaan di atas dapat diperoleh persamaan-persamaan kuadrat terkecil

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} x_{i1} + \hat{\beta}_2\sum_{i=1}^{n} x_{i2} + ... + \hat{\beta}_p\sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0\sum_{i=1}^{n} x_{i1} + \hat{\beta}_1\sum_{i=1}^{n} x_{i1}^{\,2} + \hat{\beta}_2\sum_{i=1}^{n} x_{i1}x_{i2} + ... + \hat{\beta}_p\sum_{i=1}^{n} x_{i1}x_{ip} = \sum_{i=1}^{n} x_{i1}y_i$$

$$.... \qquad ... \qquad ... \qquad .... \qquad ....$$

$$\hat{\beta}_0\sum_{i=1}^{n} x_{ip} + \hat{\beta}_1\sum_{i=1}^{n} x_{ip}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n} x_{ip}x_{i2} + ... + \hat{\beta}_p\sum_{i=1}^{n} x_{ip}^{\,2} = \sum_{i=1}^{n} x_{ip}y_i$$

Dalam persamaan tersebut, terlihat ada p + 1 persamaan normal, satu untuk setiap koefisien regresi yang tidak diketahui. Penyelesaian untuk persamaan normal menjadi estimator-estimator kuadrat tekecil dari koefisien-koefisien regresi $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$.

Hal ini lebih sederhana menyelesaikan persamaan normal jika digunakan notasi matrik. Jika model regresi linier berganda ditulis ke bentuk matrik

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon$$

dengan

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

y adalah sebuah vektor n x 1 dari observasi-observasi, **X** adalah sebuah matrik n x (p+1) dari variabel-variabel bebas, β adalah sebuah vektor (p+1) x 1 dari koefisien-koefisien regresi, dan ε adalah sebuah vektor n x 1 dari galat acak.

Untuk mendapatkan estimator vektor kuadrat terkecil, $\hat{\beta}$, yang minimum

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

L juga dapat ditulis sebagai

$$\begin{aligned} L &= \mathbf{y}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \end{aligned}$$

L diturunkan secara parsial ke-β sehingga didapat :

$$\frac{\partial L}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

yang disederhanakan menjadi

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

dan dengan mengalikan kedua ruas dengan invers $(\mathbf{X}'\mathbf{X})$, diperoleh :

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

Sehingga perkiraan model regresi tersebut adalah

$$\hat{y} = \mathbf{X}\hat{\beta}$$

Dalam bentuk skalar, perkiraan model tesebut adalah

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \qquad i = 1,2,\ldots,n$$

(Hines and Montgomery, 1990)