BAB III

MODEL LINEAR TERGENERALISASI

3.1 Model Linear

Perkembangan pemodelan stokastik, terutama model linier, dapat dikatakan dimulai pada abad ke 19 yang didasari oleh teori matematika yang dijelaskan diantaranya oleh Gauss, Boole, Cayley dan Sylvester yang terkait dengan teori invarian dalam aljabar. Teori invarian aljabar mempelajari bentuk-bentuk kuantitas yang tidak berubah terhadap suatu transformasi linear. Teori invarian ini yang mendasari perkembangan teori nilai eigen, vektor eigen, matriks determinan, metode dekomposisi dan masih banyak lagi yang lainnya. Salah satu contoh dalam statistika adalah korelasi dua peubah acak tidak berubah walaupun peubah-peubah tersebut mengalami transformasi.

Perkembangan model linear dimulai dengan perkembangan analisis regresi pada abad 19 oleh Pearson, dilanjutkan dengan perkembangan korelasi. Teori regresi ini yang menjadi dasar perkembangan teori model linear. Perkembangan model linear tidak bisa dilepaskan dengan perkembangan teori matriks atau aljabar linear. Melalui teori matriks (determinan, invers, perkalian matriks) pembahasan model linear dapat didekati secara umum. Dalam pembahasan ini perkembangan model linear lebih dititikberatkan pada dua asumsi dasar, yaitu distribusi dan independensi dari kesalahan. Sebagaimana diuraikan sebelumnya, bahwa pemodelan dimulai dari yang sederhana, yang secara matematis mudah diselesaikan, kemudian berkembang ke arah

yang lebih realistik. Hal ini dapat dilakukan dengan menerapkan berbagai asumsi yang berbeda terhadap distribusi kesalahan dalam model yang digunakan. Prinsip seperti ini telah berkembang dari model yang paling sederhana (klasik), ke model hirarkis tergeneralisasi yang saat ini merupakan pemodelan yang paling terkini. Dalam sub-bab ini diuraikan secara ringkas perkembangan model linier ditinjau dari segi distribusi dan independensi kesalahannya.

3.2 Model Linear Klasik

Pemodelan stokastik memiliki bentuk umum:

$$Y = X\beta + \varepsilon_i \qquad ...(3.2.1)$$

Dalam hal ini ε_i merupakan kesalahan atau galat yang diasumsikan merupakan peubah acak yang berasal dari suatu distribusi tertentu, misalnya normal. Peubah x adalah peubah yang bukan acak dan adalah parameter yang menentukan koefisien dari peubah peubah tetap tadi. Misalnya dalam perdagangan, dianggap bahwa sebenarnya ada hubungan yang bersifat tetap yang menentukan harga barang di pasar. Namun, selain itu masih ada lagi faktor lain yang bersifat acak yang menyebabkan harga barang tadi dalam kenyataannya dari pembeli ke pembeli mungkin menyimpang dari fungsi hubungan yang ada.

Dalam pemodelan statistika/ stokastik, kedua komponen ini (peubah acak dan peubah tetap) dipisahkan yaitu yang bersifat tetap dan fungsional dinotasikan dengan f(x, β), yang biasa disebut sebagai komponen tetap *(fixed)*, sedangkan komponen

lainnya, ε, yang bersifat acak disebut komponen acak *(random component)* atau komponen kesalahan *(error component).*

Dari segi fungsi hubungan *f*, bentuk yang paling sederhana adalah hubungan linear, sehingga dari aspek ini model yang paling sederhana yang dimiliki adalah model linier. Sedangkan dari segi komponen acaknya, yang paling sederhana adalah asumsi bahwa kesalahannya berdistribusi normal dan saling independen antara satu respon dengan respon lainnya.

Asumsi ini menghasilkan model linear normal sederhana atau *Normal Linear Models (NLM)*. Dari kedua hal tersebut lahirlah yang disebut model normal sederhana atau model linear klasik yang secara formal dapat diuraikan sebagai berikut.

Definisi 3.1: (Tirta, 2005: 177)

(Bentuk dan Asumsi Model Linear Klasik).

Model: $$y_i = \sum_{j=0}^{k} x_{ij} + \varepsilon \qquad ...(3.2.2)$$

atau untuk keseluruhan respon dapat dituliskan dalam bentuk matriks seperti persamaan (3.2.1)

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad ...(3.2.3)$$

Asumsi: x_i bukan peubah acak dan diukur tanpa kesalahan dan ε_i independen dengan $\varepsilon_i^{'}$ untuk setiap $i \neq i^{'}$ dan masing-masing berdistribusi $N\left(0, \sigma^2\right)$.

Dari asumsi di atas diperoleh bahwa secara keseluruhan ε dapat dianggap berdistribusi multivariat normal (MVN) dengan koefisen variasi konstan, yang

dinotasikan dengan $\varepsilon \sim MVN(0, \sigma^2 I)$. Model mengisyaratkan bahwa respon ke i dan ke i' adalah saling bebas (independen), yang berarti tidak ada korelasi diantaranya.

3.3 Model Linear Tergeneralisasi

Kondisi lain di lapangan yang tidak dapat ditangani langsung oleh model linear klasik adalah adanya kenyataan bahwa, distribusi respon tidak mesti normal. Memang kondisi seperti ini bisa ditanggulangi dengan mengadakan transformasi dari respon. Transformasi yang banyak dipakai adalah transformasi logaritma. Namun, ada beberapa permasalahan yang mungkin timbul sebagai efek dari transformasi ini misalnya seperti berikut ini. Respon yang sudah ditransformasi mungkin mendekati distribusi normal, tetapi akibat transformasi ada kemungkinan syarat yang lain (syarat ketidak-bergantungan) menjadi tidak terpenuhi. Adanya kerancuan dalam menafsirkan hasil penelitian oleh karena efek yang diuji adalah dalam skala logaritma, bukan dalam sekala aslinya. Hal ini menyebabkan kesimpulan terasa janggal misalnya, ”ada hubungan positif antara log-konsentrasi pemupukan dengan log-panen” (Tirta, 2005: 178).

Untuk menangani kondisi dimana respon yang ada tidak berdistribusi Normal, tetapi masih saling bebas, maka para statistisi yang dipelopori oleh Nelder dan Wedderburn (1972) telah mengembangkan model linear yang dikenal dengan *Gereralized Linear Model (GLM)*. Model linear ini menggunakan asumsi bahwa repon memiliki distribusi keluarga ekponensial. Distribusi keluarga eksponensial adalah distribusi yang sifatnya lebih umum, dimana distribusi- distribusi yang banyak

kita kenal (Normal, Gamma, Poisson) termasuk di dalamnya dan merupakan bentuk-bentuk khusus dari distribusi Keluarga Eksponensial.

Model Linear Tergeneralisasi pada dasarnya merupakan model regresi. Seperti semua model regresi, model ini terbuat dari komponen acak (yang biasanya disebut dengan eror) dan fungsi dari faktor desain (x) dan beberapa parameter (β). Dalam teori normal baku, model regresi linear berganda dituliskan sebagai berikut:

$$y = \beta_0 + \beta_i x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad ...(3.3.1)$$

Dimana bentuk eror ε diasumsikan bedistribusi normal dengan rerata 0 dan varians konstan. Rerata dari variabel respon y adalah:

$$E(y) = \mu = \beta_0 + \beta_i x_1 + \beta_2 x_2 + ... + \beta_k x_k = x'\beta \quad ...(3.3.2)$$

Model linier mempunyai beberapa hal yang sifatnya khas dan istimewa yaitu:

1) ada komponen tetap yang disebut prediktor linier

2) respon y_i berdistribusi normal dan saling independen dan

3) rerata y_i adalah $\mu_i = \sum_{j=0}^{k} X_{ij}\beta_j$

Dalam model linear tergeneralisasi, hubungan di atas mengalami perubahan atau generalisasi, sebagaimana dalam definisi berikut:

Definisi 3.2 (Tirta, 2005: 178)

Asumsi Model Linear Tergeneralisasi

Model linier tergeneralisasi adalah model yang mengandung tiga hal yaitu:

1) Komponen tetap yang disebut prediktor linier $\eta_i = \sum_{j=0}^{k} x_{ij}\beta_i$ Prediktor linear, dinotasikan dengan η_i, dari bentuk model linear tergeneralisasi, yaitu: $\eta_i = x_i^{'}\beta$ dimana x_i adalah vektor regresi untuk unit sebanyak i dengan *fixed effect* β
2) Respon y_i berdistribusi secara independen dalam keluarga eksponensial
3) Hubungan antara mean dengan prediktor linear ditunjukkan oleh fungsi g(.) yang disebut fungsi 'link' sedemikian sehingga g(μ_i) = η_i. Fungsi g() disebut fungsi hubungan *(link-function).*

Ada fungsi hubungan khusus yang disebut fungsi hubungan kanonik atau natural yang berkaitan erat dengan distribusi y. Misalnya, jika distribusinya normal maka g(.) adalah identitas. Dari hal di atas dikatakan bahwa komponen penting dalam model linear tergeneralisasi ada tiga yaitu:

1) adanya prediktor linear,
2) adanya distribusi keluarga eksponensial dan
3) adanya fungsi-hubungan.

3.3.1 Keluarga Eksponensial

Berikut ini diberikan catatan sederhana mengenai properti dari keluarga eksponensial. Misalkan $\ell_i = \ln f(y_i, \theta_i, \alpha_i)$

$$= \alpha_i\{\theta_i y_i - a(\theta_i) + b(y_i)\} - c(\alpha_i - y_i) \qquad ...(3.3.1.1)$$

$$S_i = \frac{\partial \ell_i}{\partial \theta_i} = \alpha_i\{y_i - a^{'}\theta_i\} \qquad ...(3.3.1.2)$$

$$\frac{\partial^2 \ell_i}{\partial \theta_i^2} = -\alpha_i a''(\theta_i) \qquad ...(3.3.1.3)$$

S_i disebut dengan fungsi skor (*score function*) dan memiliki properti yang menarik. Properti dari fungsi skor yang akan dibahas berikut ini akan sangat dibutuhkan dalam analisis statistik yang dilakukan.

- $E(S_i) = E\left(\frac{\partial \ell_i}{\partial \theta_i}\right) = 0$ ...(3.3.1.4)
- $E\left(\frac{\partial^2 \ell_i}{\partial \theta_i^2}\right) + E\left(\frac{\partial \ell_i}{\partial \theta_i}\right)^2 = 0$ ...(3.3.1.5)

Berdasarkan (3.3.1.4) dan (3.3.1.2) diperoleh bahwa:

$$E(y_i) = \mu_i = a'\theta_i \qquad ...(3.3.1.6)$$

dan bisa dituliskan bahwa $S_i = \alpha_i(y_i - \mu_i)$. Dari persamaan (3.3.1.3) dan (3.3.1.5) diperoleh bahwa $\text{var}(y_i) = a''(\theta_i) / a_i$

Definisi 3.3: (Tirta 2007: 2)

Suatu peubah acak Y dengan fungsi kepadatan peluang (fkp) f dan parameter θ dikatakan menjadi anggota distribusi keluarga eksponensial, jika f dapat dinyatakan sebagai:

$$f(y,\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \qquad ...(3.3.1.7)$$

Dalam keadaan khusus $a(y) = y$, maka (3.3.1.7) menjadi:

$$f(y) = \exp[yb(\theta) + c(\theta) + d(y)] \qquad ...(3.3.1.8)$$

dan persamaan (3.3.1.8) disebut dengan bentuk kanonik dari distribusi keluarga eksponensial dan $b(\theta)$ disebut parameter alami/natural dari distribusinya.

3.3.1.1 Fungsi Skor $[U]$, $E[U]$, dan $Var[U]$

Fungsi skor dari $f(y)$ terhadap θ didefinisikan sebagai $U = dl(y)/d\theta$, dengan $l(y) = \log f(y) = \ln f(y)$. Perhitungan $E[U]$ dan $Var[U]$ dibutuhkan untuk menurunkan rerata dan varians Y atau dalam bentuk yang lebih umum, $E[a(Y)]$, dan $Var[a(Y)]$

$$U = \frac{dl(y)}{d\theta} \qquad \text{...(3.3.1.1.1)}$$

$$= \frac{1}{f(y)} \frac{df(y)}{d(\theta)} \qquad \text{...(3.3.1.1.2)}$$

Dengan demikian

$$\begin{aligned} U[U] &= \int \frac{1}{f(y)} \frac{df(y)}{d(\theta)} f(y) d(y) \\ &= \int \frac{df(y)}{d\theta} d(y) \\ &= \frac{d}{d\theta} \int f(y) d(y) \qquad \text{...(3.3.1.1.3)} \\ &= \frac{d1}{d\theta} \\ &= 0 \end{aligned}$$

Berdasarkan persamaan (3.3.1.1.1) dan (3.3.1.1.2) diperoleh:

$$\frac{df(y)}{d(\theta)} = f(y) \frac{dl(y)}{d(\theta)} \qquad \text{...(3.3.1.1.4)}$$

Selanjutnya akan ditujukkan bahwa $E[U'] + E[U^2] = 0$

$$E\left[U'\right] = E\left(\frac{dU}{d\theta}\right) = \frac{d}{d\theta}E[U] \qquad \text{...(3.3.1.1.5)}$$

$$= \frac{d\,0}{d\theta} = 0 \qquad \text{...(3.3.1.1.6)}$$

Tetapi dari (3.3.1.1.4), ruas kanan dari (3.3.1.1.5) menjadi $\frac{d}{d\theta}\int \frac{dl(y)}{d\theta} f(y)dy$. Jadi bersama dengan (3.3.1.1.4) menghasilkan:

$$\begin{aligned} 0 &= \frac{d}{d(\theta)}\int \frac{dl(y)}{d\theta} f(y)d(y) \\ &= \int \frac{d^2 l(y)}{d\theta^2} f(y)dy + \int \frac{dl(y)}{d\theta}\frac{df(y)}{d\theta}dy \\ &= \int \frac{d^2 l(y)}{d\theta^2} f(y)dy + \int \left(\frac{dl(y)}{d\theta}\right)^2 f(y)dy \\ &= \int U' f(y)dy + \int U^2 f(y)dy \\ &= E\left[U'\right] + E\left[U^2\right] \end{aligned}$$

Jadi,

$$\mathrm{E}\left[-U'\right] = E\left[U^2\right]$$

dan,

$$Var(U) = E\left[-U'\right] \qquad \text{...(3.3.1.1.7)}$$

Untuk persamaan (3.3.1.7), U dan U' terhadap θ adalah

$$\begin{aligned} U &= \frac{d}{d\theta}\left[a(y)b(\theta) + c(\theta) + d(y)\right] \\ &= a(y)b'(\theta) + c'(\theta) \end{aligned} \qquad \text{...(3.3.1.1.8)}$$

dan,

$$U' = a(y)b''(\theta) + c''(\theta) \qquad \text{...(3.3.1.1.9)}$$

Teorema 3.1 (Tirta, 2007: 7)

Jika rerata dan varians *a(Y)* yang didefinisikan seperti pada Definisi 3.3 maka rarata dan varians masing-masing adalah:

$$E[a(Y)] = -\frac{c'\theta'}{b'\theta} \qquad \text{...(3.3.1.1.10)}$$

$$\text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \qquad \text{...(3.3.1.1.11)}$$

3.4 Fungsi Hubungan (*Link Function*)

Fungsi yang meghubungkan komponen sistematis η terhadap nilai rerata μ (nilai harapan dari komponen acak) dinamakan dengan fungsi hubungan (*link function*).

$$\eta = h(\mu),\ \mu = h^{-1}(\eta) \qquad \text{...(3.4.1)}$$

Berdasarkan persamaan (3.3.1.6) dapat diperoleh bahwa,

$$h^{-1}(\eta) = a'(\theta) \qquad \text{...(3.4.2)}$$

$$\theta = (a')^{-1}\{h^{-1}(\eta)\} = g(\eta) = g(X\beta) \qquad \text{...(3.4.3)}$$

Ada beberapa pilihan yang mungkin untuk *h*. Berikut ini diberikan pilihan yang paling sering digunakan,

a. Fungsi Identitas, yakni $\eta = \mu$

b. Hubungan *logit*, $\eta = \ln\frac{\mu}{1-\mu}$

c. *Probit*, $\eta = \Phi^{-1}(\mu)$, dimana Φ kepadatan normal kumulatif, $0<\mu\leq 1$

d. *Log-log link,* $\eta = \ln[-\ln(1-\mu)]$, $0 < \mu < 1$, dan

e. Hubungan power keluarga (*power family link*), $\eta = \mu^2$, jika , $\gamma \neq 0$ dan

$\eta = \log \mu$, jika $\gamma = 0$

Fungsi hubungan dimana $\theta = \mu$ disebut dengan fungsi hubungan kanonik (*canocical link function*).

3.5 Penaksiran Parameter Untuk Model Linear Tergeneralisasi

Penaksiran kemungkinan maksimum dari parameter β akan diturunkan dari yang terdapat dalam prediktor linear η. Perhatikan bahwa

$$L_i = \alpha_i \{ g(X_i^T \beta) y_i - a(g(X_i^T)) + b(y_i) + b(y_i) + c(\alpha_i, y_i) \} \quad (3.5.1)$$

dimana $\theta_i \ln \ell_i$ telah diganti oleh persamaan (3.4.3). Fungsi *log* kemungkinanya adalah

$$L = \sum_{i=1}^{n} \ell_i \quad (3.5.2)$$

Dari penaksiran kemungkinan mamksimum (*Maximum Likelihood Estimation (MLE)*) diperoleh:

a. $\dfrac{\partial L}{\partial \beta} = 0$ dan $\dfrac{\partial}{\partial \beta}\left(\dfrac{\partial L}{\partial \beta}\right) = 0$

b. $\dfrac{\partial}{\partial \beta}\left(\dfrac{\partial L}{\partial \beta}\right) = H < 0$

Dalam penurunan kemungkinan maksimum yang perlu untuk diperhatikan adalah

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = 0 \qquad \text{...(3.5.3)}$$
$$j = 1,2,\ldots,k$$

dan,

$$\sum_{i=1}^{n} x_{ij} d_i \alpha_i (y_i - \mu_i) = \sum_{i=1}^{n} x_{ij} d_i S_i = 0 \qquad \text{... (3.5.4)}$$

karena $\alpha_i > 0$, dimana $S_i = (y_i - \mu_i)$, dan $d_i = \frac{\partial \theta_i}{\partial \eta_i}$.

Persamaan yang telah diberikan pada (3.5.4) bisa dituliskan dalam notasi matriks seperti berikut ini:

$$\frac{\partial L}{\partial \beta} = X^T \Delta S \qquad (3.5.5)$$

dimana, $\Delta = diag(d_i)$ dan $S_i = (y_1 - \mu_1, y_2 - \mu_2, \ldots y_n - \mu_n)$.

Sekarang perhatikan bahwa

$$\begin{aligned} H &= \frac{\partial}{\partial \beta}\left(\frac{\partial L}{\partial \beta}\right) \\ &= \frac{\partial}{\partial \beta}\left(X^T \Delta S\right) \\ &= X^T \left\{ \Delta \frac{\partial S}{\partial \beta} + \frac{\partial \Delta}{\partial \beta} \right\} \end{aligned} \qquad (3.5.6)$$

dan perlu diingat juga bahwa

$$\frac{\partial S_i}{\partial \beta_j} = -\frac{\partial \mu_i}{\partial \beta_j}$$
$$= -\frac{\partial \mu_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \pi_i}{\partial \beta_j} \qquad (3.5.7)$$
$$= -V_i d_i X_{ij} \qquad i = 1,2,\ldots,n$$
$$j = 1,2,\ldots,k$$