

## BAB III

### MODEL REGRESI BINOMIAL NEGATIF UNTUK MENGATASI OVERDISPERSI PADA MODEL REGRESI POISSON

#### 3.1 Regresi Poisson

Regresi Poisson merupakan salah satu model regresi dengan variabel responnya tidak berasal dari distribusi normal. Variabel respon dalam regresi Poisson ini berasal dari data hitung yang diharapkan jarang terjadi, seperti kecacatan dalam barang produksi, error dalam perangkat lunak (*software*), dan zat polutan dalam lingkungan.

Fungsi peluang yang sering digunakan untuk data hitung adalah distribusi Poisson, yaitu:

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}; y = 0, 1, 2, 3 \dots \dots(3.1.1)$$

Dengan parameter  $\mu > 0$ . Model dari regresi Poisson dapat dituliskan sebagai berikut.

$$\hat{y}_i = E(Y_i) + \varepsilon_i; i = 1, 2, \dots n \dots(3.1.2)$$

Ekspektasi dari variabel responnya dapat ditulis

$$E(Y_i) = \mu_i \dots(3.1.3)$$

Dalam *Generalized Linear Model* (GLM), terdapat sebuah fungsi  $g$  yang menghubungkan rata-rata dari variabel responnya dengan sebuah prediktor linear, yaitu:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k = \mathbf{x}_i' \boldsymbol{\beta} \quad \dots(3.1.4)$$

Fungsi  $g$  biasanya disebut fungsi penghubung (*link function*). Hubungan antara rata-rata dan prediktor linear adalah sebagai berikut.

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}) \quad \dots(3.1.5)$$

Terdapat dua fungsi penghubung yang biasa digunakan dalam distribusi Poisson. *Pertama* adalah penghubung identitas (*identity link*). Kedua adalah penghubung log (*log link*). Fungsi penghubung identitas berbentuk:

$$g(\mu_i) = \mu_i = \mathbf{x}_i' \boldsymbol{\beta} \quad \dots(3.1.6)$$

Jika fungsi penghubung identitas ini digunakan, maka

$$E(Y_i) = \mu_i = \mathbf{x}_i' \boldsymbol{\beta} \text{ karena } \mu_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}) = (\mathbf{x}_i' \boldsymbol{\beta}) \quad \dots(3.1.7)$$

Fungsi penghubung selanjutnya adalah penghubung log yang berbentuk

$$g(\mu_i) = \ln(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta} \quad \dots(3.1.8)$$

Untuk fungsi penghubung log dalam persamaan (3.1.8), hubungan antara rata-rata variabel respon dan prediktor linear adalah sebagai berikut.

$$\begin{aligned} \ln(\mu_i) &= \mathbf{x}_i' \boldsymbol{\beta} \\ e^{\ln(\mu_i)} &= e^{\mathbf{x}_i' \boldsymbol{\beta}} \\ \mu_i &= e^{\mathbf{x}_i' \boldsymbol{\beta}}. \end{aligned} \quad \dots(3.1.9)$$

Fungsi penghubung log adalah fungsi penghubung yang lebih cocok digunakan karena fungsi log menjamin bahwa nilai yang diharapkan dari variabel responnya akan bernilai non negatif. Dalam pembahasan ini, fungsi penghubung yang dipilih adalah penghubung log.

## 6.2 Penaksir Parameter Regresi Poisson

Metode yang digunakan untuk menaksir parameter pada regresi Poisson adalah metode kemungkinan maksimum. Misalkan terdapat sampel acak yang terdiri atas  $n$  pengamatan pada variabel respon  $Y$  dan variabel prediktor  $X$ , maka fungsi kemungkinannya adalah sebagai berikut.

$$\begin{aligned}L(y, \beta) &= \prod_{i=1}^n f_i(y_i) \\ &= \prod_{i=1}^n \frac{\mu^{y_i} e^{-\mu_i}}{y_i!} \\ &= \frac{\prod_{i=1}^n \mu^{y_i} \exp(-\sum_{i=1}^n \mu_i)}{\prod_{i=1}^n y_i!}\end{aligned}$$

Setelah fungsi penghubung dipilih, dalam hal ini adalah fungsi penghubung log, akan dimaksimumkan fungsi log-kemungkinannya

$$\ln L(y, \beta) = \sum_{i=1}^n y_i \ln(\mu_i) - \sum_{i=1}^n \mu_i - \sum_{i=1}^n \ln(y_i!) \quad \dots(3.2.1)$$

Untuk menentukan penaksir kemungkinan maksimumnya, digunakan teknik iteratif yang cukup rumit. Beberapa perangkat lunak telah menyediakan fasilitas untuk menghitung penaksir bagi parameter regresi Poisson, di antaranya SAS, S-Plus, dan STATA. Dalam tugas akhir ini, perangkat lunak yang digunakan adalah SAS. Uji keberartian koefisien regresi juga akan dilakukan menggunakan SAS.

Setelah diperoleh penaksir parameter  $\hat{\beta}$ , model regresi Poisson berbentuk

$$\hat{y}_i = g^{-1}(x_i' \hat{\beta}) \quad \dots(3.2.2)$$

Sebagai contoh, jika penghubung identitas digunakan, maka persamaannya menjadi:

$$\hat{y}_i = g^{-1}(x_i' \hat{\beta}) = x_i' \hat{\beta} \quad \dots(3.2.3)$$

Dan jika penghubung log digunakan, maka

$$\hat{y}_i = g^{-1}(x_i' \hat{\beta}) = \exp(x_i' \hat{\beta}) \quad \dots(3.2.4)$$

### 6.3 Overdispersi pada Model Regresi Poisson

Overdispersi merupakan sebuah kondisi yang dapat terjadi ketika pemodelan menggunakan distribusi Poisson. Hal ini terjadi karena distribusi Poisson memiliki rata-rata dan varians yang sama, yaitu  $Var(Y) = E(Y)$ . Tetapi pada kenyataannya hal yang kadang terjadi adalah varians dari variabel responnya lebih besar daripada rata-ratanya,  $Var(Y) > E(Y)$ , keadaan seperti ini disebut juga dengan overdispersi.

Pengujian overdispersi pada regresi Poisson dilakukan dengan menggunakan statistik uji skor (C. B. Dean, 1992, dalam Utami, 2003:10). Prosedur yang akan dilakukan untuk melakukan pengujian tersebut sebagai berikut:

1. Perumusan Hipotesis

$H_0$ : Tidak terdapat overdispersi pada model regresi Poisson

$H_0$ : Terdapat overdispersi pada model regresi Poisson

2. Besaran-besaran yang digunakan

a. Menghitung  $\sum_{i=1}^n \{(Y_i - \mu_i)^2 - \mu_i\}$

b. Menghitung  $\{2 \sum_{i=1}^n \mu_i^2\}^{\frac{1}{2}}$

### 3. Statistik Uji

Statistik uji yang digunakan adalah:

$$S = \frac{\sum_{i=1}^n \{(Y_i - \mu_i)^2 - \mu_i\}}{\{2 \sum_{i=1}^n \mu_i^2\}^{\frac{1}{2}}} \quad \dots(3.4.1)$$

### 4. Kriteria Pengujian

Dengan mengambil taraf nyata  $\alpha$ , maka tolak  $H_0$  jika  $|S| > Z_{\alpha/2}$ .

Dalam hal ini,  $Z_{\alpha/2}$  diperoleh dari Tabel Distribusi Normal Baku dengan peluangnya  $\frac{\alpha}{2}$ .

### 5. Kesimpulan

Penafsiran dari  $H_0$  diterima atau ditolak. Jika  $H_0$  ditolak, maka dalam model regresi Poisson terdapat overdispersi sehingga model regresi Poisson dapat dikatakan kurang tepat.

### 6.4 Distribusi Binomial Negatif

Distribusi binomial negatif mempunyai peranan yang cukup penting dalam analisis statistika parametrik untuk data frekuensi yang mempunyai overdispersi. Misalkan bahwa variabel acak  $Y$  berdistribusi Poisson dengan parameter  $\lambda$  atau Poisson ( $\lambda$ ). Akan tetapi,  $\lambda$  itu sendiri merupakan peubah acak dan diasumsikan berdistribusi gamma, yaitu:

$$Y|\lambda \sim \text{Poisson}(\lambda)$$

$$\lambda \sim \text{gamma}(\alpha, \beta)$$

di mana  $gamma(\alpha, \beta)$  adalah distribusi gamma dengan rata-rata =  $\alpha\beta$  dan varian =  $\alpha\beta^2$  dengan fungsi densitas seperti pada persamaan (2.4.1). Fungsi densitas bersama bagi  $Y$  dan  $\lambda$ , yaitu:

$$f(y, \lambda) = f(y|\lambda)f(\lambda)$$

Akan tetapi, karena  $\lambda$  tidak diamati, maka  $\lambda$  harus dipisahkan melalui distribusi marginalnya, yaitu:

$$\begin{aligned} f(y, \alpha, \beta) &= \int_0^{\infty} f(y|\lambda)f(\lambda) \\ &= \int_0^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \cdot \frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}} d\lambda \\ &= \frac{1}{y! \beta^\alpha \Gamma(\alpha)} \int_0^{\infty} \lambda^{y+\alpha-1} e^{-(1+\frac{1}{\beta})\lambda} d\lambda \quad \dots(3.5.1) \end{aligned}$$

Integral di atas diselesaikan dengan menggunakan bantuan fungsi gamma, yaitu seperti pada persamaan (2.4.1), sehingga bentuk integral dalam persamaan (3.5.1) dapat ditulis menjadi:

$$\int_0^{\infty} \lambda^{y+\alpha-1} e^{-(1+\frac{1}{\beta})\lambda} d\lambda = \frac{\Gamma(y+\alpha)}{(1+\frac{1}{\beta})^{y+\alpha}} \cdot \int_0^{\infty} \left[ \left(1 + \frac{1}{\beta}\right) \lambda \right]^{y+\alpha-1} e^{-(1+\frac{1}{\beta})\lambda} d \left[ \left(1 + \frac{1}{\beta}\right) \lambda \right]$$

Dengan demikian distribusi marginalnya dapat ditulis menjadi :

$$\begin{aligned} f(y, \alpha, \beta) &= \frac{\Gamma(y+\alpha)}{y! \Gamma(\alpha)} \cdot \frac{1}{\beta^\alpha (1+\frac{1}{\beta})^{y+\alpha}} \\ &= \frac{\Gamma(y+\alpha)}{y! \Gamma(\alpha)} \cdot \frac{1}{\beta^\alpha \left(\frac{\beta+1}{\beta}\right)^\alpha \left(\frac{\beta+1}{\beta}\right)^y} \\ &= \frac{\Gamma(y+\alpha)}{y! \Gamma(\alpha)} \cdot \left(\frac{1}{1+\beta}\right)^\alpha \left(\frac{\beta}{1+\beta}\right)^y \quad \text{Untuk } y = 0, 1, 2, \dots \quad \dots(3.5.2) \end{aligned}$$

Schafer (2006) menyatakan bahwa distribusi tersebut mempunyai rata-rata:

$$E(Y) = E[E(Y|\lambda)] = E(\lambda) = \alpha\beta$$

Dan variannya dihitung sebagai berikut:

$$\begin{aligned} \text{Var}(Y) &= E[\text{Var}(y|\lambda)] + \text{Var}[E(Y|\lambda)] \\ &= \text{Var}(\lambda) + E(\lambda) \\ &= \alpha\beta + \alpha\beta^2 \end{aligned}$$

Untuk membangun model regresi, wajar untuk menyatakan distribusi binomial negatif dalam parameter  $\mu = \alpha\beta$  dan  $\kappa = \frac{1}{\alpha}$ , sehingga rata-rata dan varians menjadi:

$$E(Y) = \mu \quad \dots(3.5.3)$$

dan

$$\text{Var}(Y) = \mu + \kappa\mu^2 \quad \dots(3.5.4)$$

Sehingga fungsi densitas dari peubah acak Y adalah:

$$f(y, \mu, \kappa) = \frac{\Gamma(y+\kappa^{-1})}{y!\Gamma(\kappa^{-1})} \cdot \left(\frac{\kappa\mu}{1+\kappa\mu}\right)^y \left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa} \quad \dots(3.5.5)$$

Fungsi peluang pada persamaan (3.5.5) merupakan fungsi peluang dari distribusi binomial negatif dengan parameter  $\mu$  dan  $\kappa$ .

Penaksir kemungkinan maksimum dari parameter  $\mu$  dengan fungsi densitas seperti dalam persamaan (3.5.5) adalah:

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad \dots(3.5.6)$$

Sementara itu penaksir kemungkinan maksimum untuk parameter  $\kappa$  adalah solusi dari persamaan (3.5.7) dengan terlebih dahulu mensubstitusikan persamaann (3.5.6) ke persamaan (3.5.7).

$$\sum_{i=1}^n \sum_{v=0}^{y_i-1} \frac{v}{1+\kappa v} + n\kappa^{-2} \ln(1 + \kappa\mu) - \frac{1}{1+\kappa\mu} [\mu(\sum_{i=1}^n y_i + n\kappa^{-1})] = 0 \quad \dots(3.5.7)$$

## 6.5 Regresi Binomial Negatif

Misalkan  $y_i$  adalah nilai dari variabel respon untuk pengamatan ke-  $i$  dan  $x_i$  adalah vektor dari nilai variabel bebas untuk pengamatan ke-  $i$  dengan  $i = 1, 2, \dots, n$ . Model regresi binomial negatif mengasumsikan bahwa variabel respon ke-  $i$  mengikuti distribusi binomial negatif dengan parameter  $\mu_i$  dan  $\kappa$ , yang dinotasikan oleh  $Y_i \sim BN(\mu_i, \kappa)$  (Zafatia, 2003).

Regresi binomial negatif mengasumsikan suatu model berbentuk:

$$Y = \log \hat{\mu}_i = \eta_i = x_i' \beta$$

$$Y = \log \hat{\mu}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

### 3.5.1 Taksiran Parameter Model Regresi Binomial Negatif

Dengan mengabaikan konstanta  $y!$  pada persamaan (3.5.2), fungsi log kemungkinan pengamatan ke-  $i$  untuk model regresi binomial negatif adalah:

$$\ell_i = \log \Gamma(\kappa^{-1} + y_i) - \log \Gamma(\kappa^{-1}) + y_i \log \left( \frac{\kappa \mu_i}{1 + \kappa \mu_i} \right) + \kappa^{-1} \log \left( \frac{1}{1 + \kappa \mu_i} \right) \quad \dots(3.6.1.1)$$

dimana  $\mu_i = \exp(x_i' \beta)$

Jika nilai  $\kappa$  diketahui, maka parameter model regresi binomial negatif  $\beta$  dapat ditaksir dengan menggunakan metode *Fisher Scoring*. Nilai taksiran untuk satu iterasinya adalah:

$$\beta = (X'WX)^{-1} X'Wz \quad \dots(3.6.1.1)$$

di mana:

$$W = \text{Diag} \left[ \text{Var}(y_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 \right]^{-1} \quad \dots(3.6.1.2)$$

dan



$$z_i = \eta_i + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i) \quad \dots(3.6.1.3)$$

Seperti yang telah dikemukakan pada persamaan (3.5.4), di bawah fungsi penghubung  $\log\left(\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu}\right)$ , maka nilai bobotnya adalah:

$$w_i = \frac{\mu_i^2}{\mu_i + \kappa \mu_i^2} \quad \dots(3.6.1.4)$$

dan

$$z_i = x_i' \beta + \frac{y_i - \mu_i}{\mu_i} \quad \dots(3.6.1.5)$$

### 3.5.2 Uji Kecocokan Model Regresi Binomial Negatif

Untuk melakukan uji kecocokan model regresi binomial negatif, maka perumusan hipotesisnya sebagai berikut:

$H_0$  : Model regresi binomial negatif cocok dengan data

$H_1$  : Model regresi binomial negatif tidak cocok dengan data

Statistik uji yang digunakan adalah:

$$\chi^2 = \sum_{i=1}^n r_i^2 \quad \dots(3.6.2.1)$$

dengan residual Pearson pengamatan ke- $i$  atau  $r_i$  adalah:

$$r_i = \frac{y_i - \mu_i}{\sqrt{\mu_i + \kappa \mu_i^2}} \quad \dots(3.6.2.2)$$

Dalam uji kecocokan model regresi binomial negatif dapat juga digunakan nilai devians. Nilai devians untuk pengamatan ke- $i$  adalah:

Jika  $y_i > 0$

$$D_i = 2 \left\{ y_i \log \frac{y_i}{\mu_i} + (y_i + \kappa^{-1}) \log \left( \frac{1 + \kappa \mu_i}{1 + \kappa y_i} \right) \right\} \quad \dots(3.6.2.3)$$

Jika  $y_i = 0$

$$D_i = 2\kappa^{-1} \log (1 + \kappa \mu_i) \quad \dots(3.6.2.4)$$

Sehingga nilai devians untuk  $n$  pengamatan adalah:

$$G^2 = \sum_{i=1}^n D_i = \sum_{y_i > 0} 2 \left\{ y_i \log \frac{y_i}{\mu_i} + (y_i + \kappa^{-1}) \log \left( \frac{1 + \kappa \mu_i}{1 + \kappa y_i} \right) \right\} + \sum_{y_i = 0} 2\kappa^{-1} \log (1 + \kappa \mu_i) \quad \dots(3.6.2.5)$$

Untuk kedua statistik uji di atas, dengan  $H_0$  bahwa model regresinya adalah binomial negatif, maka tolak  $H_0$  jika nilai yang ada pada persamaan (3.6.2.1) dan (3.6.2.5) lebih besar dari  $\chi^2$  tabel dengan derajat kebebasan  $(n - p)$ , dimana  $p$  adalah banyak parameter dalam model.

