Model Regresi Zero Inflated Poisson Pada Data *Overdispersion*

Wirajaya Kusuma
Fakultas MIPA, Universitas Mataram
e-mail: Kusuma_Wirajaya@yahoo.co.id

Desy Komalasari
Fakultas MIPA, Universitas Mataram
e-mail: Desi_its@yahoo.com

Mustika Hadijati
Fakultas MIPA, Universitas Mataram
e-mail: Ika_wikan@yahoo.co.id

Abstract: Overdispersion is a phenomenon of the data variance greater than the average. One of the causes of overdispersion is too many zero value (excess zero) on the response variable. Zero inflated Poisson regression model (ZIP) is one of the method that can be used to overcome problems due to excess zeros. The purpose of this research is to estimate the regression parameters model Zero -inflated Poisson (ZIP) and applying to the data of unsuccessful students in national examinations in senior high school and vocational school in the city of Mataram. Parameter estimation Zero inflated Poisson regression model using the maximum likelihood and maximization expectation algorithm with Newton Rhapson approach. Zero inflated Poisson regression model obtained on the data is:

$\ln(\mu_i) = -0{,}3954 + 0{,}1153x_{2i}$ dan logit $(\boldsymbol{\omega}) = -4{,}2963 + 0{,}1280x_{3i}$

With X_2 is school accreditation; and X_3 is the proportion of teachers who are already certified

Keywords: Zero Inflated Poisson, Overdispersion, Maximization Expectation, Newton Rhapson, unsuccessful students SMA/SMK

1. Pendahuluan

Analisis regresi adalah suatu metode yang digunakan untuk menganalisis hubungan antara variabel *dependent* (Y) dan variabel *independent* (X). Salah satu model regresi yang dapat digunakan untuk menganalisis hubungan antara variabel *dependent* Y yang berupa data diskrit dan variabel *independent* X berupa data kontinyu, diskrit, atau campuran adalah model regresi Poisson. Dalam model regresi Poisson terdapat beberapa asumsi, salah satu asumsi yang harus terpenuhi adalah variansi dari variabel *dependent* sama dengan rataannya (*equidispersion*).

Namun, dalam analisis data sering dijumpai data yang variansinya lebih kecil atau lebih besar dari rataannya. Keadaan ini lebih dikenal dengan *underdispersion* atau *overdispersion*. Salah satu penyebab terjadinya *overdispersion* adalah terlalu banyak nilai nol (*excess zero*) pada variabel respon. Model regresi *Zero Inflated Poisson* (ZIP) merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah *overdispersion* akibat *excess zeros* pada data respon bertipe diskrit. Pada regresi ZIP terdapat dua parameter yang akan diestimasi yaitu parameter μ yang mewakili variabel yang mempengaruhi *zero state* dan parameter ω yang mewakili variabel yang mempengaruhi Poisson *state*.

Permasalahan yang diangkat dalam penelitian ini yaitu bagaimana menaksir parameter model regresi *Zero Inflated* Poisson (ZIP) dalam mengatasi *overdispersion*; dan bagaimana menerapkannya pada data. Sehingga yang menjadi tujuan pada penelitian ini yaitu menaksir parameter model regresi *Zero Inflated* Poisson (ZIP); dan menerapkan pada data ketidaklulusan siswa SMA/SMK di Kota Mataram.

2. Tinjauan Pustaka

2.1 Regresi Poisson

Regresi Poisson termasuk kedalam *Generalized linier model* (GLM) dan merupakan salah satu bentuk regresi yang digunakan untuk model data cacah. Variabel *dependent* dalam persamaan tersebut menyatakan data cacah (Hilbe [4]).

GLM didefinisikan ke dalam tiga komponen yaitu komponen acak, komponen sistematis dan fungsi penghubung. Komponen acak adalah suatu komponen yang mengidentifikasikan distribusi peluang dari variabel respon $Y = (y_1, y_2, \dots, y_n)$ diasumsikan saling bebas dan memiliki distribusi yang termasuk dalam keluarga eksponensial dengan fungsi kepadatan peluang sebagai berikut:

$$f(y;\theta_i,\phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i;\phi)\right] \tag{2.1}$$

Parameter θ_i disebut parameter natural dan parameter ϕ disebut dengan parameter dispersi. Komponen sistematis dari komponen model linier umum yang menghubungkan vektor $\boldsymbol{\eta} = [\eta_1 \quad \eta_2 \quad \cdots \quad \eta_n]^T$ kepada sekumpulan variabel prediktor melalui model linier $\boldsymbol{\eta} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$ dimana $\mathbf{X}^{\mathrm{T}}$ adalah matriks rancangan yang berisi nilai-nilai variabel prediktor untuk n buah pengamatan, $\boldsymbol{\beta}$ adalah vektor dari parameter dalam model dan vektor $\boldsymbol{\eta}$ disebut prediktor linier. Komponen ketiga dari GLM yaitu fungsi *link* yang menghubungkan komponen acak dengan komponen sistematis.

Suatu fungsi disebut fungsi *link* kanonik jika:

$$g(\mu_i) = \theta_i = \sum_j \beta_j x_{ji} \tag{2.2}$$

dengan θ_i merupakan parameter kanonik (Agresti [1]).

Jika y_i merupakan variabel acak untuk data cacah dengan $i = 1,2,\dots,n$, dimana n menyatakan banyaknya data dan y_i berdistribusi Poisson maka fungsi kepadatan peluangnya adalah:

$$f(y_i,\mu_i) = \frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!} \tag{2.3}$$

Untuk $\mu_i > 0$, dengan μ_i merupakan rataan dari variabel *dependen* Y.
Fungsi peluang Poisson termasuk keluarga eksponensial sehingga dapat ditulis:

$$f(y_i,\mu_i) = \exp\left(y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\right) \tag{2.4}$$

Dengan menggunakan fungsi *link* diperoleh model regresi Poisson berikut:

$$\eta_i = \ln(\mu_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} \tag{2.5}$$

Asumsi yang harus dipenuhi pada model regresi Poisson yaitu:

$$Var(Y_i) = E(Y_i) = \mu_i = \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$$

Dengan $\mathbf{X}_\mathrm{i}^{\mathrm{T}}$ matriks yang berukuran $1 \times p$ yang menjelaskan variabel *independen* dan $\boldsymbol{\beta}$ adalah vektor berukuran $p \times 1$ yang merupakan parameter regresi. Sehingga fungsi kepadatan peluang pada regresi Poisson adalah sebagai berikut:

$$f(y_i,\mu_i) = \frac{\exp[(y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})]}{y_i!} \tag{2.6}$$

Nilai harapan y_i bergantung pada variabel *independen* adalah μ_i. Taksiran parameter koefisien regresi Poisson dapat dilakukan dengan menggunakan *Maximum Likelihood Estimation* (MLE) (Hilbe [4]).

2.2 *Overdispersion*

Fenomena *overdispersion* dapat ditulis:

$$\mathrm{Var}(Y) > E(Y) \tag{2.7}$$

Taksiran dispersi dapat diukur dengan nilai *deviance* dan *pearson chi-square*. Data dikatakan *overdispersion* jika taksiran dispersi lebih besar dari 1 dan *underdispersion* jika taksiran dispersi kurang dari 1 (Khoshgoftaar, et al. [8]).

Terdapat dua cara yang dapat digunakan untuk mendeteksi *overdispersion*, yaitu:

1. *Deviance*

$$\Phi_1 = \frac{D^2}{db}\ ;\ D^2 = 2\sum_{i=1}^{n}\left\{y_i ln\left(\frac{y_i}{\mu_i}\right)\right\} \tag{2.8}$$

dimana db = $n - p$ dengan p merupakan banyak parameter termasuk konstanta, n merupakan banyaknya pengamatan dan D^2 adalah nilai *deviance* (Hilbe [4]).

2. *Pearson chi-square*

untuk menguji asumsi *equidispersion* pada regresi Poisson dilakukan dengan melihat nilai statistik *Pearson chi-square* yang dibagi dengan derajat bebasnya.

$$\Phi_2 = \frac{\chi^2}{db}\ ;\ \chi^2 = \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mathrm{var}(y_i)} \tag{2.9}$$

dimana $db = n - p$ dengan p merupakan banyak parameter termasuk konstanta, n merupakan banyaknya pengamatan dan χ^2 adalah nilai *pearson chi-square* .

Jika nilai Φ_1 dan Φ_2 bernilai lebih dari satu maka terjadi *overdispersion* pada data.

2.3 Regresi *Zero-Inflated* Poisson (ZIP)

Salah satu penyebab terjadinya *overdispersion* adalah lebih banyak observasi bernilai nol daripada yang ditaksir. Salah satu metode yang diusulkan untuk menaksir yaitu model regresi *Zero Inflated Poisson* (ZIP) (Jansakul & Hinde [7]).

Jika y_i adalah variabel random yang mempunyai distribusi ZIP, nilai nol pada observasi diduga muncul dalam dua cara yang sesuai untuk keadaan (*state*) yang terpisah. Keadaan pertama disebut *zero state* terjadi dengan probabilitas ω_i dan menghasilkan hanya observasi bernilai nol, sementara keadaan kedua disebut Poisson *state* terjadi dengan probabilitas $(1-\omega_i)$ dan berdistribusi Poisson dengan mean μ_i (Jansakul & Hinde [7]).

Proses dua keadaan ini memberikan distribusi campuran dua komponen dengan fungsi probabilitas sebagai berikut:

$$\mathrm{P}(\mathrm{Y_i} = y_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\mu_i}, & untuk\ \ y_i = 0 \\ \dfrac{(1-\omega_i)e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}, & untuk\ \ y_i > 0, 0 \le \omega_i \le 1 \end{cases} \tag{2.10}$$

Lambert [9] menyarankan model gabungan untuk μ dan ω, yakni:

$$\ln(\mu) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} \text{ dan logit } (\boldsymbol{\omega}) = \ln\left(\frac{\omega}{1-\omega}\right) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} \tag{2.11}$$

dengan $\mathbf{X}_i^{\mathrm{T}}$ adalah matriks variabel prediktor, $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ adalah vektor parameter yang akan ditaksir, dan $\boldsymbol{\omega}$ adalah probabilitas observasi bernilai nol (Ismail & Zamani [6]).

2.4 Metode Maksimum *Likelihood*

Pandang suatu sampel random $X_1, X_2, \ldots, X_n$ dari suatu distribusi yang memiliki p.d.f $f(x;\theta)$: $\theta \epsilon \Omega$, dimana θ merupakan suatu parameter yang tidak diketahui dan Ω adalah ruang parameter. Karena $X_1, X_2, \ldots, X_n$ sampel random, maka p.d.f bersama dari $X_1, X_2, \ldots, X_n$ adalah:

$$f(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) \quad (2.12)$$

Fungsi *likelihood* didefinisikan sebagai fungsi peluang bersama dari $X_1, X_2, \dots, X_n$ yang dapat dianggap sebagai fungsi dari θ. Misalkan fungsi *likelihood*:

$$\begin{aligned} L(\theta) = f(x_1, x_2, \dots, x_n; \theta) &= f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) \ ; \ \theta \epsilon \Omega \\ &= \textstyle\prod_{i=1}^{n} f(x_i; \theta) \end{aligned} \quad (2.13)$$

Penaksiran maksimum *likelihood* yang memaksimumkan fungsi *likelihood* $\hat{\theta}$ disebut taksiran maksimum *likelihood* dari θ. Nilai θ yang memaksimumkan $l(\theta)$ dapat diperoleh dengan mencari solusi dari persamaan $\frac{\partial l(\theta)}{\partial \theta} = 0$ (Hoog & Craig [5]).

2.5 Algoritma EM

Algoritma EM diperkenalkan oleh Dempster, Laird dan Rubin [2], merupakan salah satu metode optimasi yang digunakan sebagai alternatif dalam memaksimumkan fungsi *likelihood* yang mengandung data missing. Dua tahap dilakukan dalam Algoritma EM, yaitu tahap Ekspektasi dan Maksimalisasi. Misal diasumsikan terdapat data observasi *x* berdistribusi tertentu yang mengandung data missing *y*. Untuk mengatasinya dibentuk distribusi gabungan antara *x* dan *y*, yaitu:

$$f\left(z \middle| \hat{\theta}\right) = f\left(x, y \middle| \hat{\theta}\right) = f(x|y)\, f(y) \quad (2.14)$$

Langkah "E" pada algoritma EM adalah menghitung *complete data likelihood*. $L(\boldsymbol{\theta} \mid x, y)$, yaitu menghitung ekspektasi dari missing data dengan diketahui data yang ada (tidak *missing*). Berikut langkah-langkah algoritma EM diantaranya Menentukan inisialisasi parameter $\hat{\theta}_k$; $k = 0$; Langkah ekspektasi yaitu Menghitung *complate data likelihood* dengan cara substitusi $\hat{\theta}_k$ pada fungsi $Q\left(\hat{\theta}_k\right) = E\left\lfloor L\left(\hat{\theta}_k \middle| x, y\right) | x \right\rfloor$; Langkah maksimalisasi dilakukan dengan mengacu pada kondisi $\frac{\partial Q(\hat{\theta}_k)}{\partial \hat{\theta}_k} = 0$ untuk mendapatkan inisialisasi yang baru; Langkah E dan M dilakukan secara iteratif sampai didapatkan perbedaan antara $\left(\hat{\theta}_{k+1} - \hat{\theta}_k\right)$ lebih kecil dari kriteria l tertentu yang bernilai kecil sehingga diperoleh $\hat{\theta}$ yang konvergen dan memenuhi $\frac{\partial^2 Q(\hat{\theta})}{\partial \hat{\theta}^2} < 0$. Optimasi secara numerik dapat dilakukan pada langkah M.

3. Metode Penelitian

Berikut tahapan penelitian meliputi:

1. Menaksir Parameter Model Regresi *Zero Inflated* Poisson

   Untuk menaksir parameter model regresi *Zero Inflated* Poisson digunakan *maximum likelihood estimation* (MLE) dengan langkah-langkahnya sebagai berikut:

Menentukan fungsi probabilitas variabel respon dari model regresi *zero-inflated* Poisson; Membentuk fungsi *likelihood*; Menentukan fungsi *ln-likelihood*; Membentuk distribusi dari variabel z_i; Membentuk distribusi gabungan antara y_i dan z_i; Menentukan turunan parsial pertama dan kedua dari fungsi *ln-likelihood* distribusi gabungan y_i dan z_i; Tahap ekspektasi; dan Tahap maksimalisasi dengan iterasi *Newton-rhapson*.

2. Penerapannya Pada Data

   Pada penelitian ini, diterapkan pada data ketidaklulusan Siswa SMA/SMK dalam Ujian Nasional di Kota Mataram Tahun 2012. Data diambil dari DIKPORA Kota Mataram yang memuat 42 data sekolah dimana variabel yang digunakan dalam penelitian ini berupa Variabel Respon (Y) yaitu Jumlah siswa SMA/SMK yang tidak lulus UN di Kota Mataram tahun 2012; dan Variabel prediktor (X) meliputi X_1 (jumlah peserta UN pada tiap SMA/SMK di Kota Mataram); X_2 (akreditasi sekolah SMA/SMK di kota Mataram) yang dibagi sebagai berikut:

$$\mathrm{X}_2 = \begin{cases} 1, & jika\ SMA\ atau\ SMK\ terakreditasi\ A \\ 0, & jika\ SMA\ atau\ SMK\ terakreditasi\ selain\ A \end{cases}$$

   dan X_3 (presentasi proporsi guru yang sudah sertifikasi pada tiap SMA/SMK di kota Mataram).

4. Hasil dan Pembahasan

4.1 Menaksir Parameter Model Regresi *Zero-inflated* Poison (ZIP)

Jika y_i adalah variabel random yang mempunyai distribusi ZIP, nilai nol pada observasi diduga muncul dalam dua cara yang sesuai untuk keadaan (*state*) yang terpisah. Keadaan pertama disebut *zero state* terjadi dengan probabilitas ω_i dan menghasilkan observasi bernilai nol, sementara keadaan kedua disebut *poisson state* terjadi dengan probabilitas (1-ω_i) dan berdistribusi Poisson dengan mean μ_i.

Proses dua keadaan ini memberikan distribusi campuran dua komponen dengan fungsi probabilitas sebagai berikut:

$$\mathrm{P}(\mathrm{Yi} = y_i) = \begin{cases} \omega_i + (1-\omega_i)\exp(-\mu_i), & untuk\ \ y_i = 0 \\ \dfrac{(1-\omega_i)\exp(-\mu_i){\mu_i}^{y_i}}{y_i!}, & untuk\ \ y_i > 0, 0 \le \omega_i \le 1 \end{cases} \tag{4.1}$$

Lambert [9] menyarankan model gabungan untuk μ dan ω, yakni:

$$\ln(\mu_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} \text{ dan logit } (\omega) = \ln\left(\frac{\omega_i}{1-\omega_i}\right) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} \tag{4.2}$$

Dengan $\mathbf{X}_i^{\mathrm{T}}$ adalah matriks variabel prediktor, $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ adalah vektor parameter yang akan ditaksir, dan $\boldsymbol{\omega}$ adalah probabilitas observasi bernilai nol.

Maka dapat ditentukan nilai μ_i , ω_i dan $(1-\omega_i)$ sebagai berikut:

$$\mu_i = \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$$

$$\omega_i = \frac{exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)}{(1+exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))}$$

$$(1-\omega_i) = \frac{\omega_i}{exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma})}$$

nilai μ_i , ω_i dan $(1-\omega_i)$ disubstitusikan yaitu:

$$\mathrm{P}(\mathrm{Yi} = y_i) = \begin{cases} \dfrac{\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))} + \dfrac{1}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))}\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right), & y_i = 0 \\ \dfrac{\dfrac{1}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))}\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\left(\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)^{y_i}}{y_i!}, & y_i > 0, \end{cases} \tag{4.3}$$

Jika n buah pengamatan diasumsikan saling bebas, maka fungsi *likelihood* diperoleh dengan mengalikan semua fungsi probabilitas dari Y_i yaitu:

Untuk $y_i = 0$

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i) = \prod_{y_i=0}\left[\frac{\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)+\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))}\right] \tag{4.4}$$

Untuk $y_i > 0$

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i) = \prod_{y_i>0}\left[\frac{\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\left(\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{y_i}}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))\, y_i!}\right] \tag{4.5}$$

Dibentuk fungsi *ln-likelihood* sebagai berikut:

Untuk $y_i = 0$

$$\begin{aligned} l(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i) &= lnL(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i) \\ &= \sum_{y_i=0} ln\left(\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)+\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\right) - \sum_{y_i=0} ln\left(1+\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) \end{aligned}$$

Untuk $y_i > 0$

$$\begin{aligned} l(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i) &= lnL(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i) \\ &= ln\left\{\prod_{y_i>0}\left[\frac{\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\left(\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{y_i}}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))\, y_i!}\right]\right\} \\ &= \sum_{y_i>0}\left(y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right) - \sum_{y_i>0} ln\left(1+\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) - \sum_{y_i>0}\ln(y_i!) \end{aligned} \tag{4.6}$$

Total *ln-likelihood* dapat diberikan oleh $l_T = l(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i)_{\,y_i=0} + l(\boldsymbol{\beta},\boldsymbol{\gamma}|\, y_i)_{\,y_i>0}$ yaitu:

$$l_T = \sum_{y_i=0} ln\left(\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right) + \exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\right) - \sum_{y_i=0} ln\left(1+\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) + \sum_{y_i>0}\left(y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right) - \sum_{y_i>0} ln\left(1+\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) - \sum_{y_i>0}\ln(\,y_i!) \qquad (4.7)$$

Pada persamaan (4.7) tidak diketahui mana nilai nol yang berasal dari *zero state* dan mana yang berasal dari *poisson state* sehingga menyulitkan perhitungan dan fungsi *ln-likelihood* ini tidak dapat diselesaikan dengan metode numerik biasa.

Untuk memaksimalkan fungsi *ln-likelihood* digunakan algoritma *expectation maximization* (EM) yang merupakan salah satu metode optimasi yang banyak digunakan sebagai alternatif dalam memaksimumkan fungsi *likelihood* yang mengandung data hilang (*missing*)).

Misalkan untuk setiap Y_i berkaitan dengan variabel indikator z_i yaitu:

$$z_i = \begin{cases} 1, & jika\;\; y_i\; berasal\; dari\; zero\; state \\ 0, & jika\;\; y_i\; berasal\; dari\; poisson\; state \end{cases} \qquad (4.8)$$

Dapat dibentuk distribusi dari indikator z_i yaitu:

$\mathrm{P}(z_i = 1) = \omega_i$ dan $\mathrm{P}(z_i = 0) = 1 - \omega_i$

Sehingga z_i ~ binomial (1, ω_i).

Jika nilai variabel respon y_i = 1,2,3,...., maka nilai z_i = 0. Sedangkan jika nilai variabel respon $y_i = 0$, maka nilai z_i mungkin 0 atau 1. Jadi untuk variabel respon $y_i = 0$, maka nilai z_i belum bisa ditentukan. Oleh karena itu dibentuk distribusi gabungan antara Y_i dan z_i sebagai berikut:

$$\mathrm{P}(\,y_i, z_i|\omega_i, \mu_i) = \begin{cases} (z_i)\omega_i + (1-z_i)(1-\omega_i)\exp(-\mu_i), & untuk\;\; y_i = 0 \\ (1-z_i)\dfrac{(1-\omega_i)\exp(-\mu_i){\mu_i}^{y_i}}{y_i!}, & untuk\;\; y_i > 0, 0 \le \omega_i \le 1 \end{cases}$$

$$= \begin{cases} (z_i)\dfrac{\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))} + (1-z_i)\dfrac{1}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))}\exp(-\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})), \; y_i = 0 \\ (1-z_i)\dfrac{\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\left(\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{y_i}}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))\; y_i!}, \; y_i > 0, \end{cases} \qquad (4.9)$$

Jika n buah pengamatan diasumsikan saling bebas, maka fungsi *likelihood* diperoleh dengan mengalikan semua fungsi probabilitas dari Y_i sebagai berikut:

Untuk $y_i = 0$

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}|\, z_i, y_i) = \prod_{y_i=0}\left[(z_i)\frac{\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))} + (1-z_i)\frac{\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)}{(1+\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))}\right] \qquad (4.10)$$

Untuk $y_i > 0$

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i) = \prod_{y_i>0} \left[(1 - z_i) \frac{\exp\left(-\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\left(\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{y_i}}{(1 + \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}))\, y_i!} \right] \tag{4.11}$$

Digunakan fungsi *ln-likelihood* untuk mempermudah perhitungan sebagai berikut:

Untuk $y_i = 0$

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i) = lnL(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i)$$
$$= \sum_{y_i=0} \left(z_i \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right) - \sum_{y_i=0} (1 - z_i) \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right) - \sum_{y_i=0} \ln\left(1 + \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right)$$

Untuk $y_i > 0$

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i) = lnL(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i)$$
$$= \sum_{y_i>0} (1 - z_i)\,(\, y_i \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sum_{y_i>0} (1 - z_i) \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$$
$$- \sum_{y_i>0} ln\left(1 + \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) - \sum_{y_i>0} ln\, y_i! \tag{4.12}$$

Sehingga total *ln-likelihood* nya dapat diberikan oleh $l_T = l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i)_{y_i=0} + l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, z_i, y_i)_{y_i>0}$ yaitu:

$$l_T = \sum_{i=1}^{n} \left[z_i \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} - ln\left(1 + \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) \right] + \sum_{i=1}^{n} (1 - z_i)\left((\, y_i \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$$
$$- \sum_{i=1}^{n} (n) ln\, y_i!$$

Persamaan ini disebut *complate data likelihood*. Persamaan ini dimaksimumkan menggunakan algoritma EM, dimana parameter $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ dapat diestimasi secara terpisah. Sehingga dapat ditulis sebagai berikut:

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, z_i, y_i) = l(\boldsymbol{\gamma}, z_i, y_i) + l(\boldsymbol{\beta}, z_i, y_i) - \textstyle\sum_{i=1}^{n} (n) ln\, y_i! \tag{4.13}$$

dengan :

$$l(\boldsymbol{\gamma}, z_i, y_i) = \textstyle\sum_{i=1}^{n} \left[z_i \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} - ln\left(1 + \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right) \right]$$

(4.14)

dan

$$l(\boldsymbol{\beta}, z_i, y_i) = \textstyle\sum_{i=1}^{n} (1 - z_i)\left((\, y_i \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right) \tag{4.15}$$

Taksiran maksimum *likelihood* $\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_p$ dan $\hat{\gamma}_0, \hat{\gamma}_1, \dots \hat{\gamma}_p$, sebagai berikut:

Untuk $l(\boldsymbol{\beta}, z_i, y_i)$

Misalkan $\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right) = \exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi})$ dan $U_j(\boldsymbol{\beta})$ turunan pertama dari $l(\boldsymbol{\beta}, z_i, y_i)$ terhadap β_j $(j = 0,1,2, \dots, p)$ maka:

$$l(\boldsymbol{\beta}, z_i, y_i) = \sum_{i=1}^{n}(1 - z_i)\left(y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right) \tag{4.16}$$

$$U_p(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta}, z_i, y_i)}{\partial \beta_p} = \sum_{i=1}^{n}(1 - z_i)x_{pi}\left(y_i - \exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi})\right) = 0$$

Untuk $l(\boldsymbol{\gamma}, z_i, y_i)$

Misalkan $\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right) = \exp(\gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_p x_{pi})$ dan $U_j(\boldsymbol{\gamma})$ turunan pertama dari $l(\boldsymbol{\gamma}, z_i, y_i)$ terhadap γ_j $(j = 0,1,2, \dots, p)$ maka :

$$l(\boldsymbol{\gamma}, z_i, y_i) = \sum_{i=1}^{n}\left[z_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} - ln\left(1 + \exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)\right)\right]$$

$$= \sum_{i=1}^{n}\left[z_i(\gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_p x_{pi}) - ln\left(1 + \exp\left(\gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_p x_{pi}\right)\right)\right] \tag{4.17}$$

Selanjutnya turunan pasial kedua dari $l(\boldsymbol{\beta}, z_i, y_i)$ terhadap $\beta_0, \beta_1, \dots \beta_p$ dan $l(\boldsymbol{\gamma}, z_i, y_i)$ terhadap $\gamma_0, \gamma_1, \dots \gamma_p$ yaitu:

Untuk $l(\boldsymbol{\beta}, z_i, y_i)$

Misalkan $H_{jk}(\boldsymbol{\beta})$ adalah turunan parsial kedua dari $l(\boldsymbol{\beta}, z_i, y_i)$ terhadap $\beta_0, \beta_1, \dots \beta_p$ dimana $j, k = 0,1,2, \dots, p$ maka:

$$H_{pp}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta}, z_i, y_i)}{\partial {\beta_p}^2}$$

$$= \frac{\partial}{\partial \beta_p}\left[\sum_{i=1}^{n}(1 - z_i)x_{pi}\left(y_i - \exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)\right)\right]$$

Misalkan $H_{jk}(\boldsymbol{\gamma})$ adalah turunan parsial kedua dari $l(\boldsymbol{\gamma}, z_i, y_i)$ terhadap $\gamma_0, \gamma_1, \dots \gamma_p$ dimana $j, k = 0,1,2, \dots, p$ maka:

$$H_{pp}(\boldsymbol{\gamma}) = \frac{\partial^2 l(\boldsymbol{\gamma}, z_i, y_i)}{\partial {\gamma_p}^2}$$

$$= -\sum_{i=1}^{n}\frac{{x_{pi}}^2\exp\left(\gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_p x_{pi}\right)}{\left(1 + \exp\left(\gamma_0 + \gamma_1 x_{1i} + \dots + \gamma_p x_{pi}\right)\right)^2}$$

Atau dapat ditulis dalam bentuk matriks:

$$\mathbf{H}(\boldsymbol{\gamma}) = \begin{bmatrix} H_{00}(\boldsymbol{\gamma}) & H_{01}(\boldsymbol{\gamma}) & \dots & H_{0p}(\boldsymbol{\gamma}) \\ H_{10}(\boldsymbol{\gamma}) & H_{11}(\boldsymbol{\gamma}) & \dots & H_{1p}(\boldsymbol{\gamma}) \\ \vdots & \vdots & \ddots & \vdots \\ H_{p0}(\boldsymbol{\gamma}) & H_{p1}(\boldsymbol{\gamma}) & \cdots & H_{pp}(\boldsymbol{\gamma}) \end{bmatrix}$$

Tahap ekspektasi:

Mengganti variabel z_i dengan ${z_i}^{(m)}$ $(m = 0,1,2,\dots)$ yang merupakan ekspektasi dari z_i

$$\begin{aligned} {z_i}^{(m)} &= E\left(z_i | y_i, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\beta}^{(m)}\right) \\ &= P\left(z_i | y_i, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\beta}^{(m)}\right) \end{aligned}$$

Untuk $y_i = 0$

$$= \frac{1}{1+\exp\left(-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}^{(\mathrm{m})} - exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{(\mathrm{m})}\right)\right)} \tag{4.18}$$

Untuk $y_i > 0$

$$\begin{aligned} {z_i}^{(m)} &= P\left(z_i = 0 | y_i, \boldsymbol{\gamma}^{(m)}, \boldsymbol{\beta}^{(m)}\right) \\ &= 0 \end{aligned} \tag{4.19}$$

Tahap maksimalisasi untuk parameter $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$

Untuk mencari taksiran maksimum *likelihood* ($\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_p$ dan $\hat{\gamma}_0, \hat{\gamma}_1, \dots \hat{\gamma}_p$) digunakan metode newton-rhapson dengan prosedur sebagai berikut:

1. Pilih taksiran awal dari $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ yaitu $\widehat{\boldsymbol{\beta}}^{(0)} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}^{(0)}$

dan $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix}$ yaitu $\hat{\boldsymbol{\gamma}}^{(0)} = \begin{bmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \\ \vdots \\ \hat{\gamma}_p \end{bmatrix}^{(0)}$

2. Tentukan taksiran dari $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ pada iterasi ke- $m+1$ $(m = 0,1,2,\dots)$ yaitu $\boldsymbol{\beta}^{(m+1)}$ dan $\boldsymbol{\gamma}^{(m+1)}$, secara iteratif menggunakan formula:

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} - \left[\mathbf{H}\left(\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}\right)\right]^{-1}\mathbf{U}\left(\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}\right) \tag{4.20}$$

dan

$$\hat{\boldsymbol{\gamma}}^{(m+1)} = \hat{\boldsymbol{\gamma}}^{(m)} - \left[\mathbf{H}\left(\hat{\boldsymbol{\gamma}}^{(\mathrm{m})}\right)\right]^{-1}\mathbf{U}\left(\hat{\boldsymbol{\gamma}}^{(\mathrm{m})}\right) \tag{4.21}$$

dengan :

$\widehat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{\gamma}}^{(m)}$ adalah masing-masing taksiran dari $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ pada iterasi ke $-\, m$.

$\mathbf{U}(\widehat{\boldsymbol{\beta}}^{(m)}), U(\widehat{\boldsymbol{\gamma}}^{(m)})$ adalah masing-masing vektor turunan parsial pertama dari $l(\boldsymbol{\beta}, z_i, y_i)$ dan $l(\boldsymbol{\gamma}, z_i, y_i)$ dengan elemen $\mathrm{U}_j(\boldsymbol{\beta}) = \frac{\partial l(\beta, z_i, y_i)}{\partial \beta_j}$ dan $\mathrm{U}_j(\boldsymbol{\gamma}) = \frac{\partial l(\gamma, z_i, y_i)}{\partial \gamma_j}$ (dengan $j = 0,1,2,\ldots,p$) dihitung pada $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(m)}$ dan $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}^{(m)}$.
$\mathbf{H}(\widehat{\boldsymbol{\beta}}^{(m)}), H(\widehat{\boldsymbol{\gamma}}^{(m)})$ adalah masing-masing matriks turunan parsial kedua dari $l(\boldsymbol{\beta}, z_i, y_i)$ dan $l(\boldsymbol{\gamma}, z_i, y_i)$ dengan elemen $H_{jk}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_j}$ dan $H_{jk}(\boldsymbol{\gamma}) = \frac{\partial^2 l(\gamma, z_i, y_i)}{\partial \gamma_k \partial \gamma_j}$ (dengan $j, k = 0,1,2,\ldots,p$) dihitung pada $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(m)}$ dan $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}^{(m)}$.

3. Ganti $\boldsymbol{\beta}^{(m)}$ dan $\boldsymbol{\gamma}^{(m)}$ dengan $\boldsymbol{\beta}^{(m+1)}$ dan $\boldsymbol{\gamma}^{(m+1)}$ pada iterasi selanjutnya kemudian kembali lakukan tahap ekspektasi.

4. Tahap ke- 2 dan ke- 3 dilakukan secara berulang-ulang. Hentikan proses iterasi jika $\|\widehat{\boldsymbol{\beta}}^{(m+1)} - \widehat{\boldsymbol{\beta}}^{(m)}\| < 10^{-5}\ dan\ \|\widehat{\boldsymbol{\gamma}}^{(m+1)} - \widehat{\boldsymbol{\gamma}}^{(m)}\| < 10^{-5}$, lalu ambil $\widehat{\boldsymbol{\beta}}^{(m+1)}$ sebagai taksiran $\widehat{\boldsymbol{\beta}}$ dan $\widehat{\boldsymbol{\gamma}}^{(m+1)}$ sebagai taksiran $\widehat{\boldsymbol{\gamma}}$.

Dengan menggunakan taksiran maksimum *likelihood* $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ dan $\hat{\gamma}_0, \hat{\gamma}_1, \ldots \hat{\gamma}_p$ dapat dibentuk model ZIP nya yaitu:

$$\ln(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi} \qquad i = 1,2,\ldots,n \tag{4.22}$$

dan

$$\text{logit}\,(\widehat{\boldsymbol{\omega}}) = \hat{\gamma}_0 + \hat{\gamma}_1 x_{1i} + \cdots + \hat{\gamma}_p x_{pi} \tag{4.23}$$

Untuk menyelesaikan persamaan iterasi pada umumnya digunakan *software*.

4.1.2 Taksiran Matriks Variansi-Kovariansi

Taksiran matriks variansi-kovariansi $\widehat{\boldsymbol{\beta}}$ dan $\widehat{\boldsymbol{\gamma}}$ dinyatakan dengan $\mathbf{V}(\widehat{\boldsymbol{\beta}})$ dan $\mathbf{V}(\widehat{\boldsymbol{\gamma}})$:

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}) \approx -[\mathbf{H}(\widehat{\boldsymbol{\beta}})]^{-1} \text{ dan } \mathbf{V}(\widehat{\boldsymbol{\gamma}}) \approx -[\mathrm{H}(\widehat{\boldsymbol{\gamma}})]^{-1} \tag{4.24}$$

Elemen diagonal utama ke-j masing-masing matriks $\mathbf{V}(\widehat{\boldsymbol{\beta}})$ dan $\mathbf{V}(\widehat{\boldsymbol{\gamma}})$ menunjukkan taksiran variansi dari $\hat{\beta}_j$ dan $\hat{\gamma}_j$ yang dinyatakan dengan $\mathrm{Var}(\hat{\beta}_j)$ dan $\mathrm{Var}(\hat{\gamma}_j)$, elemen non diagonalnya menunjukkan taksiran kovariansi dari $(\hat{\beta}_j\ dan\ \hat{\beta}_k)$ dan $(\hat{\gamma}_j\ dan\ \hat{\gamma}_k)$ yang dinyatakan dengan $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k)$ dan $\mathrm{Cov}(\hat{\gamma}_j, \hat{\gamma}_k)$; $j, k = 0,1,2,\ldots,p$. Dari matriks $\mathbf{V}(\widehat{\boldsymbol{\beta}})$ dan $\mathbf{V}(\widehat{\boldsymbol{\gamma}})$ diperoleh taksiran standar error dari $\hat{\beta}_j$ dan $\hat{\gamma}_j$ yaitu:

$$SE(\hat{\beta}_j) = \sqrt{\mathrm{Var}(\hat{\beta}_j)} \text{ dan } SE(\hat{\gamma}_j) = \sqrt{\mathrm{Var}(\hat{\gamma}_j)} \qquad j = 0,1,2,\ldots,p \tag{4.25}$$

Nilai $SE(\hat{\beta}_j)$ dan $SE(\hat{\gamma}_j)$ akan digunakan dalam pengujian signifikansi dari tiap parameter dalam model.

4.2 Penerapan Pada Data

4.2.1 Model *Zero-Inflated* Poisson (ZIP)

Data mengalami *overdispersion* maka untuk menganalisisnya digunakan ZIP. Dengan menggunakan software diperoleh estimasi parameter model regresi ZIP sebagai berikut:

$$\ln(\mu_i) = -0{,}3954 - 0{,}1584x_{1i} + 0{,}1153x_{2i} + 7{,}8536x_{3i} \qquad i = 1,2,..,42$$
$$\text{logit}\,(\boldsymbol{\omega}) = -4{,}2963 - 0{,}0056x_{1i} + 1{,}2618x_{2i} + 0{,}1280x_{3i} \qquad i = 1,2,..,42$$

Selanjutnya diuji signifikansi model apakah model tersebut dapat digunakan untuk menggambarkan hubungan antara jumlah siswa yang tidak lulus UN dengan jumlah peserta UN, akreditasi sekolah dan proporsi guru yang sudah sertifikasi. Menggunakan nilai *fits statistics* regresi ZIP diperoleh -2 Log Likelihood sebesar 35,20; nilai AIC sebesar 51,2; nilai AICC sebesar 55,6; dan nilai BIC sebesar 65,1. Sehingga diperoleh nilai $G = 35{,}2 > \chi^2{}_{0{,}05;3} = 12{,}592$ H_0 ditolak yang artinya model regresi Poisson signifikan dan model ZIP dapat digunakan untuk menggambarkan hubungan antara jumlah siswa yang tidak lulus UN dengan jumlah peserta UN, akreditasi sekolah dan proporsi guru yang sudah sertifikasi.

Selanjutnya akan diuji signifikansi dari tiap parameter model regresi Poisson menggunakan uji Wald untuk model $\ln(\mu_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$. Diperoleh hanya parameter β_2 yang signifikan pada $\alpha = 0{,}05$ sedangkan parameter yang lain tidak signifikan. Berdasarkan kesimpulan tersebut didapat model sebagai berikut:

$$\ln(\mu_i) = -0{,}3954 + 0{,}1153x_{2i} \qquad i = 1,2,\dots,42$$

Dengan x_{2i} adalah akreditasi sekolah SMA/SMK.

Untuk model $\text{logit}(\boldsymbol{\omega}) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma}$; Pengujian signifikansi parameter secara individu digunakan uji Wald. Diperoleh hanya parameter β_3 yang signifikan pada $\alpha = 0{,}05$ sehingga dapat dibentuk model logitnya sebagai berikut:

$$\text{logit}\,(\boldsymbol{\omega}) = -4{,}2963 + 0{,}1280x_{3i} \qquad i = 1,2,..,n$$

dengan X_3 adalah proporsi guru tiap SMA/SMK yang sudah sertifikasi.

4.2.2. Interpretasi Hasil

Berdasarkan uji signifikansi parameter regresi ZIP diperoleh model

$$\ln(\mu_i) = -0{,}3954 + 0{,}1153x_{2i} \text{ dan logit } (\boldsymbol{\omega}) = -4{,}2963 + 0{,}1280x_{3i}$$
$$i = 1,2,\dots,42$$

Artinya bahwa setiap sekolah yang terakreditasi A menyebabkan penurunan nilai harapan siswa yang tidak lulus sebesar $\exp(0{,}1153) = 1{,}12$ kali sekolah yang

terakreditasi selain A dan penambahan proporsi guru SMA/SMK yang sudah sertifikasi sebesar 1% akan menurunkan jumlah siswa yang tidak lulus UN sebesar
$= 100\left(e^{0,1280)} - 1\right) = 13,66\%$.

5. Kesimpulan dan Saran

Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa:

1. Hasil estimasi parameter pada model regresi ZIP berbentuk implisit. Metode yang digunakan yaitu Algoritma *Expectation Maximization* (EM). Pada tahapan maksimalisasi digunakan pendekatan Newton Rhapson untuk memaksimalkan fungsi Likelihood yang diperoleh dari tahapan Ekspektasi.
2. Model regresi *Zero Inflated* Poisson yang diperoleh pada data ketidaklulusan siswa SMA/SMK di Kota Mataram tahun 2012 sebagai berikut:

   $\ln(\mu_i) = -0,3954 + 0,1153x_{2i}$ dan logit $(\boldsymbol{\omega}) = -4,2963 + 0,1280x_{3i}$

   dengan

   X_2 adalah akreditasi sekolah SMA/SMK

   X_3 adalah proporsi guru SMA/SMK yang sudah sertifikasi

Saran

Penelitian lanjutan dapat menggunakan model regresi *Zero Inflated Negative Binomial* (ZINB) dan model regresi *Zero Inflated Generalized Poisson* (ZIGP) untuk mengatasi masalah *overdispersion* pada regresi Poisson.

Daftar Pustaka

[1] Agresti, A. 1990. *Categorical Data Analysis.* New York: John Wiley & Sons, Inc.

[2] Dempster, A. P, Laird, N. M, Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977. Vol. 39. No. 1: 1-38.

[3] Hall, D, B and Shen, J. Robust Estimation for Zero-Inflated Poisson Regression. *Scandinavian Journal of Statistics*, Blackwell Publishing Ltd. 2009: 1-16.

[4] Hilbe, J, M. 2011. *Negatif Binomial Regression 2nd Edition*. Cambridge: Cambridge University Press.

[5] Hoog, R, V and Craig, A, T. 1995. *Introduction to Mathematical Statistics*. New Jersey: Prentice-Hall Inc.

[6] Ismail, N and Zamani, H. 2013. Estimation of Claim Count Data using Negative Binomial, Generalized Poisson, Zero-Inflated Negative Binomial and Zero-Inflated generalized Poisson Regression Models. *Casualty Actuarial Society E-Forum.* 41(20): 1-28.

[7] Jansakul, N and Hinde, J, P. Score Tests for Zero-Inflated Models. *Computational Statistics and Data Analysis.* 2002. Vol. 40: 75-96.

[8] Khoshgoftaar TM, Gao K, Szabo RM. 2004. Comparing Software Fault Predictions of Pure and Zero-Inflated Poisson Regression Models. *International Journal of System Science* 36(11): 705-715.

[9] Lambert, D. Zero-Inflated Poison Regression with an Application to Defects in Manufacturing. *Technometrics*. 1992. Vol.34: 1-14.