JURNAL GAUSSIAN, Volume 1, Nomor 1, Tahun 2012, Halaman 55-64
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

PEMODELAN REGRESI *ZERO-INFLATED NEGATIVE BINOMIAL* (ZINB) UNTUK DATA RESPON DISKRIT DENGAN *EXCESS ZEROS*

Bayu Ariawan[1], Suparti[2], Sudarno[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staf Pengajar Jurusan Statistika FSM UNDIP

ABSTRACT

Zero-Inflated Negative Binomial (ZINB) regression is one of the methods used in troubleshooting overdispersion due to excessive zero values in the response variable (excess zeros). ZINB regression model was based on the negative binomial distribution resulting from a mixture distribution between Poisson distribution (μ) with μ is value of random variable which gamma distributed.

ZINB regression parameter estimation can be performed by using Maximum Likelihood Estimation (MLE) method then is followed by the EM algorithm (Expectation maximization) procedure and Newton Rhapson. Test the suitability of the model simultaneously performed using Likelihood Ratio test and significance testing parameters individually performed with Wald test statistics. The model is applied to the case of car insurance obtained PT. Insurance of Sinar Mas Semarang Branch in 2010 in the form of data many policyholders filed claims to the PT. Sinar Mas Semarang Branch Insurance. Response variable is the number of claims submitted to the PT. Insurance of Sinar Mas Semarang Branch, while the predictor variable is the age car and the type of coverage that consists of All Risk, Total Lost Only (TLO), and the joint between All Risk and Total Lost Only (TLO). From the analytical result obtained the conclution that the age of the car and the type of coverage affects number of claims filed by the policyholder to the PT. Insurance of Sinar Mas Semarang Branch in 2010.

Keywords: Overdispersion, Excess zeros, Negative Binomial Distribution, Zero-Inflated Negative Binomial (ZINB) Regression

1. PENDAHULUAN

Analisis regresi merupakan teknik analisis yang digunakan untuk menganalisis hubungan antara variabel bebas dan variabel respon dalam suatu penelitian.Pada umumnya analisis regresi digunakan untuk menganalisis data variabel respon yang berupa data kontinu. Namun dalam beberapa aplikasinya, data variabel respon yang akan dianalisis dapat berupadata diskrit. Salah satu model regresi yang dapat digunakan untuk menganalisis hubungan antara variabel respon Y yang berupa data diskrit yang menunjukkan hubungan antara proses diskrit dengan fungsi peluang yang dihasilkan dari N kejadian yang terbentuk dari distribusi poisson disebut regresi Poisson [11].Tetapi model regresi Poisson memiliki keterbatasan pada asumsi variannya yaitu untuk observasi i (i= 1….,n), Var (Y_i) = E(Y_i), sementara untuk data yang bertipe diskrit terkadang terjadi overdispersi yaitu nilai varian lebih besar dari nilai mean atau underdispersi yaitu nilai mean lebih besar dari nilai variansi. Penanganan model yang dapat digunakan untuk mengatasi masalah overdispersi pada data respon bertipe diskrit antara lainadalah model regresi Binomial Negatif , model regresi Quasi-Likelihood dan model regresi *Generalized Poisson*. Salah satu penyebab terjadinya overdispersi yaitu banyaknya nilai nol yang berlebih pada variabel respon (*excess zeros*), sehingga penanganan model yang dapat digunakan untuk mengatasi masalah overdispersi akibat *excess zeros*pada data respon bertipe diskrit antara lain adalah model regresi *Zero-Inflated Poisson* (ZIP), model regresi *Zero-InflatedNegative Binomial* (ZINB), model regresi *Zero-InflatedGeneralized Poisson* (ZIGP) dan model regresi Hurdle.

Dalam penulisan ini, permasalahan yang dibahas adalah penggunaan model regresi *Zero-InflatedNegative Binomial* (ZINB) untuk mengatasi overdispersi pada regresi Poisson, estimasi parameter, analisis kesesuaian model dan signifikansi koefisien *Zero-InflatedNegative Binomial* (ZINB).Penerapannya dalam kasus asuransi mobil PT. Asuransi Sinar Mas Cabang Semarang tahun 2010.

2. TINJAUAN PUSTAKA

2.1. Distribusi Poisson

Distribusi Poisson adalah distribusi nilai-nilai bagi suatu variabel random *Y*, yaitu banyaknya sukses

selama selang waktu tertentu atau dalam daerah tertentu. Misalkan y_i, i = 1,2,… merupakan jumlah kejadian yang muncul dalam selang waktu dengan rata-rata μ_i. Jika Y adalah variabel acak Poisson dengan parameter $\mu > 0$, maka fungsi massa peluangnya adalah

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

μ menyatakan rata-rata banyaknya sukses yang terjadi dalam selang waktu atau daerah tertentu tersebut. Distribusi poisson mempunyai Rata-rata dan variansi keduanya sama dengan μ[1].

2.2. Distribusi Binomial Negatif

Percobaan binomial negatif terdiri atas beberapa usaha dan tiap usaha dengan dua kemungkinan hasil yang dapat diberi nama *sukses* atau *gagal* dan dilakukan sampai tercapai sejumlah sukses tertentu[2]. Fungsi massa peluangnya adalah

$$b^*(y;r,p) = \frac{\Gamma(y+r)}{y!\Gamma(r)}\,p^r(1-p)^y \qquad 0 \le p \le 1,\ r = 1, 2, 3, \ldots, y = 0,1,2,....$$

Distribusi binomial negatif *b*(y;r, p)* mempunyai rataan dan variansi

$$\mu = \frac{r(1-p)}{p} \text{ dan } \sigma^2 = \frac{r(1-p)}{p^2}$$

Distribusi binomial negatif juga dapat terbentuk dari suatu distribusi campuran poisson gamma[6]. Misalkan bahwa variabel acak Y berdistribusi poisson dengan parameter μ dengan μmerupakan nilai dari variabel random yang berdistribusi gamma, yaitu:

$$Y \mid \mu \sim \ poisson(\mu) \text{dan } \mu \sim Gamma(\alpha,\beta)$$

Fungsi massa peluangnya adalah

$$f(y \mid \alpha,\beta) \quad = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(1-\frac{1}{1+\beta}\right)^{y} \qquad \alpha > 0, \quad \beta > 0, y = 0,1,2,..$$

Rataan dan variansinya adalah

$$E[Y] = \alpha\beta \qquad \text{dan} \qquad V[Y] = \alpha\beta + \alpha\beta^2$$

2.3. Metode Maksimum Likelihood

Misalkan $X_1, X_2, \ldots X_n$ adalah sampel random dari populasi dengan densitas f(x; **θ**) dengan $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^T$ maka fungsi likelihood didefinisikan sebagai fungsi densitas bersama dari $x_1,x_2,\ldots,x_n$, sehingga

$$L(\boldsymbol{\theta}) = f(x_1,x_2,\ldots,x_n;\boldsymbol{\theta}) = f(x_1;\boldsymbol{\theta}).f(x_2;\boldsymbol{\theta})\ldots.f(x_n;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

Estimator maksimum likelihood $\hat{\boldsymbol{\theta}}$ adalah nilai $\boldsymbol{\theta}$ yang memaksimalkan fungsi likelihood $L(\boldsymbol{\theta})$. Untuk memperoleh nilai $\hat{\boldsymbol{\theta}}$ yang memaksimumkan $L(\boldsymbol{\theta})$ harus diderivatifkan dengan langkah-langkah sebagai berikut :

1. Nilai $\hat{\boldsymbol{\theta}}$ diperoleh dari derivatif pertama
   $\frac{\partial}{\partial\theta_j}L(\boldsymbol{\theta}) = 0$dengan j=1,2,..,p
2. Nilai $\hat{\boldsymbol{\theta}}$ dikatakan memaksimumkan $L(\boldsymbol{\theta})$ jika
   $\frac{\partial^2}{\partial\theta_j^{\,2}}L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} < 0$ dengan j=1,2,..,p

Selain memaksimumkan fungsi likelihood,nilai$\hat{\boldsymbol{\theta}}$ juga dapat diperoleh dengan memaksimumkan log natural-likelihood (*ln L(θ)*). Dalam banyak kasus dengan diferensiasi digunakan, akan lebih mudah bekerja pada logaritma natural yang dinotasikan dengan *l(θ) = ln L(θ).* Untuk memperoleh nilai $\hat{\boldsymbol{\theta}}$ yang memaksimumkan *ln L(θ)* dapat dilakukan dengan langkah-langkah yang sama seperti dalam memperoleh nilai $\hat{\boldsymbol{\theta}}$ yang memaksimumkan *L(θ)*[2].

2.4. *Generalized Linear Model* (GLM)

Analisis regresi yang responnya termasuk salah satu keluarga eksponensial disebut Generalisasi Model Linier atau lebih dikenal dengan GLM (*Generalized Linear Models*).*Generalized Linear Model* (GLM) merupakan perluasan dari proses pemodelan linier untukpemodelan data yang mengikuti distribusi

probabilitas selain distribusi normal, seperti Poisson, Binomial, multinomial, dan lain-lain.

Ada tiga komponen utama dalam analisis GML seperti diuraikan berikut[1]:

1. Komponen random

Variabel respon Y = ($y_1 y_{2,}\ldots,y_n$) saling bebas dan memiliki distribusi yang termasuk dalam keluarga eksponensial

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

Parameter θ_i disebut dengan parameter natural dan nilainya dapat berbeda untuk i=1,2,..,n.

2. Komponen Sistematis

Kontribusi variabel prediktor dalam model dinyatakan dalam bentuk kombinasi linier antara parameter (η) dengan parameter regresi yang akan diestimasi.

$$\eta_i=\beta_0+\beta_1 x_{i1}+....+\beta_p x_{ip}=\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}$$

Atau dalam matriks dituliskan dalam bentuk

$\boldsymbol{\eta = X\beta}$

η adalah vektor (n x 1), X adalah matriks (n x c) dari variabel bebas, β adalah matriks (c x 1) dari koefisien regresi, dengan *c=p+1*

3. Fungsi *link*

Fungsi *link, g(.)* adalah fungsi yang menghubungkan ekspektasi variabel respon $E[Y_i]=\mu_i$ dengan prediktor linier.

$\eta_i = g(\mu_i)$ dengan i=1,2,…,n

2.5. Regresi Poisson

Model regresi Poisson adalah model regresi nonlinear yang berasal dari distribusi Poisson yang merupakan penerapan dari *Generalized Linear Model* (GLM) yang menggambarkan hubungan antara variabel dependen dengan variabel independen, dengan variabel dependen merupakan bentuk diskrit. Regresi Poisson mempunyai asumsi E(Y) = Var(Y). Berdasarkan konsep GLM untuk distribusi Poisson bahwa pada saat $g(\mu_i)$ sama dengan parameter natural θ_i ($(g(\mu_i)=\theta_i=\ln(\mu_i))$, sehingga kanonikal link (fungsi yang mentransformasikan nilai mean ke parameter natural) adalah log natural link : $g(\mu_i)=\ln(\mu_i)$. Sehingga hubungan μ_i dengan prediktor linier η_i, dinyatakan dengan $\ln(\mu_i)=\eta_i$. Dengan mengunakan fungsi *link* log natural tersebut diperoleh model regresi Poisson dalam bentuk :

$$\ln\mu_i=\eta_i$$

$$\ln\mu_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}$$

$$\mu_i=\exp(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip})$$

denganμ_i nilai ekspektasi y_i berdistribusi Poisson dengan i = 1,2,3,…,n. Penaksiran koefisien parameter-parameter dalam regresi poisson menggunakan metode *Maximum Likelihood Estimation* (MLE) dan diiterasikan dengan menggunakan metode iterasi Newton-Rhapson.

2.6. Permasalahan Pada Regresi Poisson

2.6.1. Overdispersi

Overdispersi adalah nilai variansnya lebih besar dari nilai meannya. Untuk mendeteksi terjadinya masalah overdispersi dalam model regresi poisson dapat dilihat dengan menguji hubungan antara varian dan mean dalam bentuk persamaan :

$$V(\mu_i)=\phi\mu_i$$

Untuk menghitung nilai ϕ dilakukan dengan melakukan pendekatan nilai *Pearson's Chi Square* yang didefinisikan sebagai berikut :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\text{var}(\hat{\mu}_i)} \quad \text{dengan} \quad \hat{\phi} = \frac{\chi^2}{n - p - 1}$$

2.6.2. *Excess zeros*

Salah satu permasalahan pada regresi poisson yaitu nilai nol yang berlebih *(Excess zeros)*.Pada Variabel respon pada data diskrit mungkin ditemukan data untuk kosong/tak terisi (bernilai nol).Akan tetapi, dalam banyak kasus, kosong memiliki arti penting pada penelitian yang bersangkutan.Jika nilai nol memiliki arti penting dalam data diskritmaka data tersebut harus dimasukkan dalam analisis.*Excess zeros* dapat dilihat pada proporsi variabel respon yang bernilai nol lebih besar dari data diskrit lainnya. *Excess zeros* merupakan salah satu penyebab terjadinya overdispersi.

2.7. Regresi *Zero-Inflated Negative Binomial* (ZINB)

Regresi*Zero-Inflated Negative Binomial* (ZINB) merupakan model yang dibentuk dari distribusi campuran poisson gamma [6].

Jika Y_i adalah variabel random independen yang diskrit dengan i = 1,2,3,…,n , nilai nol pada observasi diduga muncul dalam dua cara yang sesuai untuk keadaan (*state*) yang terpisah. Keadaan pertama disebut *zero state* terjadi dengan probabilitas p_i dan menghasilkan hanya observasi bernilai nol, sementara keadaan kedua disebut *Negative Binomialstate* terjadi dengan probabilitas (1 - p_i) dan berdistribusi Binomial Negatif dengan mean μ , dengan $0 \leq p_i \leq 1$. Proses dua keadaan ini dengan variabel Y_i memberikan distribusi campuran dua komponen dan didapat fungsi probabilitas sebagai berikut :

$$P(Y_i = yi) = \begin{cases} p_i + (1-p_i)\left(\frac{1}{1+k\mu_i}\right)^{1/k} & , \text{untuk } y_i = 0 \\ (1-p_i)\frac{\Gamma(y_i + 1/k)}{\Gamma(1/k)\Gamma(y_i+1)}\left(\frac{1}{1+k\mu_i}\right)^{1/k}\left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i} & , \text{untuk } y_i = 1,2,.. \end{cases}$$

Dengan i = 1,2,3,…,n; $0 \leq p_i \leq 1$, $\mu_i \geq 0$, k adalah parameter tersebar dengan $1/k > 0$ dan Γ(.) adalah fungsi gamma. Mean dan variansinya didefinisikan $E(Y_i) = (1-p_i)\mu_i$ dan $Var(Y_i) = (1-p_i)\mu_i(1 + \mu_i k + p_i\mu_i)$. Ketika $p_i = 0$, variabel random Y_i berdistribusi binomial negatif dengan mean μ_i dan parameter dispersi k, sehingga $Y_i \sim NB(\mu_i, 1/k)$. Diasumsikan bahwa parameter μ_i dan p_i masing - masing bergantung pada variabel x_i dan z_i, sehingga model dari regresi ZINB dibagi menjadi dua komponen model yaitu:

1. Model data diskrit untuk μ_i adalah

   $\ln(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, $\mu_i \geq 0$, i=1,…,n.

x_i adalah matriks variabel yang memuat himpunan-himpunan yang berbeda dari faktor eksperimen yang berhubungan dengan peluang pada *mean Negative Binomial* pada *Negative Binomial state*.

2. Model *zero-Inflation* untuk p_i adalah

   $\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{z}_i^T \boldsymbol{\gamma}$, $0 \leq p_i \leq 1$, i=1,…,n.

z_i adalah matriks variabel yang memuat himpunan-himpunan yang berbeda dari faktor eksperimen yang berhubungan dengan peluang pada *zero state.*

Pengaruh dari masing - masing matriks kovariat x_i dan z_i terhadap μ_i dan p_i bisa sama atau tidak sama, jika masing – masing matriks kovariat memberikan pengaruh yang sama terhadap μ_i dan p_i maka matrix $x_i = z_i$, sehingga modelnya menjadi :

1. Model data diskrit untuk μ_i adalah

   $\ln(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, $\mu_i \geq 0$, i=1,…,n.

2. Model z*ero-Inflation* untuk p_i adalah

   $\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$, $0 \leq p_i \leq 1$, i=1,…,n.

x_i adalah matriks variabel yang memuat himpunan-himpunan yang berbeda dari faktor eksperimen yang berhubungan dengan peluang *zero state* dan *mean Negative Binomial* pada *Negative Binomial state*, sedangkan β dan γ adalah parameter regresi yang akan ditaksir[4].

2.8. Estimasi Parameter Regresi *Zero-Inflated Negative Binomial (*ZINB)

Estimasi parameter regresi ZINB menggunakan metode *Maximum Likelihood Estimation* (MLE) dengan prosedur Algoritma EM (*Expectation Maximization*) dan Newton Rhapson.Metode ini biasanya digunakan untuk menaksir parameter suatu model yang diketahui fungsi densitasnya.sehingga fungsi log-likelihood dari fungsi probabilitas ZINB adalah :

$$\ln L(\boldsymbol{\theta}\mid y_i)=\begin{cases}\sum_{i=1}^{n}\ln\left\{e^{x_i^T\gamma}+\left(\frac{1}{1+ke^{x_i^T\beta}}\right)^{1/k}\right\}-\sum_{i=1}^{n}\ln\left[1+e^{x_i^T\gamma}\right] & \text{, untuk } y_i=0\\ -\sum_{i=1}^{n}\ln\left[1+e^{x_i^T\gamma}\right]+\sum_{i=1}^{n}\ln\left[\Gamma(1/k+y_i)\right]-\sum_{i=1}^{n}\ln\left[\Gamma(y_i+1)\right]-\sum_{i=1}^{n}\ln\left[\Gamma(1/k)\right] & \\ +y_i\sum_{i=1}^{n}\ln\left(\frac{e^{x_i^T\gamma}}{1+e^{x_i^T\gamma}}\right)^{y_i}+1/k\sum_{i=1}^{n}\ln\left(\frac{1}{1+ke^{x_i^T\beta}}\right)^{1/k} & \text{,untuk } y_i=1,2,..\end{cases}$$

dengan i = 1,2,3,…,n. Estimasi dengan maksimum likelihood rasio dihitung dengan memaksimalkan log-likelihoodny. Karena fungsi log-likelihoodnya tidak linier jika tidak digunakan nilai awal yang bagus, sehingga fungsi likelihood ini tidak dapat diselesaikan dengan metode numerik biasa.Sehingga digunakanlah algoritma EM (*Expectation Maximization*)[4].

Misalkan variabel y_i (i = 1,2,3,…,n) berkaitan dengan vektor variabel indikator $W = (w_1,\ldots,w_n)^T$yaitu:

$$w_i=\begin{cases}1, \text{ jika } y_i \text{ berasal dari } \textit{zero state}\\ 0, \text{ jika } y_i \text{ berasal dari } \textit{Negative Binomial state}\end{cases}$$

dengan i = 1,2,3,…,n , jika nilai variabel respon y_i=1,2,… maka nilai w_i = 0. Sedangkan jika nilai variabel respon y_i = 0, maka nilai w_i mungkin 0 mungkin 1. Oleh karena itu, nilai w_i dianggap hilang. Peluang dari w_i dapat dinyatakan :

$P(w_i = 1) = p_i$

$P(w_i = 0) = 1- p_i$

dengan i = 1,2,3,…,n,Sehingga distribusi dari variabel W adalah $w_i \sim$ Binomial (1, p_i) mempunyai rataan dan variansi $E(w_i) = p_i$ dan $var(w_i) = p_i(1- p_i)$. Distribusi gabungan antara y_i dan w_i yang terbentuk yaitu

$$f(w_i, y_i \mid p_i, \mu_i)= (p_i)^{w_i}(1-p_i)^{(1-w_i)}\left[\frac{\Gamma(y_i+1/k)}{\Gamma(1/k)\Gamma(y_i+1)}\left(\frac{1}{1+k\mu_i}\right)^{1/k}\left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i}\right]^{(1-w_i)}$$

didapat persamaan log-likelihoodnya :

$$\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}|y_i,w_i)=\sum_{i=1}^{n}\left\{w_i x_i^T\gamma-\ln\left[1+\exp(x_i^T\gamma)\right]+(1-w_i)\ln\left[g(y_i;\beta,1/k)\right]\right\}$$

dimana $g(y_i;\beta,1/k)=\frac{\Gamma(y_i+1/k)}{\Gamma(1/k)\Gamma(y_i+1)}\left(\frac{1}{1+k\mu_i}\right)^{1/k}\left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i}$, dan $\mu_i = e^{x_i^T\boldsymbol{\beta}}$dengan i = 1,2,3,…,n.

Algoritma EM dibagi menjadi dua langkah yaitu

1. Tahap ekspektasi (*E-Step*)

   Mengganti variabel w_i dengan $w_i^{(m)}$ yang merupakan ekspektasi dari w_i

   $$w_i^{(m)} = E(w_i|y_i,\boldsymbol{\gamma}^{(m)},\boldsymbol{\beta}^{(m)})$$

   $$=\begin{cases}\left(1+(e^{x_i^T\gamma^{(m)}})\left[\frac{1}{1+k^{(m)}e^{x_i^T\beta^{(m)}}}\right]^{1/k^{(m)}}\right)^{-1}, & \textit{jika } y_i=0\\ 0, & \textit{jika } y_i=1,2,..\end{cases}$$

   Sehingga

$$Q(\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)})=E_{\theta^{(k)}}\{\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}|y_i,w_i)\,|y_i,\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}\}=\sum_i^n \ln L(\boldsymbol{\gamma}^{(m)}\,|y_i,w_i^{(m)})+\sum_i^n \ln L(\boldsymbol{\beta}^{(m)}\,|y_i,w_i^{(m)})$$

   dimana

$$\ln L(\boldsymbol{\gamma}^{(m)} \left| y_i, w_i^{(m)} \right.) = \sum_{i=1}^{n} \left[w_i^{(m)} x_i^T \gamma - \ln\left(1 + e^{x_i^T \gamma}\right) \right]$$

$$\ln L(\boldsymbol{\beta}^{(m)} \left| y_i, w_i^{(m)} \right.) = \sum_{i=1}^{n} (1 - w_i^{(m)}) \left\{ \frac{\Gamma(1/k + y_i)}{\Gamma(y_i + 1)\Gamma(1/k)} \left(\frac{e^{x_i^T \gamma}}{1 + e^{x_i^T \gamma}} \right)^{y_i} \left(\frac{1}{1 + k e^{x_i^T \beta}} \right)^{1/k} \right\}$$

2. Tahap maksimalisasi (*M-step*)

Memaksimalkan $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ dari hasil *E-Step* dengan menghitung $\boldsymbol{\beta}^{(m+1)}$ dan $\boldsymbol{\gamma}^{(m+1)}$ dengan metode Newton-Raphson (Hall, 2000)

2.9. Pengujian Parameter Regresi ZINB

2.9.1. Pengujian Kesesuaian Model Regresi ZINB

Pengujian kesesuaian model regresi ZINB adalah dengan menggunakan Likelihood Ratio (LR) Test dengan prosedur pengujian :

Hipotesis :

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = \gamma_1 = \gamma_2 = \ldots = \gamma_p = 0$

H_1 : paling sedikit ada satu $\beta_j \neq 0$ atau $\gamma_j \neq 0$, dengan j = 1, 2, …, p

dengan β_j adalah parameter ke-j dari model $\ln(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ dengan i=1,…,n , γ_j adalah parameter ke-j dari model logit(p_i)= $\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$ dengan i=1,…,n.

Statistika uji :

$G = -2 \ln\left[\frac{L_0}{L_1}\right] = -2 (\ln L_0 - \ln L_1)$

$= -2\{[\ln L(\beta_0 | y_i, w_i) + \ln L(\gamma_0 | y_i, w_i)] - [\ln L(\boldsymbol{\beta} | y_i, w_i) + \ln L(\boldsymbol{\gamma} | y_i, w_i)]\}$

$G \sim \chi^2_p$

Kriteria uji :

Tolak H_0 pada taraf signifikansi α jika $G_{hitung} > \chi^2_{\alpha;2p}$

2.9.2. Pengujian Signifikansi Parameter Regresi ZINB secara Individu

a. Pengujian signifikansi parameter model $\ln(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ dengan i=1,…,n.

Hipotesis :

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$,

Untuk setiap j = 1,2,…, p

Statistika uji :

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)^2$$

$W_j \sim \chi^2_1$

Kriteria uji :

Tolak H_0 pada taraf signifikansi α jika $W_j > \chi^2_{\alpha;1}$

b. Pengujian signifikansi parameter model $\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$ dengan i=1,…,n.

Hipotesis :

$H_0 : \gamma_j = 0$

$H_1 : \gamma_j \neq 0$

Untuk setiap j = 1,2,…, p

Statistika uji :

$$W_j = \left(\frac{\hat{\gamma}_j}{SE(\hat{\gamma}_j)} \right)^2$$

$W_j \sim \chi^2_1$

Kriteria uji :

Tolak H_0 pada taraf signifikansi α jika $W_j > \chi^2_{\alpha;1}$

3. METODOLOGI PENELITIAN

3.1. Jenis dan Sumber Data

Data yang digunakan pada penulisan ini berupa data sekunder tentang asuransi mobil yang diperoleh dari PT. Asuransi Sinar Mas Cabang Semarang tahun 2010 [10].Data yang digunakan berupa data banyaknya klaim yang diajukan pemegang polis kepada pihak PT. Asuransi Sinar Mas Cabang Semarang, umur mobil, dan jenis pertanggungan asuransi. Jenis pertanggungan asuransi terdiri dari *All Risk*, *Total Lost Only* (TLO), serta gabungan antara *All Risk* dan *Total Lost Only* (TLO) dengan total sebanyak 406 data.

3.2. Variabel Data

Variabel data yang digunakan dalam penulisan ini yaitu banyaknya klaim yang diajukan kepada pihak PT. Asuransi Sinar Mas Cabang Semarang sebagai variabel respon (Y) dan variabel prediktor (X) meliputi :

1. Umur mobil (X_1)
2. Jenis pertanggungan asuransi 1 (X_2)dengan variabel dummy dengan dua kategori yaitu 1 untuk jenis pertanggungan*Total Lost Only* (TLO) dan 0 untuk jenis pertanggungan lainnya.
3. Jenis pertanggungan asuransi 2 (X_3) dengan variabel dummy dengan dua kategori yaitu 1 untuk jenis pertanggungan gabungan *All Risk* dan *Total Lost Only* dan 0 untuk jenis pertanggungan lainnya.

3.3. Teknik Pengolahan Data

Data yang digunakan diolah dengan menggunakan *software* R 2.15 (dengan menggunakan *packagefield*, *MASS*, *pscl*, dan *lmtest*)[12]. Langkah-langkah analisis data yang digunakan dalam penulisan ini adalah sebagai berikut:

1. Melakukan uji Kolmogorov-Smirnov untuk menguji apakah variabel respon Y mengikuti distribusi Poisson atau tidak.
2. Menentukan model regresi Poisson.
3. Menguji asumsi equidispersi model regresi Poisson dengan uji *Pearson's chi-square*. Menentukan model akhir regresi Poisson jika asumsi equidispersi terpenuhi.
4. Jika terjadi overdispersi, kemudian melihat apakah variabel respon Y mengalami *excess zeros* atau tidak dengan melihat proporsi nilai nol.
5. Menentukan model regresi ZINB.
6. Melakukan mengujian kesesuaian model regresi ZINB.
7. Melakukan pengujian signifikansi parameter secara individu regresi ZINB.

4. HASIL DAN PEMBAHASAN

4.1. Pengujian Distribusi Poisson pada Variabel Respon Y

Pengujian distribusi Poisson pada variabel respon Y yaitu banyaknya klaim yang diajukan kepada pihak perusahaan asuransidilakukan dengan uji Kolmogorov-Smirnov dengan prosedur pengujian yaitu :

Hipotesis

H_0 = Data variabel respon Y mengikuti distribusi Poisson

H_1 = Data variabel respon Y tidak mengikuti distribusi Poisson

Dengan taraf signifikansi α = 5%diperoleh nilai D = 0.030 dannilai *asymp.sig* (*2-tailed*) = 0.845 .Pada Tabel Kolmogorov Smirnov didapat nilai $D^*_{(0.05)}$ untuk n = 406 yakni sebesar 0.068.Karena nilai $D < D^*_{(0.05)}$(0.030 < 0.068) atau p-value > α (0.845>0.05) , maka H_0 diterima dan disimpulkan bahwa data variabel respon Y mengikuti distribusi Poisson.

4.2. Pemodelan Regresi Poisson

Estimasi parameter didapatkan model regresi Poisson seperti yang terlihat sebagai berikut :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.618867   0.204580  -3.025  0.00249 **
X1           0.001357   0.039600   0.034  0.97266
X2          -1.528625   0.256587  -5.958 2.56e-09 ***
X3          -0.079836   0.205319  -0.389  0.69740
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model regresi poisson yang terbentuk :

$$\mu_i = e^{(-0.618867+0.001357X_{i1}-1.528625X_{i2}-0.079836X_{i3})}, \mu_i \geq 0,\ i=1,\ldots,406.$$

$\mu_i$adalah nilai harapan dari banyaknya klaim yang ke-i.

4.3. Pengujian Asumsi Equidispersi Model Regresi Poisson

Pengujian equidispersi ini dapat dilakukan menggunakan uji Pearson chi-square dengan prosedur pengujian sebagai berikut :

Hipotesis

$H_0 : \phi \leq 1$, (tidak terjadi overdispersi)

$H_1 : \phi > 1$, (terjadi overdispersi)

Dari output R 2.15 didapat nilai pearson 1.04278. Karena 1.04278 > 1, maka H_0 ditolak dan disimpulkan data variabel respon terjadi overdispersi, karena terjadi overdispersi maka dilanjutkan pengujian apakah variabel respon mengalami excess zeros atau tidak.

4.5. Pengujian *Excess zeros* pada Variabel Respon

Pengujian apakah variabel respon mengalami excess zeros atau tidak dapat dilihat pada Tabel 1sebagai berikut :

Tabel 1*Excess zeros* pada Variabel Respon Y

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	67	16.5	72.0	72.0
	2	23	5.7	24.7	96.8
	3	3	.7	3.2	100.0
	Total	93	22.9	100.0	
Missing	0	313	77.1		
Total		406	100.0		

Dari Tabel 1 dapat dilihat bahwa nilai nol mempunyai proporsi tertinggi dengan 77.1% (313) dan melebihi proporsi nilai diskrit lainnya, sehingga dapat disimpulkan bahwa variabel respon Y mengalami *excess zeros*. Salah satu model yang digunakan untuk menangani keadaan overdispersi dan mengalami *excess zeros* yaitu model regresi *Zero-InflatedNegative Binomial* (ZINB).

4.6. Pemodelan Regresi *Zero-InflatedNegative Binomial*(ZINB)

Dari estimasi parameter didapatkan model regresi *Zero-Inflated Negative Binomial*(ZINB) seperti yang terlihat sebagai berikut:

```
Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.64202    0.30515   2.104  0.03539 *
X1          -0.17495    0.05872  -2.980  0.00289 **
X2          -1.00345    0.41865  -2.397  0.01653 *
X3          -0.50493    0.26829  -1.882  0.05983 .

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0023     0.6435   3.112 0.001860 **
X1           -0.6896     0.1991  -3.464 0.000532 ***
X2            1.4520     0.8539   1.701 0.089029 .
X3           -1.7910     0.8972  -1.996 0.045927 *
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model awal *Zero-Inflated Negative Binomial*(ZINB) yang terbentuk :

1. Model data diskrit untuk μ_i yaitu

$$\mu_i = e^{(0.64202-0.17495X_{i1}-1.00345X_{i2}-0.50493X_{i3})}, \ \mu_i \geq 0, \ i=1,\ldots,406.$$

$\mu_i$adalah nilai harapan dari banyaknya klaim yang ke-i.

2. Model *zero-inflation* untuk p_i yaitu

$$p_i = \frac{e^{(2.0023-0.6896X_{i1}+1.4520X_{i2}-1.7910X_{i3})}}{1+e^{(2.0023-0.6896X_{i1}+1.4520X_{i2}-1.7910X_{i3})}}, 0 \leq p_i \leq 1, \ i=1,\ldots,406.$$

p_i adalah peluang resiko pemegang polis tidak mengajukan klaim yang ke-i.

Dengan

X_{i1} = umur mobil
X_{i2} = jenis pertanggungan *Total Lost Only* (TLO)
X_{i3} = jenis pertanggungan gabungan *All Risk* dan *Total Lost Only*(TLO)
Interpretasi dari koefisien regresi untuk model $\ln(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ dimana i=1,2,3,…,406:
X_{i1} → setiap perubahan satu tahun dalam umur mobil menyebabkan penurunan nilai harapan banyaknya klaim sebesar $e^{(-0.17495)} = 0.8395$.
X_{i2} → setiap pemilihan jenis pertanggungan TLO menyebabkan penurunan nilai harapan banyaknya klaim $e^{(-1.00345)} = 0.3666$ kali lebih kecil dari jenis pertanggungan *All Risk.*
X_{i3} → setiap pemilihan jenis pertanggungan gabungan *All Risk* dan TLO menyebabkan penurunan nilai harapan banyaknya klaim sebesar $e^{(-0.50493)} = 0.6035$ kali lebih kecil dari jenis pertanggungan *All Risk.*

Interpretasi dari koefisien regresi untuk model logit(p_i)= $\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$ dimana i=1,2,3,…,406 :
X_{i1} → setiap perubahan satu tahun dalam umur mobil menyebabkan kecenderungan menurunnya peluang resiko pemegang polis tidak mengajukan klaim sebesar $e^{(-0.6896)} = 0.50177$.
X_{i2} → setiap pemilihan jenis pertanggungan TLO menyebabkan kecenderungan meningkatnya peluang resiko pemegang polis tidak mengajukan klaim $e^{(1.4520)} = 4.2716$ kali lebih besar dari jenis pertanggungan *All Risk*.
X_{i3} → setiap pemilihan jenis pertanggungan gabungan *All Risk* dan TLO menyebabkan kecenderungan menurunnya peluang resiko pemegang polis tidak mengajukan klaim $e^{(-1.7910)} = 0.16679$ kali lebih kecil daripada jenis pertanggungan *All Risk*.

4.6.1. Pengujian Kesesuaian Model Regresi *Zero-Inflated Negative Binomial*(ZINB)

Pengujian ketepatan model regresi *Zero-Inflated Negative Binomial*(ZINB) adalah dengan menggunakan uji Likelihood Ratio (LR) dengan prosedur pengujian:
Hipotesis
$H_0 : \beta_1 = \beta_2 = \beta_3 = \gamma_1 = \gamma_2 = \gamma_3 = 0$
H_1 : paling sedikit ada satu $\beta_j \neq 0$ atau $\gamma_j \neq 0$, dengan j = 1,2,3.
Menggunakan Taraf Signifikansidiambil α = 5%. Dari Output R 2.15 diperoleh nilai

$$G = -2(-283.87 - (-250.66)) = 66.42$$

dengan P-value =$2.211e^{-12}$,tabel χ^2, nilai $\chi^2_{(0.05;6)} = 12.59$. H_0 ditolak karena $G_{hitung} > \chi^2_{(\alpha;2p)}$(66.42>12.59) atau P-value <α ($2.211e^{-12} < 0.05$) dan disimpulkan model regresi *Zero-Inflated Negative Binomial*(ZINB) dapat digunakan.

4.6.2. Pengujian Signifikansi Parameter Regresi ZINB secara Individu

Tolak H_0 jika $W_j > \chi^2_{\alpha;1}$ atau p-value < α, dari tabel χ^2, nilai $\chi^2_{(0.05;1)} = 3.841$.

Tabel 2 Pengujian Signifikansi Parameter Regresi ZINB secara Individu

Parameter β

		Zj	Wj	Pvalue	Keputusan	Kesimpulan
β1	Umur Mobil	-2.98	8.8804	0.00289	H0 Ditolak	Koefisien Signifikan
β2	Jenis Pertanggungan TLO	-2.396	5.74082	0.01653	H0 Ditolak	Koefisien Signifikan
β3	Jenis Pertanggungan gabungan All Risk dan TLO	-1.882	3.54192	0.05983	H0 Diterima	Koefisien Tidak Signifikan

Karena koefisien $\beta_1$dan $\beta_2$signifikan, maka ada pengaruh umur mobil dan jenis pertanggungan TLO masing-masing terhadap besarnya nilai harapan banyaknya klaim.Sedangkan koefisien β_3 tidak signifikan, maka tidak ada pengaruh jenis pertanggungan gabungan antara *All Risk* dan TLO terhadap besarnya nilai harapan banyaknya klaim.

Parameter γ

		Zj	Wj	Pvalue	Keputusan	Kesimpulan
γ1	Umur Mobil	-3.464	11.9993	0.00053	H0 Ditolak	Koefisien Signifikan
γ2	Jenis Pertanggungan TLO	1.701	2.8934	0.08903	H0 Diterima	Koefisien Tidak Signifikan
γ3	Jenis Pertanggungan gabungan All Risk dan TLO	-1.996	3.98402	0.04593	H0 Ditolak	Koefisien Signifikan

Koefisien γ_1 dan γ_3 signifikan, maka ada pengaruh umur mobil terhadap besarnya peluang resiko tidak mengajukan klaim.Sedangkan Koefisien γ_2 tidak signifikan, maka tidak ada pengaruh jenis pertanggungan TLO terhadap besarnya peluang resiko tidak mengajukan klaim.

5. KESIMPULAN

Salah satu penyebab terjadinya overdispersi adalah banyaknya nilai nol (*excess zeros*) pada variabel respon. Salah satu metode yang digunakan dalam mengatasi masalah overdispersi dan mengalami *excess zeros* tersebut adalah metode regresi *Zero-InflatedNegative Binomial* (ZINB).Distribusi yang digunakan dalam model regresi *Zero-InflatedNegative Binomial* (ZINB) adalah distribusi binomial negatif yang dihasilkan dari distribusi campuran poisson gamma.Model regresi *Zero-InflatedNegative Binomial* (ZINB) yang terbentuk dibagi menjadi dua komponen model yaitu:

a. Model data diskrit untuk μ_i adalah

$$\ln(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} \text{ atau } \mu_i = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}, \qquad \mu_i \geq 0,\ i=1,\ldots,n.$$

b. Model z*ero-Inflation* untuk p_i adalah

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\gamma} \text{ atau } p_i = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}},\ 0 \leq p_i \leq 1,\ i=1,\ldots,n$$

Dari hasil analisis terhadap dalam kasus asuransi tahun 2010 di PT. Asuransi Sinar Mas Cabang Semarang untuk menguji pengaruh umur mobil dan jenis pertanggungan asuransi terhadap banyaknya klaim yang diajukan pemegang polis kepada pihak PT. Asuransi Sinar Mas Cabang Semarang. Adapun model regresi ZINB yang diperoleh adalah sebagai berikut :

1. Model data diskrit untuk μ_i yaitu

$$\mu_i = e^{(0.64202-0.17495X_{i1}-1.00345X_{i2}-0.50493X_{i3})}, \quad \mu_i \geq 0,\ i=1,\ldots,406.$$

$\mu_i$adalah nilai harapan dari banyaknya klaim yang ke-i.

2. Model *zero-inflation* untuk p_i yaitu

$$p_i = \frac{e^{(2.0023-0.6896X_{i1}+1.4520X_{i2}-1.7910X_{i3})}}{1+e^{(2.0023-0.6896X_{i1}+1.4520X_{i2}-1.7910X_{i3})}}, 0 \leq p_i \leq 1,\ i=1,\ldots,406.$$

p_i adalah peluang resiko pemegang polis tidak mengajukan klaim yang ke-i.

Dengan

X_{i1} = umur mobil

X_{i2} = jenis pertanggungan *Total Lost Only*(TLO)

X_{i3} = jenis pertanggungan gabungan *All Risk* dan *Total Lost Only*(TLO)

6. DAFTAR PUSTAKA

1. Agresti, A. 2002. *Categorical Data Analysis*. Second Edition. New York : John Wiley and Sons, Inc.
2. Casella, G and Berger, R. L. 1990. *Statistical Inference*. California :Wadsworth, INC.
3. Daniel, W.W. 1989. *Statistika Nonparametrik Terapan*. Jakarta : PT Gramedia.
4. Garay, A.M., Hashimoto, E.M. 2011. *On Estimation And Influence Diagnostics For Zero-Inflated Negative Binomial Regression Models*. Computational Statistics and Data Analysis Vol.55. pp. 1304–1318.
5. Hall, D.B. 2000. "*Zero-Inflated Poisson and Binomial Regression with Random Effects : A Case Study*". *Biometrics*. Vol.56. pp. 1030-1039.
6. Hilbe, J.M. 2007. *Negative Binomial Regression*. New York : Cambridge University Press.
7. Istiana, N. 2011. *Count Regression Models*. (http://www.nofitaistiana.wordpress.com, diakses 13 April 2012).
8. Jiang, J. 2007. *Linier and Generalized Linear Mixed Model and Their Applications*. New York : Springer Science+Business Media, LLC.
9. McLachlan, G.J., Krishran, T. 2008. *The EM Algorithm and Extensions 2nd Edition*. New York :a John Wiley &Sons, Inc.
10. Taufan, M. 2011. *Pemodelan Regresi Zero-Inflated Poisson Tentang Faktor-Faktor Yang Mempengaruhi Banyaknya Klaim Asuransi Kendaraaan Bermotor.* Semarang:Undip Press
11. Winkelmann, R. 2008. *Econometric Analysis of Count Data 5th edition*. Berlin: Springer.
12. Zuur, A.F., Leno, E.N., Walker, N.J. 2009. *Mixed Effects Models and Extensions in Ecology with R*. New York : Springer Science+Business Media, LLC.