JKMA | Jurnal Kajian Matematika dan Aplikasinya
VOLUME 1 NOMOR 1, July 2020
e-ISSN: 2722-7650. http://journal2.um.ac.id/index.php/jkma

ESTIMASI PARAMETER REGRESI *ZERO-INFLATED NEGATIVE BINOMIAL* DENGAN METODE ALGORITMA *EXPECTATION MAXIMIZATION* (EM) (STUDI KASUS: PENYAKIT DIFTERI DI JAWA BARAT TAHUN 2016)

Dimas Adi Pradana[1], Trianingsih Eni Lestari[1, *]
[1] Jurusan Matematika, FMIPA, Universitas Negeri Malang

Email: dimasadip12345@gmail.com (*D. A. Pradana*), trianingsih.eni.fmipa@um.ac.id (*T. E. Lestari*)
* Corresponding Author

Abstract
This study aims to determine the parameter estimation method of the ZINB model with the Expectation Maximization (EM) algorithm, and also to overcome the case of overdispersion caused by excess zeros in diphtheria cases in West Java in 2016 and to know the factors that significantly influence. Based on the results of the analysis, the factors that significantly influence the number of diphtheria cases in West Java Province in 2016 for the data count are the percentage of DPT immunization (X1), the percentage of residents who have access to safe drinking water (X3) and the percentage of TPM that meet sanitary hygiene requirements (X6). Whereas for zero-inflation data is the percentage of residents who have access to decent drinking water (X3), number of puskesmas (X5) and the percentage of TPM that meet sanitary hygiene requirements (X6).

Keywords: *regression, ZINB, overdispersion, diphtheria, Expectation Maximization (EM) algorithm*

Submitted:14 January 2020; Revised: 03 April 2020; Accepted Publication: 29 May 2020;
Published Online: July 2020
DOI: 10.17977/um055v1i1p18-26

PENDAHULUAN

Analisis regresi merupakan suatu metode untuk melihat hubungan variabel terikat dengan faktor-faktor yang mempengaruhinya. Pada analisis regresi terdapat kasus dimana variabel terikat tidak mengikuti sebaran normal dan berupa data *count*. Salah satu model yang dapat digunakan pada kondisi tersebut adalah regresi Poisson.

Regresi Poisson memiliki suatu asumsi yaitu nilai varian dan rata-rata dari variabel terikat harus sama (*equidispersion*) (Hilbe, 2011). Namun kenyataannya sering ditemukan data yang nilai variansinya lebih kecil dari rata-rata (*underdispersion*) atau nilai variansinya lebih besar dari rata-rata (*overdispersion*). Salah satu penyebab *overdispersion* pada regresi Poisson adalah kelebihan nilai nol pada variabel terikat (*excess zeros*).

Lambert (1992:1) mengatakan bahwa data yang terdapat kasus *excess zeros* dapat dimodelkan dengan menggunakan regresi *Zero-Inflated Poisson* (ZIP). Regresi ini hanya mampu untuk menangani data *excess zeros*, tidak untuk kasus *overdispersion.* Oleh karena itu Ismail & Zamani (2013:3) menyarankan bahwa data yang terdapat *excess zeros* dan *overdispersion* lebih sesuai menggunakan regresi *Zero-Inflated Negative Binomial* (ZINB) dari pada *Zero-Inflated Poisson* (ZIP). Hal ini dikarenakan jika terdapat kasus *overdispersion*, maka estimasi parameter pada ZIP dapat menjadi bias. Berikut adalah beberapa penelitian terdahulu menggunakan ZINB seperti Weng, et al. (2016) untuk melihat hubungan faktor-faktor penyebab jumlah orang meninggal pada kasus kecelakaan kapal laut di Laut Cina Selatan tahun 2001-2010 dimana estimasi parameternya menggunakan suatu teknik iterasi, yaitu algoritma *Expectation-Maximization* (EM) yang mana stabil dan konvergen. Selain pada kasus kecelakaan, data *count* juga terdapat di kasus-kasus penyakit tertentu seperti penyakit

difteri. Penelitian terkait difteri telah dilakukan sebelumnya. Salah satunya Faidah & Pontoh (2015) yang menerapkan metode Hurdle Poisson (kombinasi model logit dan *truncated poisson*) pada kasus difteri yang memiliki nilai nol berlebih (*excess zeros*) dan terjadi overdispersi.

Model *zero-inflated* adalah campuran dari dua distribusi yang terdiri dari distribusi delta (*perfect state*) dan suatu distribusi pada bilangan bulat tidak negatif (*imperfect state*). Secara umum, *perfect state* dimisalkan dengan probabilitas p dan *imperfect state dengan* probabilitas 1-p. Dalam model regresi *Zero Inflated Negative Binomial* (ZINB), model binomial negatif digunakan untuk bagian *imperfect state.* Fungsi kepadatan peluang dari model regresi ZINB adalah sebagai berikut:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}, \text{untuk } y_i = 0 \\ (1-p_i)\frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}, \text{untuk } y_i > 0 \end{cases} \quad (1)$$

Dan mean serta variannnnya adalah

$$E(Y_i) = (1-p_i)\mu_i \;\; \& \;\; Var(Y_i) = (1-p_i)\mu_i[1+\kappa\mu_i + p_i\mu_i] \quad (2)$$

dengan $0 \leq p_i \leq 1$, $\mu_i \geq 0$, κ adalah parameter dispersi dengan κ>0 dan $\Gamma(.)$ adalah fungsi gamma. Garay, *et al.,* (2011:1305) mengemukakan bahwa model regresi ZINB dibagi menjadi dua komponen yaitu model data *count* untuk μ_i dan model *zero-inflation* untuk p_i yaitu:

$$\ln(\mu_i) = \mathbf{x_i^T\beta} \;\; \& \;\; logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{z_i^T\gamma} \quad (3)$$

dengan $i = 1, \dots, n$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ sebagai parameter model *count* dan $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^T$ sebagai parameter model *zero-inflation*. Serta estimasi parameter dengan menggunakan MLE via algoritma EM.

Pengujian parameter regresi ZINB secara simultan dengan uji G. Statistik G (uji rasio *likelihood*) adalah sebagai berikut (Agresti, 2002:12):

$$LR = -2ln\left(\frac{L_0}{L_1}\right) = -2(\mathcal{L}_0 - \mathcal{L}_1)$$

dimana $\mathcal{L}_0$ adalah fungsi l*og-likelihood* tanpa variabel bebas dan $\mathcal{L}_1$ adalah fungsi *log-likelihood* model penuh. Untuk kriteria statistik ujinya adalah H_0 ditolak, jika $LR > \chi^2_{(1-\alpha,db)}$. Kemudian Pengujian parameter regresi ZINB secara parsial menggunakan uji Wald. Statistik uji yang digunakan pada uji Wald adalah sebagai berikut (Agresti, 2002:11):

$$W = \left(\frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}\right)^2 \text{ dan } W = \left(\frac{\widehat{\gamma}_j}{SE(\widehat{\gamma}_j)}\right)^2$$

Untuk kriteria statistik ujinya adalah H_0 ditolak, jika $W > \chi^2_{(1-\alpha,1)}$.

Pemilihan model terbaik pada regresi ZINB dapat dilihat dari nilai pearson *chi-square*, *devians*, dan nilai AIC (*Akaike's Information Criterion*) (Ismail & Jemain, 2007). Nilai AIC didefinisikan sebagai berikut: $AIC = -2l + 2p$. Dengan l adalah nilai *log-likelihood* dari model dan p adalah banyaknya parameter dari model.

METODE

Data yang digunakan dalam penelitian ini adalah data sekunder dari Dinas Kesehatan Provinsi Jawa Barat tahun 2016 meliputi kasus difteri (Y), persentase imunisasi DPT (Difteri, Pertusis, Tetanus) (X1), persentase rumah tangga Berperilaku Hidup Bersih dan Sehat (ber-PHBS) (X2), persentase penduduk yang memiliki akses air minum layak (X3), persentase Tempat-Tempat Umum (TTU) memenuhi syarat kesehatan (X4), jumlah puskesmas (X5) dan persentase Tempat Pengelolaan Makanan (TPM) yang memenuhi syarat higiene sanitasi (X6). Langkah-langkah analisis data yang digunakan pada penelitian ini antara lain:

1. Estimasi parameter regresi ZINB dengan metode algoritma EM (*Expectation Maximization*).
2. Melakukan uji distribusi Poisson & uji multikolinearitas

3. Melakukan Pendugaan Parameter Regresi Poisson
4. Uji overdispersi dan pemeriksaan *excess zeros*.
5. Pendugaan parameter regresi ZINB
6. Pengujian koefisien regresi secara serentak dan juga secara parsial.
7. Pemilihan model terbaik regresi ZINB yang berdasarkan nilai AIC terkecil

HASIL DAN PEMBAHASAN

Estimasi Parameter Model Regresi ZINB

Estimasi parameter regresi *Zero Inflated Negative Binomial* (ZINB) pada penelitian ini menggunakan suatu teknik iteratif, yaitu algoritma *Expectation-Maximization* (EM). Sebelum itu dapat diketahui bahwa *Maximum Likehood Estimation* (MLE) dari pdf ZINB pada Persamaan 1 adalah sebagai berikut:

$$\log L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^{n} log\left[p_i + (1-p_i)(\frac{1}{1+\kappa\mu_i})^{\frac{1}{\kappa}}\right], y_i = 0 \\ \sum_{i=1}^{n} log\left[(1-p_i)\frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\right], y_i > 0 \end{cases}$$

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} log\left[p_i + (1-p_i)\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}\right] I(y_i = 0) +$$

$$\sum_{i=1}^{n} log\left[(1-p_i)\frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\right] I(y_i > 0)$$

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[l_1\left(\frac{1}{\kappa}, \mathbf{x_i^T}, \boldsymbol{\beta}, \mathbf{z_i^T}, \boldsymbol{\gamma}\right) I(y_i = 0) + l_2\left(\frac{1}{\kappa}, \mathbf{x_i^T}, \boldsymbol{\beta}, \mathbf{z_i^T}, \boldsymbol{\gamma}\right) I(y_i > 0)\right] \quad (4)$$

MLE dari parameter vektor tak diketahui $\widehat{\boldsymbol{\theta}}$ dapat dihitung dengan memaksimumkan *log-likelihood* pada persamaan 4. Namun penjumlahan fungsi *log-likelihood* tidak dapat diselesaikan dengan metode numerik biasa, karena fungsi *log-likelihood*nya tidak linier dan hasil pemaksimuman secara langsung akan tidak konvergen jika tidak menggunakan nilai awal yang bagus (Garay, *et al.,* 2011:1306). Oleh karena itu dapat digunakan algoritma EM.yang mana stabil dan konvergen. Sebelum masuk ke algoritma EM, terlebih dahulu harus menentukan *complete-data log-likelihood.*

Hall (2000) memisalkan variabel indikator **w**, yaitu:

$$w_i = \begin{cases} 1, & \text{jika } y_i \text{ dari keadaan nol} \\ 0, & \text{jika } y_i \text{ dari keadaan binomial negatif} \end{cases}$$

Dengan $P(w_i = 1) = p_i$ dan $P(w_i = 0) = 1 - p_i$. Kemudian membentuk distribusi gabungan $y_i$dan w_i sebagai berikut:

$$P(\mathbf{Y_c}|\boldsymbol{\theta}) = (p_i)^{w_i}(1-p_i)^{1-w_i}\left[\frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\right]^{1-w_i} \quad (5)$$

Dari persamaan 5 maka fungsi *log-likelihood*-nya didapati:

$$\ell_c(\boldsymbol{\theta}|\mathbf{Y_c}) = \sum_{i=1}^{n}\left[w_i z_i \boldsymbol{\gamma} - \log[1 + e^{z_i\boldsymbol{\gamma}}] + (1-w_i) log\left[g(y_i|\boldsymbol{\beta}, \frac{1}{\kappa})\right]\right] \quad (6)$$

dimana $\mathbf{Y_c} = (\mathbf{y}, \mathbf{w})$, $\mu_i = e^{x_i^T\boldsymbol{\beta}}$ dan $g\left(y_i\middle|\boldsymbol{\beta}, \frac{1}{\kappa}\right) = \frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}$.

Persamaan 6 ini disebut *complete-data log-likelihood,* yang mana persamaan ini dimaksimumkan dengan algoritma EM. Langkah-langkah algoritma EM adalah sebagai berikut:

1. Tahap Ekspetaksi

Ganti variabel **w** dengan $\widehat{w_i}^{(k)}(k = 0,1,2 \ldots)$ yang merupakan ekspetasi bersyarat dari w_i dengan diberikan $\mathbf{y}, \widehat{\boldsymbol{\beta}}^{(\mathbf{k})}, \widehat{\boldsymbol{\gamma}}^{(\mathbf{k})}$. Ekspetasi bersyarat ini didefinisikan sebagai berikut:

$$\widehat{w_i}^{(k)} = E(w_i|\mathbf{y_i}, \widehat{\boldsymbol{\beta}}^{(\mathbf{k})}, \hat{\boldsymbol{\gamma}}^{(\mathbf{k})}) = \begin{cases} \left[1 + e^{-\mathbf{z_i}\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}} \left(\frac{1}{1+\hat{\kappa}^{(k)} e^{-\mathbf{x_i}\hat{\boldsymbol{\beta}}^{(\mathbf{k})}}}\right)^{\frac{1}{\hat{\kappa}^{(k)}}}\right]^{-1} & \text{jika } y_i = 0 \\ 0, \text{jika } y_i > 0 \end{cases} \quad (7)$$

Sehingga *complete-data log-likelihood* (persamaan 6) yang mana **w** telah diganti dengan $\widehat{w_i}^{(k)}$ adalah sebagai berikut

$$Q\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right) = \sum_{i=1}^{n} Q_1(\boldsymbol{\gamma}|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}) + \sum_{i=1}^{n} Q_2(\boldsymbol{\beta}, \kappa^{-1}|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})})$$

dimana

$$Q_1(\boldsymbol{\gamma}|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}) = \widehat{w_i}^{(k)} \mathbf{z_i}\boldsymbol{\gamma} - \log[1 + e^{\mathbf{z_i}\boldsymbol{\gamma}}] \quad (8)$$

$$Q_2(\boldsymbol{\beta}, \kappa^{-1}|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}) = (1 - \widehat{w_i}^{(k)}) log \left[\frac{\Gamma\left(y_i + \frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right) y_i!} \left(\frac{1}{1+\kappa e^{\mathbf{x_i^T}\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}} \left(\frac{\kappa e^{\mathbf{x_i^T}\boldsymbol{\beta}}}{1+\kappa e^{\mathbf{x_i^T}\boldsymbol{\beta}}}\right)^{y_i}\right] \quad (9)$$

2. Tahap Maksimalisasi

Memaksimalkan $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ pada persamaan 8 dan 9 dengan menghitung $\widehat{\boldsymbol{\beta}}^{(\mathbf{k+1})}$ dan $\hat{\boldsymbol{\gamma}}^{(\mathbf{k+1})}$ menggunakan metode Newton-Raphson. Misalkan $\widehat{\boldsymbol{\beta}}^{(\mathbf{k})}$ dan $\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}$ adalah aproksimasi parameter pada iterasi ke-k untuk $\widehat{\boldsymbol{\beta}}$ dan $\hat{\boldsymbol{\gamma}}$. Dengan menggunakan metode Newton-Raphson maka:

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{k+1})} = \widehat{\boldsymbol{\beta}}^{(\mathbf{k})} - \left(H\left(\widehat{\boldsymbol{\beta}}^{(\mathbf{k})}\right)\right)^{-1} U(\widehat{\boldsymbol{\beta}}^{(\mathbf{k})})$$

$$\hat{\boldsymbol{\gamma}}^{(\mathbf{k+1})} = \hat{\boldsymbol{\gamma}}^{(\mathbf{k})} - \left(H\left(\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}\right)\right)^{-1} U(\hat{\boldsymbol{\gamma}}^{(\mathbf{k})})$$

3. Ganti $\widehat{\boldsymbol{\beta}}^{(\mathbf{k})}$ dan $\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}$ dengan $\widehat{\boldsymbol{\beta}}^{(\mathbf{k+1})}$ dan $\hat{\boldsymbol{\gamma}}^{(\mathbf{k+1})}$ pada iterasi selanjutnya, kemudian kembali lakukan tahap ekspetasi.
4. Lakukan iterasi ini sampai diperoleh penaksiran parameter yang konvergen ($\left|\widehat{\boldsymbol{\beta}}^{(\mathbf{k+1})} - \widehat{\boldsymbol{\beta}}^{(\mathbf{k})}\right|$ dan $\left|\hat{\boldsymbol{\gamma}}^{(\mathbf{k+1})} - \hat{\boldsymbol{\gamma}}^{(\mathbf{k})}\right|$ cukup kecil).

Penerapan Data

Hasil estimasi parameter regresi ZINB dengan algoritma EM akan diterapkan pada kasus difteri di Jawa Barat tahun 2016. Sebelum dilakukan analisis, langkah awal yang dilakukan adalah pengujian distribusi untuk variabel kasus difteri (Y) menggunakan uji *Kolmogorov-Smirnov*. Dengan bantuan *software* Easyfit, hasilnya adalah nilai $P_{value} = 0.15822 > 0.05$ dan $D_{hitung} = 0.3389 < 0.40925 = d_{(0.05;10)}$. Sehingga kesimpulannya terima H_0 yang berarti variabel jumlah kasus difteri (Y) berdistribusi Poisson. Dari hasil pengujian ini juga menunjukkan bahwa variabel terikat Y berdistribusi binomial negatif dengan $P_{value} = 0.24725 > 0.05$ dan $D_{hitung} = 0.30693 < 0.40925 = d_{(0.05,10)}$. Karena variabel terikat Y mengikuti distribusi Poisson, maka telah memenuhi asumsi awal untuk pemodelan regresi Poisson.

Pada pemodelan regresi Poisson, variabel-variabel bebas yang digunakan harus memenuhi asumsi tidak terjadi multikolinearitas. Salah satu cara mendeteksi multikoliearitas dengan melihat korelasi antara variabel bebas. Untuk melihat lebih jelas hubungan antar variabel bebas dapat dilihat di Tabel 1:

Tabel 1. Analisis Korelasi

	Y	X1	X2	X3	X4	X5
X1	-0.086					
X2	-0.392	0.01				
X3	-0.057	-0.129	0.186			
X4	-0.081	0.074	0.154	0.216		
X5	0.364	-0.129	-0.357	-0.003	0.157	
X6	0.017	-0.032	0.181	0.283	0.465	0.107

Berdasarkan Tabel 1 dapat diketahui bahwa hubungan antar variabel bebas tidak ada yang memiliki hubungan yang kuat antar variabelnya sehingga dapat disimpulkan tidak ada kasus multikolinearitas.

Setelah diketahui hubungan antar variabel bebas tidak ada kasus multikolinearitas dan distribusi dari variabel kasus difteri (Y) adalah Poisson, maka selanjutnya adalah pemodelan regresi Poisson antara variabel-variabel bebas terhadap variabel kasus difteri. Hasil estimasi parameter model regresi Poisson dapat dilihat pada Tabel 2 sebagai berikut:

Tabel 2. Hasil Estimasi Parameter Model Regresi Poisson

Parameter	Koefisien	SE Koefisien	$\text{Wald}_{\text{hitung}}$	P_{value}
β_0	5.689533	1.934106	8.655364	0.003264
β_1	-0.010525	0.020991	0.251001	0.616096
β_2	-0.074685	0.019296	14.98464	0.000109
β_3	-0.001469	0.008961	0.026896	0.869738
β_4	-0.022434	0.012490	3.225616	0.072482
β_5	0.001441	0.001028	1.965604	0.160920
β_6	0.012951	0.009705	1.779556	0.182063
Devians: 86.967			db: 20	
Statistik uji G: 33.13972			AIC : 130.96	

Berdasarkan Tabel 2, dapat diketahui bahwa uji serentak didapatkan $\text{G} = 33.13972 > 31.41 = \chi^2_{(0.05;20)}$, sehingga keputusan dari pengujian parameter secara serentak menyatakan tolak H_0, yang artinya bahwa ada salah satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon. Selanjutnya untuk uji parameter secara individual, variabel yang berpengaruh secara signifikan hanya persentase rumah tangga berperilaku hidup bersih dan sehat (X2) dengan $wald_{\text{hitung}} = 14.98464 > 3.814 = \chi^2_{(0.05;1)}$ dan $\text{P}_{\text{value}} = 0.000109 < 0.05$.

Salah satu asumsi dari regresi Poisson adalah tidak terjadi kasus *overdispersion.* Pengujian *overdispersion* pada regresi Poisson dapat dilakukan dengan melihat hasil bagi *devians* dengan derajat bebas. Karena hasil nilai $\frac{Devians}{db} = \frac{86.967}{20} > 1$, maka dapat disimpulkan bahwa model regresi Poisson terjadi *overdispersion*. Selanjutnya dilihat apakah terdapat tambahan kasus *excess zeros* atau tidak. Hasil pemeriksaan *excess zeros* pada variabel jumlah kasus difteri (Y) didapati persentase amatan yang bernilai nol lebih dari 50% yaitu sebesar 62.962% dari total data. Sehingga disimpulkan data terdapat *excess zeros* Maka dari itu, model regresi yang diharapkan dapat mengatasi data variabel terikat Y yang berdistribusi binomial negatif dan *excess zero* serta dilanjutkan terdapat *overdispersion* di pemodelan regresi Poisson adalah dengan menggunakan regresi *Zero-Inflated Negative Binomial* (ZINB).

Model regresi *Zero-Inflated Negative Binomial* (ZINB) bertujuan untuk memperbaiki data yang terdapat *overdispersion* dan *excess zero* pada variabel terikat. Hasil estimasi parameter regresi ZINB, nilai uji statistik G dan uji Wald serta AIC adalah sebagai berikut:

Tabel 3. Estimasi Parameter Model *Zero Inflated Negative Binomial*

Variabel	Nilai Estimasi Parameter	Standar Error	$\text{Wald}_{\text{hitung}}$	p-value
	Model *Count*			
Intercept	8.694467	2.507285	12.02702	0.000525
X_1	-0.052863	0.024218	4.765489	0.029050
X_2	-0.035799	0.028356	1.592644	0.206776
X_3	-0.054446	0.013085	17.31392	3.17e-05
X_4	0.006576	0.041952	0.024649	0.87543

Variabel	Nilai Estimasi Parameter	Standar Error	$Wald_{hitung}$	p-value
X_5	-0.001189	0.001918	0.3844	0.535279
X_6	0.038918	0.018185	4.5796	0.032344
Log (*theta*)	9.741395	29.538004	0.1089	0.741557
Model *Zero Inflation*				
Intercept	3.63230	8.79405	0.170569	0.680
X_1	-0.05874	0.09427	0.388129	0.533
X_2	0.11313	0.12574	0.81000	0.368
X_3	-0.17281	0.07655	5.094049	0.024
X_4	0.10513	0.07396	2.019241	0.155
X_5	-0.07093	0.06203	1.308736	0.253
X_6	0.07708	0.05115	2.271049	0.132
Loglikelihood $= -23.15$	Statistik uji G=36.54		AIC=76.3	

Berdasarkan Tabel 3, pengujian signifikansi parameter secara silmultan model regresi ZINB dengan G didapati nilai 36.54. Karena nilai $G = 36.54 > 24.995 = \chi^2_{(0.05;15)}$, yang berarti terdapat minimal satu variabel bebas yang berpengaruh secara signifikan terhadap variabel terikat. Sedangkan untuk uji signifikansi parameter secara individual dengan uji Wald hanya variabel persentase imunisasi DPT (X1), persentase penduduk yang memiliki akses air minum layak (X3) dan persentase TPM (Tempat Pengelolaan Makanan) memenuhi syarat higiene sanitasi (X6) yang berpengaruh secara signifikan terhadap variabel jumlah kasus difteri (Y) dengan $wald_{\text{hitung}} > 3.814 = \chi^2_{(0.05;1)}$ dan $p - value < 0.05$. Kemudian untuk model *zero inflation,* hanya variabel persentase penduduk yang memiliki akses air minum layak (X3) yang berpengaruh secara signifikan dengan $wald_{\text{hitung}} = 5.094049 > 3.814 = \chi^2_{(0.05;1)}$ dan $p - value < 0.05$.

Tahap selanjutnya adalah pemilihan model terbaik pada regresi ZINB dengan melihat nilai AIC (*Akaike's Information Criterion*). Untuk mendapatkan AIC yang baik, model akan direduksi menggunakan *backward elimination* dengan mengeluarkan variabel yang tidak signifikan. Hasil estimasi model ZINB dengan AIC terkecil setelah direduksi sebagai berikut:

Tabel 4. Estimasi Parameter Model *ZINB* dengan AIC Terkecil

Variabel	Nilai Estimasi Parameter	Standar Error	$Wald_{hitung}$	*p-value*
Model *Count*				
Intercept	7.25717	1.47235	24.29504	8.27e-07
X_1	-0.05178	0.01601	10.45876	0.001221
X_3	-0.05633	0.01270	19.68697	9.13e-06
X_6	0.04086	0.01130	13.06823	0.000301
Log (*theta*)	10.27636	56.73196	0.032761	0.856259
Model *Zero Inflation*				
Intercept	15.35174	7.61110	4.068289	0.0437
X_3	-0.20186	0.09418	4.592449	0.0321
X_5	-0.15452	0.07544	4.194304	0.0405
X_6	0.11379	0.05222	4.748041	0.0293
Loglikelihood $= -25.96$	Statistik uji $\boldsymbol{G}$ =30.92		AIC = 69.92	

Pengujian signifikansi parameter secara simultan regresi ZINB dilakukan dengan uji G. Dari Tabel 4, Nilai G yang didapat adalah 30.92, karena nilai $G = 30.92 > 16.919 =$

$\chi^2_{(0.05;9)}$ maka hasil uji menyatakan tolak H_0. Yang berarti terdapat minimal satu variabel bebas yang berpengaruh secara signifikan terhadap variabel terikat. Dari Tabel 4 terlihat bahwa semua variabel bebas untuk model *count* maupun model *zero-inflation* berpengaruh secara signifikan dengan $wald_{\text{hitung}} > 3.814 = \chi^2_{(0.05;1)}$ dan $p - value < 0.05$. Untuk melihat kasus overdispersi telah teratasi dengan menggunakan pearson statistik. Nilai pearson statistik yang diperoleh adalah 19.48139, yang mana $\frac{P}{n-k} = \frac{19.48139}{19} = 1.0253 \approx 1$ sehingga dapat disimpulkan kasus overdispersi telah teratasi.

Dengan demikian model ZINB terbaik berdasarkan AIC yang terbentuk adalah sebagai berikut:

Model *count* (μ_i) untuk data aktual $y_i > 0$ yaitu

$$\mu_i = \exp(7.25717 - 0.05178\,\text{X1} - 0.05633\,\text{X3} + 0.04086\,\text{X6}) \quad (10)$$

Dan model *zero-inflation* (p_i) untuk data aktual $y_i = 0$ yaitu

$$p_i = \frac{\exp(15.35174 - 0.20186\,X3 - 0.15452\,X5 + 0.11379\,X6)}{1+\exp(15.35174 - 0.20186\,X3 - 0.15452\,X5 + 0.11379X6)} \quad (11)$$

Interpretasi model yang terbentuk dari ZINB didasarkan pada nilai dari nilai exp(**β**). Dan berikut adalah interpretasi model terbaik regresi ZINB:

a. Untuk model *count* μ_i, yaitu:

Setiap penambahan 1% imunisasi DPT (X1) maka akan menurunkan rata-rata ditemukannya kasus difteri sebesar 0.9495 kali dari rata-rata kasus difteri semula jika faktor lain dianggap konstan. Setiap penambahan 1% penduduk yang memiliki akses air minum (X3) maka akan menurunkan rata-rata ditemukannya kasus difteri sebesar 0.9452 kali dari rata-rata kasus difteri semula jika faktor lain dianggap konstan. Dan setiap penambahan 1% TPM (Tempat Pengelolaan Makanan) memenuhi syarat higiene sanitasi (X6) maka akan meningkatkan rata-rata ditemukannya kasus difteri sebesar 1.0417 kali dari rata-rata kasus difteri semula jika faktor lain dianggap konstan.

b. Untuk model data *zero-infllation* p_i, yaitu:

Setiap penambahan 1% penduduk yang memiliki akses air minum (X3) maka akan menurunkan peluang ditemukannya kasus difteri sebesar 0.8172 kali dari rata-rata kasus difteri semula jika faktor lain dianggap konstan. Setiap penambahan 1unit puskesmas (X5) maka akan menurunkan peluang ditemukannya kasus difteri sebesar 0.8568 kali dari rata-rata kasus difteri semula jika faktor lain dianggap konstan. Setiap penambahan 1% TPM (Tempat Pengelolaan Makanan) memenuhi syarat higiene sanitasi (X6) maka akan meningkatkan peluang ditemukannya kasus difteri sebesar 1,1205 kali dari rata-rata kasus difteri semula jika faktor lain dianggap konstan.

PENUTUP

Berdasarkan hasil analisis data dan pembahasan yang telah dilakukan maka kesimpulan yang dapat diambil adalah:

1. Estimasi parameter model ZINB dengan metode MLE via algoritma EM didapatkan hasil sebagai berikut:

   a. Tahap Ekspektasi

   Hasil *complete-data log-likelihood* yang mana **w** telah diganti dengan $\widehat{w_i}^{(k)}$ sebagai berikut:

$$\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right) = \sum_{i=1}^{n} \widehat{w_i}^{(k)} \mathbf{z_i}\boldsymbol{\gamma} - \log[1 + e^{\mathbf{z_i}\boldsymbol{\gamma}}] \; + \sum_{i=1}^{n} (1 - \widehat{w_i}^{(k)}) log \left[\frac{\Gamma\left(y_i + \frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right) y_i!} \left(\frac{1}{1+\kappa e^{\mathbf{x_i}^{\mathbf{T}}\boldsymbol{\beta}}} \right)^{\frac{1}{\kappa}} \left(\frac{\kappa e^{\mathbf{x_i}^{\mathbf{T}}\boldsymbol{\beta}}}{1+\kappa e^{\mathbf{x_i}^{\mathbf{T}}\boldsymbol{\beta}}} \right)^{y_i} \right]$$

dimana $\widehat{w_i}^{(k)} = E(w_i|\mathbf{y_i}, \widehat{\boldsymbol{\beta}}^{(\mathbf{k})}, \hat{\boldsymbol{\gamma}}^{(\mathbf{k})}) = \begin{cases} \left[1 + e^{-\mathbf{z_i}\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}} \left(\frac{1}{1+\hat{\kappa}^{(k)} e^{-\mathbf{x_i}\widehat{\boldsymbol{\beta}}^{(\mathbf{k})}}}\right)^{\frac{1}{\hat{\kappa}^{(k)}}}\right]^{-1} & \text{jika } y_i = 0 \\ 0, \text{jika } y_i > 0 \end{cases}$

b. Tahap Maksimisasi

Memaksimalkan $Q\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right)$ dengan metode Newton-Raphson didapatkan:

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{k+1})} = \widehat{\boldsymbol{\beta}}^{(\mathbf{k})} - \left(H\left(\widehat{\boldsymbol{\beta}}^{(\mathbf{k})}\right)\right)^{-1} U(\widehat{\boldsymbol{\beta}}^{(\mathbf{k})}) = \widehat{\boldsymbol{\beta}}^{(\mathbf{k})} - \left(\frac{d^2 Q\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right)}{d\boldsymbol{\beta}^{\mathbf{T}} d\boldsymbol{\beta}}\right)^{-1} \frac{d\, Q\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right)}{d\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\gamma}}^{(\mathbf{k+1})} = \hat{\boldsymbol{\gamma}}^{(\mathbf{k})} - \left(H\left(\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}\right)\right)^{-1} U(\hat{\boldsymbol{\gamma}}^{(\mathbf{k})}) = \hat{\boldsymbol{\gamma}}^{(\mathbf{k})} - \left(\frac{d^2 Q\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right)}{d\boldsymbol{\gamma}^{\mathbf{T}} d\boldsymbol{\gamma}}\right)^{-1} \frac{d\, Q\left(\boldsymbol{\theta}\middle|\widehat{\boldsymbol{\theta}}^{(\mathbf{k})}\right)}{d\boldsymbol{\gamma}}$$

Yang kemudian lakukan iterasi sampai diperoleh penaksiran parameter yang konvergen ($\left|\widehat{\boldsymbol{\beta}}^{(\mathbf{k+1})} - \widehat{\boldsymbol{\beta}}^{(\mathbf{k})}\right|$ dan $\left|\hat{\boldsymbol{\gamma}}^{(\mathbf{k+1})} - \hat{\boldsymbol{\gamma}}^{(\mathbf{k})}\right|$ cukup kecil).

2. Model terbaik regresi *Zero-Inflated Negative Binomial* (ZINB) yang dapat menjelaskan hubungan variabel terikat dengan variabel bebas pada jumlah kasus difteri di Provinsi Jawa Barat tahun 2016 adalah sebagai berikut:

   Untuk model data diskrit:

   $$\mu_i = \exp(7.25717 - 0.05178\,\text{X1} - 0.05633\,\text{X3} + 0.04086\,\text{X6})$$

   Dan untuk model *zero-inflation:*

   $$p_i = \frac{\exp(15.35174 - 0.20186\,X3 - 0.15452\,X5 + 0.11379X6)}{1+\exp(15.35174 - 0.20186\,X3 - 0.15452\,X5 + 0.11379X6)}$$

   Dengan faktor-faktor yang mempunyai pengaruh secara signifikan terhadap jumlah kasus difteri di Provinsi Jawa Barat tahun 2016 untuk data *count* adalah persentase imunisasi DPT (X1), persentase penduduk yang memiliki akses air minum layak (X3) dan persentase TPM yang memenuhi syarat higiene sanitasi (X6). Sedangkan untuk data *zero-inflation* adalah persentase penduduk yang memiliki akses air minum layak (X3), jumlah puskesmas (X5) dan persentase TPM yang memenuhi syarat higiene sanitasi (X6).

Saran

Pada penelitian ini penulis hanya mendapati 3 variabel bebas yang berpengaruh secara signifikan terhadap kasus difteri di Jawa Barat tahun 2016 pada model *count* maupun *zero-inflation*. Oleh karena itu, untuk penelitian selanjutnya disarankan menambahkan variabel bebas supaya didapati hasil yang lebih baik.

DAFTAR RUJUKAN

Agresti, A. 2002. *Categorical Data Analysis Second Edition*. 2nd edn. Canada: A John Wiley & Sons Inc.

Faidah, D. Y. & Pontoh, R. S. 2015. Pendekatan Hurdle Poisson Pada Excess Zero Data. *Seminar Nasinonal Matematika dan Pendidikan Matematika*, pp. 131–136.

Garay, A. M., Hashimoto, E.M., Ortega, E.M., & Lachos, V.H. 2011. On Estimation and Influence Diagnostics for Zero-Inflated Negative Binomial Regression Models. *Computational Statistics and Data Analysis*, 55, pp. 1304–1318. doi: 10.1016/j.csda.2010.09.019.

Hall, D. B. 2000. Zero-Inflated Poisson and Binomial Regression with Random Effects : A Case Study. *Biometrics*, 56, pp. 1030–1039.

Hilbe, J. M. 2011. *Negative Binomial Regression Second Edition*. 2nd edn. New York: Cambridge University Press.

Ismail, N. & Jemain, A. A. 2007. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. *Casual Actuarial Society Forum*, pp. 103–1581.

Ismail, N. & 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*, pp. 1–28.

Lambert, D. 1992. Zero-Inflated Poisson Regression, with an Applicationto Defects in Manufacturing. *Technometrics*, 34(1), pp. 1–14.

Weng, J., Ge, Y. E., & Han, H. 2016. Evaluation of Shipping Accident Casualties using Zero-inflated Negative Binomial Regression Technique. *The Journal of Navigation*, 69(2), pp. 433–448. doi: 10.1017/S037346331500078