ISSN : 2460 – 7797
e-ISSN :2614-8234
Website : jurnal.umj.ac.id/index.php/fbc
Email : fibonacci@umj.ac.id

FIBONACCI
JurnalPendidikanMatematikadanMatematika

PENANGANAN OVERDISPERSI PADA PEMODELAN DATA CACAH DENGAN RESPON NOL BERLEBIH (*ZERO-INFLATED*)

Viarti Eminita[1)*], Anang Kurnia[2)], Kusman Sadik[3)]
[1)]Pendidikan Matematika, Fakultas Ilmu Pendidikan, Universitas Muhammadiyah Jakarta, Jln. KH Ahmad Dahlan, 15419
[2,3)]Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, IPB University Bogor, 16680

[*]*phiartea.emn@gmail.com*

Abstrak

Overdispersi pada data cacah yang disebabkan karena kasus nol berlebih tidak dapat ditangani dengan metode model linier umum biasa seperti Poisson dan Binomial Negatif. Penanganan overdispersi karena nol berlebih dapat dilakukan dengan menggunakan model Zero-Inflated. Zero-Inflated Poisson (ZIP) dan Zero-Inflated Binomial Negatif (ZIBN) telah diyakini performanya dalam menangani masalah ini. Selain menangani masalah tersebut kedua model ini juga dapat memberikan informasi mengenai penyebab nol berlebih pada data respon. Performa ke Empat model tersebut dibandingkan dalam menduga model dari jumlah anak yang tidak sekolah dalam keluarga di Provinsi Jawa Barat pada tahun 2017. Berdasarkan nilai dari ukuran Pearson Chi-Squares, Likelihood Ratio Chi-Square, dan Akaike Information Crieteria (AIC). Pearson Chi-Squares, model ZIP lebih baik dibandingkan ZIBN dan model lainnya, walaupun berbeda sedikit dengan ZIBN.

Kata Kunci: *Overdispersi, Zero-Inflated Poisson, Zero-Inflated Negative Binomial*

PENDAHULUAN

Data cacah biasanya memiliki karakteristik bersebaran Poisson yang dimodelkan dengan model standar, dengan asumsi varians respon diharapkan sama dengan rata-rata. Tetapi McCullagh dan Nelder (1989) menunjukkan bahwa overdispersi tidak jarang terjadi dalam prakteknya. Overdispersi harus dipertimbangkan dengan hati-hati dalam memodelkan data respon cacah. Model linear umum Poisson biasa (GLM) yang dikembangkan oleh Palmgren (1981) tidak dapat digunakan dengan baik jika terjadi overdispersi. Overdispersi pada data biasanya disebabkan karena efek cluster (Nelder & Weddern (1972), McCullagh dan Nelder (1989)).

Metode GLM yang dapat menangani overdispersi adalah model Quasi-Poisson dan model Binomial Negatif (BN). Hausman *et. al.* (1984) mengklaim bahwa model

Binomial negatif lebih baik dari model Poisson ketika ada overdispersi. Seiring dengan berkembangnya permasalahan data, metode yang ada sebelumnya tidak bisa lagi menangani overdispersi karena nilai nol yang berlebih pada data, sehingga Lambert (1992) mengembangkan metode *Zero-Inflated* yang memperhatikan nol berlebih dan menganggap bahwa nilai nol pada data sangat bermanfaat dan dapat memberikan informasi yang lebih mengenai data. Model *Zero-Inflated* Poisson baik dalam menangani overdispersi karena nol berlebih, namun kurang baik jika penyebab lainnya (Jeong, 2018). Jiang *et. al* (2017) *Zero-Inflated* Binomial Negatif (ZIBN) juga baik dalam menangani overdispersi karena nol berlebih dan ukuran contoh yang semakin besar. Zeileis *et. al.* (2008) mengimplementasikan Zero inflated Regression models in R program yang membandingkan zero-inflated models dengan beberapa distribusi untuk menangani overdispersi karena nol berlebih pada data cacah.

Pada paper ini dikaji karakteristik pendugaan model pada data cacah yang diidentifikasi memiliki nilai nol yang berlebih sehingga menyebabkan terjadinya overdispersi. Keempat model yaitu model Poisson, model BN, model ZIP, dan model ZIBN dibandingkan dalam menduga model dari jumlah anak yang tidak sekolah dalam keluarga di Provinsi Jawa Barat pada tahun 2017.

Generalized Linear Model (GLM)

GLM merupakan pengembangan dari model linier yang mensyaratkan terpenuhinya asumsi galat yang menyebar normal. Asumsi tersebut dapat dilonggarkan ke sebaran keluarga eksponensial yang dijadakan dasar dalam pendugaan kemungkinan maksimum (Nelder dan Wedderburn, 1972). GLM juga mengakomodir semua peubah respon dan penjelas yang diukur dengan skala nominal, ordinal, dan kontinu (Dobson, 2002). GLM memiliki 3 komponen utama yang menyusun model, yaitu komponen acak (E[$\mathbf{Y}$] = $\boldsymbol{\mu}$), komponen sistematik ($\boldsymbol{\eta}$), dan funsi penghubung $g(\cdot)$ yang menghubungkan komponen acak dengan komponen sistematik ($\boldsymbol{\eta} = g(\boldsymbol{\mu})$) (McCullagh dan Nelder, 1989).

Fungsi kemungkinan untuk GLM yang mengasumsikan bahwa Y_i mempunyai sebaran dari keluarga eksponensial dengan fungsi kepadatan peluang yang dapat dinyatakan dengan persamaan (1) berikut

$$f_i(y_i;\lambda,\phi) = \exp\left(\frac{y_i \cdot \lambda_i - b(\lambda_i)}{a_i(\phi)} + c(y_i,\phi)\right) \quad (1)$$

Dimana $a_i(.)$, $b_i(.)$, dan $c_i(.)$ merupakan suatu fungsi dan λ_i adalah parameter kanonik dari keluarga eksponensial dengan ϕ diketahui. Nilai tengah dan ragam dari Y_i adalah E[Y_i] = μ_i = $b'(\lambda_i)$ dan var[Y_i] = $b''(\lambda_i)$ $a_i(\phi)$. Dari persamaan 2.1 diperoleh fungsi log kemungkinan dari Y_i, yaitu

$$l(\lambda_i;y_i,\phi) = \sum_{i=1}^{n} \log f_i(y_i;\lambda_i,\phi) = \sum_{i=1}^{n} \frac{y_i \cdot \lambda_i - b(\lambda_i)}{a_i(\phi)} + c_i(y_i,\phi) \quad (2)$$

Model Poisson

Data cacah biasanya merupakan peubah diskrit Y yang mempunyai distribusi dengan fungsi massa peluang hanya pada nilai integer non-negatif saja, yaitu distribusi Poisson (Ismail dan Jemain, 2007). Misalkan Y_i merupakan peubah acak yang berdistribusi Poisson dengan fungsi kepadatan peluang

FIBONACCI : Jurnal Pendidikan Matematika dan Matematika. Vol. 5 (1), pp: 71 - 80.

$$f_i(y_i;\lambda) = \frac{{\lambda_i}^{y_i}\exp(-\lambda_i)}{y_i!}, y_i = 0, 1, \ldots \quad (3)$$

dengan nilai tengah dan ragam, $E(Y_i) = Var(Y_i) = \lambda_i$.

Regresi Poisson menghubungkan peubah respon *Y* dengan kovariat mempunyai fungsi penghubung kanonik $g(\mu_i) = \log(\mu_i)$, sehingga nilai tengah diasumsikan mempunyai sifat multiplikatif, yaitu $E(Y_i|\mathbf{x}_i) = \lambda_i = e_i \exp(\mathbf{x}_i{}^{\mathrm{T}}\beta)$, dengan e_i merupakan ukuran eksposur, $\mathbf{x}_i$ merupakan vektor kovariat *p*x1 dan **β** parameter regresi *p*x1. Persamaan skor kemungkinan maksimum untuk menduga parameter **β** adalah

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \ell(\beta_j)} = \sum_j (y_i - \lambda_i)x_{ij} = 0, j = 1, 2, \ldots, p \quad (4)$$

Selanjutnya **β** akan diduga menggunakan persamaan di atas dengan regresi Kuadrat terkecil (*maximum likelihood* (ML)) menggunakan algoritma kuadrat terkecil terboboti iteratif (*iterative weighted least squares* (IWLS)).

Sebaran Poisson memiliki sifat yaitu nilai tengah dan ragamnya memiliki nilai yang sama (E[*Y*] = Var[*Y*]). Jika nilai ragam dari *Y* melebihi dari nilai harapannya, maka kondisi ini sering disebut overdispersi. Keragamana data pada *Y* biasanya ditunjukkan dengan rasio dispersi (τ), yaitu ukuran penyebaran data terhadap nilai tengahnya sedemikian sehingga $E(Y) = \tau\, Var(Y)$. Jika nilainya kecil, maka data memiliki ragam yang homogen, jika sebaliknya maka data memiliki ragam yang heterogen. Jika $\tau > 1$, maka data cacah diidentifikasi mengalami overdispersi.

Model Binomial Negatif

Salah satu pemodelan yang dilakukan untuk mengatasi overdispersi pada data cacah adalah dengan mengasumsikan bahwa data bersebaran Binomial Negatif. Hal ini dikarenakan parameter dispersi pada model ini diasumsikan bernilai 1 ($\tau = 1$). Misalkan peubah acak *Y* bersebaran Poisson($\nu_i\lambda_i(\mathbf{x}_i,\boldsymbol{\beta})$) dengan υ diasumsikan bersebaran Gamma dengan $E[\upsilon_i] = 1$ dan $Var[\upsilon_i] = \theta$, sehingga *Y* memiliki sebaran Binomial Negatif dengan fungsi kepekatan peluangnya, yaitu:

$$f(y;\lambda,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!} \cdot \left(\frac{\theta}{\theta+\lambda}\right)^{\theta}\left(1-\frac{\theta}{\theta+\lambda}\right)^{y_i}$$

dengan $E(Y) = \lambda$, dan $var(Y) = \lambda + \lambda^2/\theta$ dengan θ merupakan parameter *shape* sebaran Gamma dan $\Gamma(\cdot)$ adalah fungsi gamma dan $\frac{1}{\theta}$ merupakan parameter dispersi. Fungsi kemungkinan maksimum bagi λ adalah

$$\ell(\lambda,\theta;y) = \sum_{i=1}^{n}\left\{y_i \ln\lambda_i + \theta\ln\theta - (\theta+y_i)\ln(\theta+\lambda_i) + \ln\frac{\Gamma(\theta+y_i)}{\Gamma(\theta)} - \ln y_i!\right\}$$

θ diasumsikan bernilai tetap, sehingga sebaran BN merupakan anggota keluarga eksponensial.

Regresi menghubungkan peubah respon *Y* dengan kovariat mempunyai fungsi penghubung kanonik $g(\lambda_i) = \ln(\lambda_i) = \eta_i$, pemodelan λ_i dengan predictor linier $\eta_i = \mathbf{x}_i{}^{\mathrm{T}}\beta$, $\mathbf{x}_i$ merupakan vektor kovariat *p*x1 dan **β** parameter regresi *p*x1. Persamaan skor untuk kemungkinan maksimum pada pendugaan parameter **β** dengan θ tetap adalah:

$$\frac{\partial^2\ell}{\partial^2\beta_j} = \sum_{i=1}^{n}\frac{(y_i-\lambda_i)}{\lambda_i\left(1+\frac{\lambda_i}{\theta}\right)}\frac{1}{g'(\lambda_i)}x_{ij}$$

Pendugaan **β** biasanya dilakukan dengan metode *Iterative Reweighted Least Square* (IRLS) dengan θ tetap dan $V(\lambda) = \lambda + \lambda^2/\theta$.

Model *Zero-Inflated* Poisson

Lambert (1992) menyatakan bahwa peubah respon $Y = (Y_1, Y_2, \dots, Y_n)'$ yang saling bebas dalam regresi ZIP memiliki sebaran yaitu

$Y_i \sim 0$ dengan peluang p_i
$Y_i \sim Poisson(\lambda_i)$ dengan peluang $1 - p_i$

Dalam hal ini berarti bahwa nilai nol diasumsikan muncul dengan peluang *p* yang sering disebut *structural zeros* dan data cacah menyebar Poisson pada parameter λ dengan peluang (1-*p*) yang disebut dengan *sampling zeros* (Jansakul dan Hinde, 2002). Sehingga fungsi masa peluang Y_i menyebar ZIP adalah

$$P(Y = y_i) = \begin{cases} p + (1-p)e^{-\lambda}, y_i = 0 \\ (1-p)\dfrac{e^{-\lambda}\lambda^{y_i}}{y_i!}, y_i = 1,2,\dots, dan\ 0 \le p \le 1 \end{cases} \quad (5)$$

dengan paramaeter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)'$ dan $\mathbf{p} = (p_1, p_2, \dots, p_n)'$ dengan fungsi penghubung

$$\ln(\lambda) = \mathbf{B\beta}$$

dan

$$logit(\mathbf{p}) = \ln\left(\frac{\mathbf{p}}{1-\mathbf{p}}\right) = \mathbf{G\gamma} \quad (6)$$

Dimana **B** dan **G** merupakan matriks kovariat . Jika **B** = **G** serta λ dan p tidak memiliki hubungan fungsional, maka regresi ZIP membutuhkan parameter dua kali lipat dibandingkan regresi Poisson. Sedangkan pada kasus lainnya, yaitu jika peluang dari kondisi sempurna tidak bergantung pada kovariat, maka **G** adalah matriks 1 kolom dan regresi ZIP membutuhkan minimal satu parameter dibanding regresi Poisson.

Nilai harapan dan ragam dari *Y* sebagai berikut

$$E(Y) = (1-p)\lambda = \mu \quad (7)$$

dan

$$Var(Y) = \mu + (\frac{p}{1-p})\mu^2 \quad (8)$$

Overdispersi pada *Y* terjadi jika sebaran dari marginal *Y* nilai $p > 0$ yang mengindikasikan terjadi peningkatan nilai nol pada peubah respon *Y* dan pada Persamaan (7) dan (8) terlihat bahwa Var(Y) > E(Y) yang mengindikasikan bahwa regresi ZIP dapat mengatasi overdispersi.

Metode kemungkinan maksimum digunakan untuk menduga parameter koefisien regresi ZIP dengan fungsi log-kemungkinan

$$\begin{aligned} \ell = \ell(\lambda; p; y) = \sum_{i=1}^{n} \{ & I_{(y=0)} \ln[p \\ & + (1-p)e^{-\lambda}] \\ & + I_{(y>0)}[\ln(1-p) \\ & - \lambda \\ & + y \ln \lambda - \ln(y_i!)] \} \end{aligned} \quad (9)$$

dengan $I_{(.)}$ adalah fungsi indikator kejadian tertentu. Penduga parameter bagi β dan γ pada Persamaan (6) diperoleh dengan menggunakan algoritma *Expectation Maximization* (EM).

Model *Zero-Inflated* Binomial Negatif (ZIBN)

Fungsi masa peluang Y_i menyebar ZIBN adalah (Jiang dan House, 2017):

$$P(Y = y_i) = \begin{cases} p + (1-p)\left(\frac{\theta}{\theta+\lambda}\right)^{\theta} I_{(y=0)}, & , y_i = 0 \\ (1-p)\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\theta+\lambda}\right)^{\theta}\left(1-\frac{\theta}{\theta+\lambda}\right)^{y_i} I_{(y>0)}, & , y_i > 0 \end{cases} \quad (10)$$

dengan λ adalah nilai tengah dari sebaran Binomial Negatif dan $\frac{1}{\theta}$ adalah parameter dispersi. Peubah acak Y memiliki sifat bahwa $E[Y\} = (1-p)\mu, Var(Y) = (1-p)\mu\left(1 + \frac{\lambda}{\theta} + p\lambda\right)$. Fungsi penghubung dari model regresi binomial negatif sama dengan fungsi penghubung model regresi Poisson atau sebaran binomial negatif konvergen ke sebaran Poisson jika $\to \infty$.

Identifikasi Sebaran Y

Identifikasi terhadap sebaran Y dilakukan menggunakan uji Skor dan uji *Chi-Square*. Uji skor bertujuan untuk memeriksa berlebih atau tidaknya peluang

FIBONACCI : Jurnal Pendidikan Matematika dan Matematika. Vol. 5 (1), pp: 71 - 80.

nol pada peubah respon. Hipotesis yang akan diuji adalah

$$H_0 : \omega = 0 \text{ dan } H_1: \omega > 0 \qquad (11)$$

dengan ω adalah peluang nol pada peubah respon dan statistik ujinya adalah:

$$S_\omega = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} \qquad (12)$$

dengan n_0 adalah banyaknya nilai nol, n adalah ukuran data, $p_0 = \exp(\hat{\lambda}_0)$ dengan $\hat{\lambda}_0$ merupakan penduga parameter Poisson di bawah kondisi H_0 atau $\bar{y}$, dan $\bar{y}$ adalah nilai rataan dari peubah respon. Statistik uji S_ω pada persamaan (12) bersebaran chi-square (χ^2) dengan derajat bebas 1. Jika $S_\omega > \chi^2_{\alpha,1}$, maka tolak H_0 pada taraf nyata (α) yang berarti bahwa terjadi peluang nol berlebih pada peubah respon, yang menyebabkan overdispersi.

Uji *Chi-square* digunakan untuk memeriksa kesesuaian sekumpulan data terhadap sebaran tertentu. Dalam paper ini, uji ini digunakan untuk menguji apakah sekumpulan data cacah bersebaran Poisson dan ZIP. Hipotesis dalam uji ini adalah

$$H_0 : p = p^0 \text{ dan } H_1: p \neq p^0 \qquad (13)$$

dengan p adalah peluang amatan dan p^0 adalah peluang sebaran Poisson dan ZIP. Statistik uji Chi-square diperoleh menggunakan formula berikut:

$$\chi^2 = \sum_{l=0}^{m} \frac{(n_l - np_l)^2}{np_l} \qquad (14)$$

dengan n_l adalah frekuensi yang diamati untuk setiap kategori ke-*l*, p_l adalah fungsi massa peluang dari sebaran Poisson dan ZIP, *n* adalah ukuran contoh, dan *m* adalah jumlah kategori yang diamati. Sebaran asimtotik tatistik uji χ^2 bersebaran χ^2 dengan derajat bebas (*m-p)*, dan *p* adalah jumlah parameter diduga oleh data, dalam hal ini penduga parameternya berjumlah 1, yaitu λ. Jika $\chi^2 > \chi^2_{\alpha,(m-p-1)}$, maka H_0 ditolak pada α berarti bahwa tidak terdapat kecocokan antara peluang amatan dengan peluang sebaran Poisson atau dalam hal ini peubah respon tidak memiliki sebaran Poisson atau ZIP.

Goodeness of Fit Tests

Ukuran kebaikan model yang digunakan dalam penelitian ini adalah Pearson *Chi-Squares*, *Likelihood Ratio Chi-Square*, dan *Akaike Information Crieteria* (AIC). *Pearson chi-squares* merupakan ukuran kebaikan yang sering digunakan dalam *Generalized Linear Models* (GLM). Hipotesis pada uji ini adalah:

$$H_0 : \tau = 1 \text{ dan } H_1: \tau > 1 \qquad (15)$$

dengan statistik uji Pearson *chi-square* adalah:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{Var(Y_i)}$$

Sebaran asimtotik dari statistik uji ini menyebar *chi-squares* dengan derajat bebas *n-p*, dengan *n* adalah banyaknya amatan dan *p* jumlah parameter. Rasio dispersi (τ) untuk mengukur keragaman data terhadap regresi Poisson dan ZIP adalah

$$\tau = \frac{\chi^2}{n-k} \qquad (16)$$

LR *Chi-Square* merupakan salah satu statistik uji untuk menilai *Goodness of Fit* dalam statistika multivariat seperti regresi logistik, dan ketakbebasan dalam tabel kontingensi dan formula statistik ini yaitu (Ozdemir dan Eyduran, 2005):

$$G = 2\sum_{i=1}^{n} f \cdot \ln\frac{f}{f_i}$$

dengan f adalah frekuensi amatan dan f_i frekuensi harapan. Model terbaik adalah model dengan LR *Chi-square* yang kecil. Ukuran ketiga adalah AIC yang merupakan salah satu metode yang dapat memberikan performa dari model kemungkinan maksimum dapat digunakan menyesuaikan data. AIC didefinisikan sebagai berikut:

$$\text{AIC} = -2\ell + 2p$$

Dengan ℓ menyatakan log kemungkinan yang dievaluasi pada $\boldsymbol{\mu}$ dan p merupakan jumlah parameter. Model terbaik adalah model dengan AIC yang lebih kecil.

METODE PENELITIAN

Dalam paper ini dibandingkan performa ke empat metode pemodelan data cacah, yaitu Poisson, Negatif Binomial, ZIP, dan ZIBN dalam memodelkan data daftar anggota rumah tangga yang diperoeh dari Data Survey Demografi dan Kesehatan Indonesia (SDKI) Tahun 2017. Adapun ukuran contoh yang digunakan adalah 4731 Rumah Tangga. Data jumlah anak yang tidak sekolah usia 7-15 dalam suatu keluarga di Provinsi Jawa Barat merupakan peubah respon (Y) dengan peubah penjelas yaitu:

Tabel 1. Data peubah penjelas dan karakteristiknya

No	Peubah Penjelas	Keterangan
1	Indeks Kekayaan (*Wealth Index Composit* (WIC))	(1) Poorest (2) Poorer (3) Midle (4) Richer (5) Richest
2	Tipe Tempat Tinggal (TPR)	(1) Urban (2) Rural
3	Tingkat Pendidikan Orang Tua (TPO)	(1) SD (2) SMP (3) SMA (4) D3 (5) ≥ S1 (8) Tidak Tahu

Secara garis besar adapun langkah-langkah metode penelitian pada paper ini adalah:

1. Identifikasi karakteristik data pada peubah Y (ATS) dengan menghitung nilai p (peluang nol) dan n.
2. Eksplorasi peubah Y secara deskriptif dengan histogram untuk mengetahui indikasi dari kondisi sebaran Poisson.
3. Melakukan uji *chi-square* pada peubah Y untuk mengidentifikasi peubah Y menyebar Poisson atau ZIP.
4. Melakukan uji skor pada peubah Y untuk mengetahui terjadinya peluang nol berlebih atau tidak.
5. Melakukan analisis regresi Poisson, BN, ZIP, dan ZINB kemudian menguji penduga koefisien parameter regresi dengan uji Wald. Analisis menggunakan *R Program* versi 3.5.1
6. Membandingkan dengan mengevaluasi *Goodness of Fit* Model
7. Melakukan analisis regresi terbaik dan menguji penduga koefisien parameter regresi dengan uji Wald.
8. Melakukan uji Pearson Chi-Square pada regresi terbaik untuk mengetahui terjadi overdispersi atau tidak.

HASIL DAN PEMBAHASAN

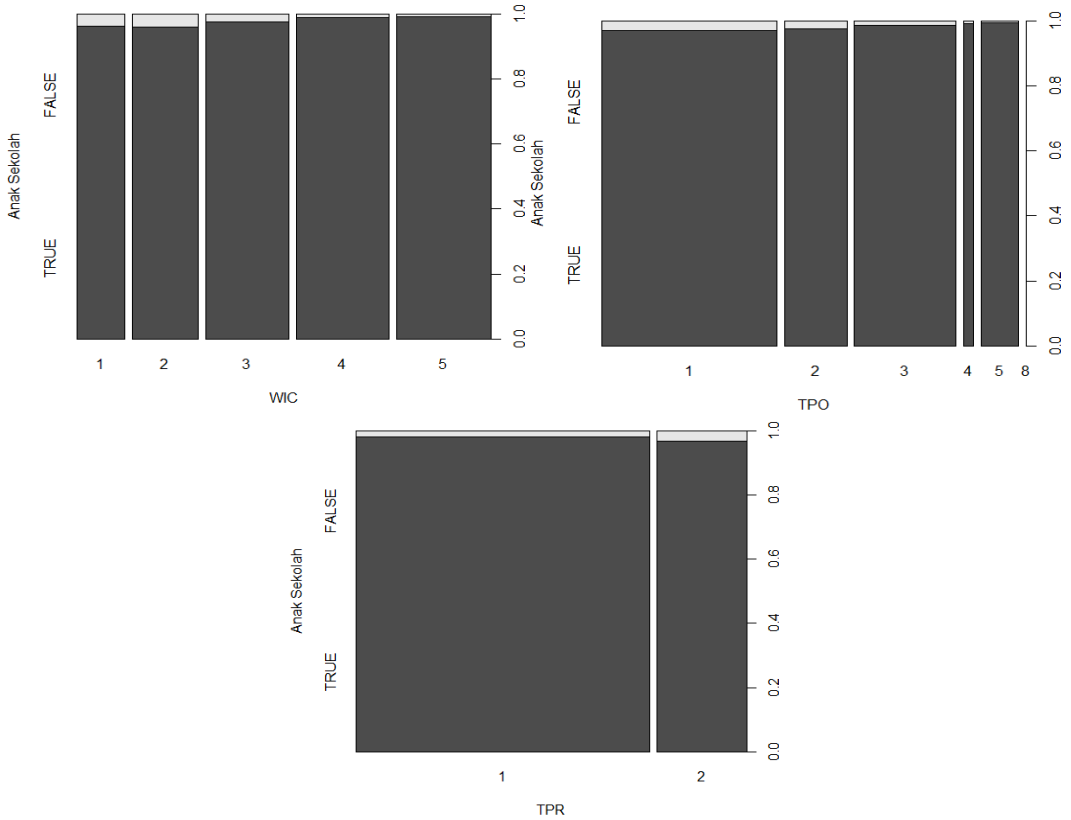
Identifikasi terhadap peubah Y diperlihatkan pada Tabel 2 yang menunjukkan bahwa peubah respon Y diidentifikasi mempunyai nol berlebih yaitu dengan frekuensi 4633 atau sebesar 97.93% dan penduga rata-rata kejadian ($\hat{\lambda}$) adalah 0.022 yang nilainya hampir mendekati nilai 0. Namun, hal ini di identifikasi lebih lanjut melalui uji skor.

Tabel 2. Eksplorasi sebaran data Y

Data jumlah anak yang tidak sekolah	Jumlah	Persentase
0	4633	97.93%
1	93	1.97%
2	5	0.10%
N	4731	100.00%
$\hat{\lambda}$	0.022	

FIBONACCI : Jurnal Pendidikan Matematika dan Matematika. Vol. 5 (1), pp: 71 - 80.

Identifikasi awal dari pengaruh peubah penjelas terhadap peubah respon dapat dilihat dari Spinogram pada Gambar 1. Gambar 1 memperlihatkan bahwa setiap tingkatan kategori pada peubah penjelas berpotensi memberikan peluang nol berlebih pada peubah respon Y, terlihat dari digram batang yang berwarna hitam bernilai 0 untuk peubah repon Y.



Gambar 1. Spinogram dari peubah penjelas

Tabel 3 menunjukkan bahwa jumlah kejadian anak tidak sekolah pada usia 7-15 dalam rumah tangga tidak menyebar Poisson dan ZIP pada α sebesar 0.05. Namun, jika dipilih dari kedua sebaran tersebut, sebaran Poisson menghasilkan nilai χ^2 yang lebih kecil dibandingkan ZIP, maka *Y* dapat dikatakann mendekati sebaran Poisson. Perhatikan juga bahwa hasil uji skor yang menolak H_0 pada $\alpha = 0.05$ karena $\omega = 13.972$ yang lebih besar dari $\chi^2_{0.05,1} = 3.841$ menunjukkan bahwa terjadinya peluang nol berlebih sebagai penyebab terjadinya overdispersi pada paubah Y yaitu berkisar 97.93%. Oleh karena adanya pelanggaran asumsi dalam regresi Poisson yaitu E[*Y*] > Var[*Y*] dan hasil uji skor yang berbeda dengan hasil uji *Chi-square*, maka pada contoh kasus ini penanganan overdispersi menggunakan dua model regresi ZIP dan ZIBN, yang juga akan dibandingkan dengan model regresi Poisson dan regresi Binomial Negatif dari peubah respon Y dengan peubah bebas WIC, TPR, dan TPO.

Tabel 3. Identifikasi Sebaran Y

Tipe sebaran	$\chi^2_{0.05;1}$	χ^2_{hitung}	**Keputusan**
Poisson	3.841	12.898	Tolak Ho
ZIP		5051.110	Tolak Ho

Pada tabel 4 terlihat bahwa model regresi ZIP dengan peubah TPR yang merupakan model terbaik untuk penanganan overdispersi. Hal ini dilihat dari nilai AIC dan BIC paling kecil diantara model lainnya, yaitu 944.11 dan 1028.1, begitu juga dengan nilai LR *Chi-Square*, yaitu 918.11, walaupun model ZIBN dengan peubah TPR memiliki nilai yang hampir sama dengan model ini. Rasio dispersi untuk model ini adalah 0.878, nilai ini hampir mendekati 1, walaupun rasio dispersinya tidak lebih baik dibandingkan model regresi Poisson, namun model ZIP dapat menangani overdispersi karena peluang nol yang berlebih (Naya *et. al.*, 2008). Uji pengaruh peubah penjelas terhadap Y adalah

Tabel 4. Pemilihan Model Terbaik

Model	AIC	LR Chisq	Rasio Dispersi
Poisson	947.31	925.31	0.924
Binomial Negatif	945.93	921.93	0.889
ZIP(WIC+TPR)	948.72	914.72	0.876
ZIP(WIC)	952.53	920.53	0.893
ZIP(TPR)	**944.11**	**918.11**	**0.878**
ZIBN(WIC+TPR)	950.72	914.72	0.876
ZIBN(WIC)	954.53	920.53	0.894
ZIBN(TPR)	946.11	918.11	0.879

Hasil pendugaan parameter menggunakan model ZIP ditunjukkan pada Tabel 5. Terdapat 2 peubah penjelas yang signifikan terhadap peubah Y, yaitu WIC3, WIC4, WIC5, dan TPR2. Dugaan untuk peubah WIC3, WIC4, dan WIC5 berturut-turut adalah -0.646, -1.545, dan -1.503. Hal ini berarti bahwa Indeks Kekayaan untuk kategori "*Poorest*/Termiskin" yang dijadikan sebagai referensi, berpengaruh paling besar dalam meningkatkan jumlah anak yang tidak sekolah dalam keluarga dibanding Indeks kekayaan lainnya. Keluaga dengan indeks kekayaan "*Middle* / Menengah" memiliki kecenderungan untuk meningkatkan jumlah anak tidak sekolah adalah $e^{-0.069}$ atau 0.933 kali dibanding keluarga dengan indeks kekayaan "*Poorest*/Termiskin" dan memiliki pengaruh yang signifikan terhadap peningkatan jumlah. Jika dibandingkan dengan keluarga dengan indeks kekayaan "Termiskin", indeks kekayaan "Menengah" memberikan pengaruh yang hampir sama dengan indeks kekayaan termiskin sedangkan indeks kekayaan "Richer/Lebih kaya" memberikan pengaruh yang paling rendah dibanding yang lain, yaitu $e^{-1.545}$ atau 0.213 kali dibanding indeks kekayaan "Termiskin".

Tabel 5. Dugaan Parameter Model ZIP

Peubah	Derajat bebas	Dugaan	W_i	Keputusan
	Model data diskret untuk λ			
Intercept	1	-3.015	-6.183	Tolak H_0
WIC2	1	-0.069	-0.244	Terima H_0
WIC3	1	-0.646	-2.058	Tolak H_0
WIC4	1	-1.545	-3.886	Tolak H_0
WIC5	1	-1.503	-3.267	Tolak H_0
TPR2	1	-1.387	-2.925	Tolak H_0

Peubah	Derajat bebas	Dugaan	W_i	Keputusan
TPO2	1	0.102	0.366	Terima H_0
TPO3	1	-0.118	-0.375	Terima H_0
TPO4	1	-0.271	-0.253	Terima H_0
TPO5	1	-0.921	-1.180	Terima H_0
TPO8	1	-10.632	-0.014	Terima H_0
Model *zero-inflation* untuk *p*				
Intercept	1	-1.182	2.135	Tolak H_0
TPR2	1	-15.758	-0.009	Terima H_0

Peubah penjelas lain yang berpengaruh terhadap jumlah anak tidak sekolah dalam keluarga adalah Tipe Tempat Tinggal dengan kategori "*Rural*/Pedesaan" yang memiliki kecenderungan untuk meningkatkan jumlah anak tidak sekolah sebesar $e^{-1.387}$ atau 0.250 kali dari keluarga yang tinggal di wilayah "*Urban*/Perkotaan". Berdasarkan Tabel 5 di atas, maka model regresi ZIP pada peubah WIC, TPR, dan TPO terhadap Y (ATS) adalah:

1. Model data diskret untuk λ adalah

$$\hat{\lambda}_i = \exp(-3.015 - 0.069\text{WIC2} - 0.646\text{WIC3} - 1.545\text{WIC4} - 1.503\text{WIC5} - 1.387\text{TPR2} + 0.102\text{TPO2} - 0.118\text{TPO3} - 0.271\text{TPO4} - 0.921\text{TPO5} - 10.632\text{TPO8})$$

2. Model *zero-inflation* untuk *p* adalah

$$\hat{p}_i = \frac{\exp(1.182 - 15.758\text{TPR2})}{1 + \exp(1.182 - 15.758\text{TPR2})}$$

dengan penduga *y* adalah $\hat{y}_i = (1 - \hat{p}_i)\hat{\lambda}_i$. Ukuran kebaikan dari model ini adalah 944.11 untuk AIC, dan nilai LR *Chi-square* sebesar 918.11. Sedangkan nilai τ = 0.878 dengan Statistik uji $\chi^2 = 4144.929$ bernilai lebih kecil jika dibandingkan dengan sebaran χ^2 dengan derajat bebas 4718 nilai $\chi^2 = 4878.908$, hal ini berarti bahwa keputusannya adalah tidak tolak H_0 pada α, sehingga hasil ini berarti bahwa peubah *Y* tidak terjadi overdipersi pada α=0.05. Hasil terbukti dari nilai rasio τ sebesar 0.878 yang menunjukkan bahwa rasio τ bernilai kurang dari 1.

SIMPULAN

Berdasarkan ukuran *Goodness of Fit* model ZIP memberikan performa yang cukup baik dibanding model Poisson, Binomial Negatif, dan ZIBN. Walaupun berdasarkan identifikasi dari sebaran *Y* tidak mengikuti sebaran Poisson dan ZIP, namun berdasarkan uji skor, data *Y* terbukti mempunyai nilai amatan nol yang berlebih, sehingga pemodelan *Zero-Inflated* dapat digunakan untuk menangani overdispersi karena nilai nol berlebih. Model BN baik dalam mengatasi masalah overdispersi dibanding model Poisson, namun jika diidentifikasi data cacah memiliki nilai nol berlebih, model BN belum cukup baik dibandingakan dengan model *Zero-Inflated.*

DAFTAR PUSTAKA

Hausman, J, BH. Hall and Z Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica.*Vol. 52 (4), pp: 909-938.

Ismail, N and Abdul AJ. 2007. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. Virginia: *Casualty Actuarial Society Forum,* Winter 2007.

Jansakul N, Hinde JP. 2002. "Score Test for Zero-Inflated Poisson Models". *Computational Statistics and Data Analysis*. Vol. 40 (1) :75-96.

Jeong, KM. 2017. "Modelling Count Responses with Overdispersion". *Communication of the Korean Statistical Society* Vol. 19 (6), pp: 761-770.

Jiang, Y. and L. House. 2017. "Comparison of the Performance of Count Data Models under Different Zero-Inflation Scenarios Using Simulation Studies". *In 2017 Annual Meeting, July 30-August 1, 2017. Chicago.* Agricultural & Applied Economics Association.

Lambert, D. 1992. "Zero-Inflated Poisson Regression with Application to Defects in Manufacturing". *Technometrics*. Vol. 34 (1), pp: 1-14.

McCullagh, P. and J. Nelder. 1989. *Generalized Linear Models* (second ed.). London: Chapman and Hall.

Naya H, Urioste JI, Chang YM, Motta MR, Kremer R, Gianola D. 2008. "A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in *Corriedale* sheep". *Genetics Selection Evolution.* Vol. 40 (4), pp: 379-394.

Nelder, J.A. and *Wedderburn*, R.W.M. *1972*. "Generalized Linear Models". *Journal of the Royal Statistical Society, Series A.* Vol. 135 (3), pp: 370-384.

Özdemir, T and Ecevit E. 2005. "Comparison of Chi-Square and Likelihood Ratio Chi-Square Tests: Power of Test". *Journal of Applied Sciences Research*. Vol. 1 (2), pp: 242-244.

Palmgren, Juni. 1981. "The Fisher Information Matrix for Log-Linear Models Arguing Conditionally in the Observed Explanatory Variables". *Biometrika.* Vol. 68 (2), pp: 563-566.

Zeiless *et. al.* 2008. "Regression Models for Count Data in R". *Journal of Statistical Software* Vol. 27 (8), pp: 1-25.