PEMODELAN FREKUENSI KLAIM ASURANSI KENDARAAN MENGGUNAKAN MODEL REGRESI *ZINB*

Hantrisa Nurrizki, *Retno Budiarti, Nur Agustiani, dan Siswandi

Sekolah Sains Data, Matematika, dan Informatika,
Institut Pertanian Bogor, Jl. Meranti, Kampus IPB Dramaga Bogor.
hantrisanurrizki@apps.ipb.ac.id, *retnobu@apps.ipb.ac.id, nur_agustiani@apps.ipb.ac.id, siswandi@apps.ipb.ac.id

Abstrak

Pemodelan frekuensi klaim asuransi kendaraan dilakukan untuk mendukung penetapan premi dan pengelolaan risiko yang lebih akurat. Penelitian ini berfokus pada data frekuensi klaim asuransi kendaraan yang memiliki permasalahan *zero inflation* dan overdispersi. Data yang digunakan terdiri dari 6661 polis asuransi kendaraan bermotor Spanyol pada tahun 2016 hingga 2017. Analisis dilakukan menggunakan model regresi *Zero Inflated Negative Binomial* (ZINB). Hasil penelitian menunjukkan bahwa model regresi ZINB relatif lebih baik dibandingkan dengan model binomial negatif dalam memprediksi frekuensi klaim berdasarkan nilai AIC dan MAE. Variabel yang signifikan memengaruhi frekuensi klaim adalah umur pemegang polis, kapasitas mesin, daya kendaraan, dan premi, sementara nilai kendaraan memengaruhi peluang tidak mengajukan klaim. Hasil penelitian ini diharapkan dapat membantu perusahaan asuransi dalam memahami risiko klaim serta mendukung penetapan premi dan pengelolaan risiko.

Kata kunci: Asuransi kendaraan, Frekuensi klaim, Overdispersi, *Zero inflation*, ZINB

1 Pendahuluan

Seiring dengan meningkatnya jumlah kendaraan di jalan raya, angka kecelakaan lalu lintas juga mengalami peningkatan. Kecelakaan lalu lintas ini menjadi salah satu penyebab kematian dan berada pada peringkat ke-9 di dunia [1]. Tingginya angka kecelakaan tersebut menimbulkan risiko yang besar, baik berupa kerusakan kendaraan maupun kerugian finansial. Untuk meminimalkan dampak kerugian tersebut, asuransi kendaraan dapat menjadi salah satu upaya perlindungan dalam mengurangi risiko yang mungkin timbul akibat kecelakaan. Berdasarkan laporan Asosiasi Asuransi Umum Indonesia (AAUI), pada tahun 2017 hingga 2018 asuransi kendaraan menempati peringkat kedua sebagai penyumbang pendapatan premi terbesar dalam kategori asuransi umum [2]. Asuransi kendaraan memberikan perlindungan finansial bagi pemilik kendaraan dari kerugian akibat kecelakaan ataupun kerusakan kendaraan.

Salah satu kunci penting dalam industri asuransi adalah kemampuan untuk memprediksi frekuensi klaim secara akurat. Hal ini menjadi dasar perusahaan asuransi untuk menetapkan kebijakan yang tepat dan menilai risiko yang terjadi di masa depan. Namun, pada kenyataannya, frekuensi klaim asuransi yang diajukan oleh pemegang polis

2020 Mathematics Subject Classification: 62P05.
Diajukan: 02/06/2026, diterima: 26/06/2026. DOI: https://doi.org/10.29244/milang.22.1.97-110
MILANG Journal of Mathematics and Its Applications, Vol.22, No.1, pp.97-110 ISSN: 2963-5233

bervariasi. Frekuensi klaim merupakan data berbentuk cacah yang merepresentasikan banyaknya klaim dalam suatu periode tertentu dan dapat bernilai nol, satu, atau lebih [3]. Untuk data cacah dengan peluang berhasil relatif kecil dalam kasus ini seperti terjadinya klaim, model yang umum digunakan adalah regresi Poisson, yang memiliki karakteristik bahwa rata-rata kejadian sama dengan variansnya [4]. Namun, dalam data frekuensi klaim asuransi kendaraan, dijumpai nilai varians lebih besar daripada rata-rata yang disebut dengan overdispersi sehingga model regresi Poisson kurang cocok digunakan untuk data frekuensi klaim asuransi kendaraan. Model yang dapat mengatasi masalah overdispersi adalah model regresi binomial negatif. Selain itu, pada asuransi kendaraan terdapat periode di mana pemegang polis tidak mengajukan klaim sama sekali sehingga menghasilkan proporsi nilai nol dalam data lebih besar daripada yang diperkirakan dalam model binomial negatif yang disebut dengan *zero inflation*. Model *Zero Inflated Negative Binomial* (ZINB) dapat digunakan untuk mengatasi masalah overdispersi sekaligus mengatasi masalah *zero inflation* dalam pemodelan frekuensi klaim asuransi kendaraan [5].

Model ZINB sendiri telah banyak digunakan pada berbagai kasus data cacah dengan adanya nilai nol yang berlebih pada variabel respon. Penelitian sebelumnya yang dilakukan oleh Purnama [6], menunjukkan bahwa model ZINB mampu mengendalikan permasalahan nol yang berlebih dan overdispersi pada data jumlah konsumsi rokok harian dengan nilai *Akaike's Information Criterion* (AIC) terkecil. Selain itu, model ZINB juga digunakan untuk memprediksi frekuensi kematian akibat kecelakaan lalu lintas dan memberikan hasil yang lebih baik dibandingkan regresi Poisson [7].

Walaupun model ZINB telah banyak digunakan pada berbagai penelitian, penerapannya pada data klaim asuransi kendaraan yang memiliki karakteristik adanya overdispersi dan *zero inflation* masih relatif terbatas. Penelitian sebelumnya yang membandingkan kinerja berbagai model data cacah, seperti regresi Poisson, regresi binomial negatif, *Zero Inflated Poisson* (ZIP), ZINB, Hurdle Poisson, dan Hurdle *Negative Binomial*, dengan beberapa metode *machine learning* seperti *Artificial Neural Network* (ANN), *Random Forest*, dan *Support Vector Machine* (SVM) telah dilakukan oleh Alomair [5]. Hasil penelitian tersebut menunjukkan bahwa pada kelompok model data cacah, model terbaik dihasilkan oleh ZIP dan ZINB sedangkan secara keseluruhan metode SVM memberikan kinerja yang terbaik dalam menangani permasalahan *zero inflation* pada data klaim asuransi kendaraan berdasarkan nilai *Mean Absolute Error* (MAE). Berbeda dengan penelitian tersebut yang menekankan perbandingan model regresi dengan metode *machine learning*, penelitian ini berfokus pada pemodelan berbasis regresi data cacah, khususnya penerapan model ZINB, dengan menggunakan regresi binomial negatif sebagai pembanding. Evaluasi model dilakukan tidak hanya berdasarkan *Mean Absolute Error* (MAE), tetapi juga menggunakan *Akaike's Information Criterion* (AIC) untuk menilai kecocokan model dan memilih model yang lebih sesuai dalam menangani overdispersi dan nilai nol yang berlebih pada data frekuensi klaim.

Pemilihan model yang tidak tepat dalam analisis klaim asuransi dapat berdampak buruk terhadap hasil prediksi. Bagi perusahaan asuransi, hal ini berpotensi menimbulkan ketidaktepatan dalam penentuan premi maupun pengelolaan risiko. Oleh karena itu, penelitian ini bertujuan: (1) melakukan pemodelan frekuensi klaim asuransi kendaraan menggunakan regresi *Zero Inflated Negative Binomial* (ZINB) serta membandingkan kinerjanya dengan model regresi binomial negatif, (2) penelitian ini juga bertujuan untuk mengetahui faktor-faktor yang berpengaruh terhadap frekuensi klaim kendaraan bermotor.

Dari sisi akademis, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan literatur terkait pemodelan data cacah dengan permasalahan *zero inflation* dan overdispersi.

2 Metode Penelitian

2.1 Sumber Data

Data yang digunakan pada penelitian ini berasal dari *platform* Kaggle dengan judul *"3-Year Non-Life Motor Insurance Dataset"* yang dipublikasikan oleh pengguna Kaggle dengan nama akun Jocelyn Dumlao. *Dataset* ini dikembangkan dan disediakan oleh Josep Lledó dari *Universitat de Valencia*, Spanyol. Data tersebut berisi informasi mengenai aktivitas transaksi polis asuransi kendaraan bermotor non jiwa di Spanyol. Penelitian ini menggunakan data pada periode tahun polis 2016 hingga 2017, sebanyak 6661 observasi dan terdiri atas 30 variabel. Meskipun *dataset* ini terdiri dari 30 variabel, penelitian ini hanya menggunakan delapan variabel. Variabel yang dipilih adalah variabel yang berpotensi memengaruhi frekuensi klaim serta mengutamakan variabel dengan tipe data numerik atau *integer*.

2.2 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari satu variabel respon dan tujuh variabel prediktor. Variabel respon yang digunakan adalah frekuensi klaim asuransi kendaraan, sedangkan variabel prediktor yang digunakan disajikan pada Tabel 1.

Tabel 1. Variabel penelitian

Variabel	Notasi	Tipe Data	Keterangan
Frekuensi klaim asuransi	Y	*Integer*	Frekuensi klaim asuransi kendaraan bermotor yang diajukan oleh pemegang polis dalam satu periode polis.
Umur pemegang Polis	X_1	Numerik	Usia pemegang polis dalam tahun.
Lama pengalaman mengemudi	X_2	Numerik	Lama pengalaman pemegang polis dalam mengemudi.
Umur kendaraan	X_3	Numerik	Usia kendaraan dalam tahun.
Nilai kendaraan	X_4	Numerik	Nilai pasar kendaraan dalam EUR.
Kapasitas mesin kendaraan	X_5	*Integer*	Kapasitas mesin kendaraan dalam cc.
Daya kendaraan	X_6	*Integer*	Daya maksimum kendaraan dalam HP (*Horsepower*).
Premi	X_7	Numerik	Premi yang dibayarkan oleh pemegang polis selama tahun berjalan dalam EUR.

2.3 Model Regresi Binomial Negatif

Model regresi binomial negatif merupakan salah satu terapan dari metode GLM (*Generalized Linear Model*). Model ini terdiri dari tiga komponen, yakni komponen

random, komponen sistematik, dan *link function*. Misalkan Y adalah variabel respon yang berdistribusi binomial negatif dengan asumsi bahwa $var(Y) > E[Y]$. Distribusi binomial negatif merupakan campuran distribusi Poisson-gamma dengan fungsi massa peluang sebagai berikut [8]:

$$Y|\Lambda \sim \text{Poisson}\ (\lambda)$$
$$\Lambda \sim \text{gamma}\ (\alpha, \beta), \lambda, \alpha, \beta > 0$$

$$P(Y = y|\alpha, \beta) = \frac{\Gamma(y + \alpha)}{y!\,\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{y}, y = 0,1,2,\ldots \quad (1)$$

$$P(Y = y|\alpha, \beta) = \frac{(y + \alpha - 1)!}{y!\,(\alpha - 1)!}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{y} \quad (2)$$

$$P(Y = y|\alpha, \beta) = \binom{y + \alpha - 1}{y}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{y}. \quad (3)$$

Untuk membentuk model regresi pada distribusi binomial negatif, rata-rata dinyatakan sebagai:

$$\mu = E[Y] = \alpha\beta \quad (4)$$

sehingga varians dapat ditulis ulang menjadi:

$$\text{Var}(Y) = \alpha\beta + \alpha\beta^2 = \mu + \frac{\mu^2}{\alpha} \quad (5)$$

Misalkan $\theta = \frac{1}{\alpha}$, dengan $\theta > 0$, maka varians menjadi:

$$\text{Var}(Y) = \mu + \theta\mu^2 \quad (6)$$

Dengan demikian fungsi massa peluang binomial negatif dapat dinyatakan sebagai:

$$P(y; \mu, \theta) = \frac{\Gamma\left(y + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)\Gamma(\text{y} + 1)}\left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(1 - \frac{1}{1+\theta\mu}\right)^{y}, y = 0,1,2,\ldots \quad (7)$$

$$P(y; \mu, \theta) = \frac{\Gamma\left(y + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)y!}\left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(1 - \frac{1}{1+\theta\mu}\right)^{y}, y = 0,1,2,\ldots \quad (8)$$

dengan y adalah variabel respon, μ adalah rata-rata, dan θ adalah parameter dispersi. Dengan menggunakan *link function*, diperoleh model regresi binomial negatif untuk memodelkan data cacah adalah sebagai berikut:

$$g(\mu_i) = \ln(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, i = 1,2,\ldots,n \quad (9)$$

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\right), \quad (10)$$

dengan β_j adalah parameter model ke-j dan x_{ji} adalah nilai variabel ke-j pada observasi ke-i dengan $j = 0,1,2,\ldots,k$ dan k banyaknya parameter [9].

2.4 Model Regresi *ZINB*

Model regresi *Zero Inflated Negative Binomial* (ZINB) memiliki asumsi bahwa data terbentuk melalui dua proses pendugaan yang berbeda yang ditentukan oleh uji coba

Bernoulli [6]. Misalkan Y_i adalah variabel acak diskret yang saling bebas, maka kemunculan nilai nol dianggap berasal dari dua proses terpisah. Proses pertama adalah nol murni dengan peluang p_i. Proses kedua adalah nol yang dihasilkan dari distribusi binomial negatif dengan rata-rata μ dan peluang sebesar $(1-p_i)$. Peluang total munculnya nilai nol merupakan kombinasi dari peluang nol yang berasal dari kedua proses tersebut dengan fungsi massa peluang sebagai berikut [10]:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)(1+\theta\mu_i)^{-\frac{1}{\theta}}, & y_i = 0 \\ (1-p_i)\dfrac{\Gamma\left(y_i + \frac{1}{\theta}\right)(\theta\mu_i)^{y_i}}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\theta}\right)(1+\theta\mu_i)^{y_i+\frac{1}{\theta}}}, & y_i = 1,2,\dots \end{cases} \tag{11}$$

dengan $i = 1,2,3,\dots,n$, $0 \le p_i \le 1$, $\mu_i \ge 0$, dan θ adalah parameter dispersi.

Terdapat dua komponen model regresi ZINB [11], yaitu:

1. Model data diskret untuk μ_i:
$$\ln(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{12}$$
$$\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_a x_{ai}). \tag{13}$$
dengan β_j adalah parameter model dengan $j = 0,1,2,\dots,a$, $i = 1,2,\dots,n$, dan a banyaknya parameter untuk model data diskret.
2. Model *zero inflation* untuk p_i:
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\gamma}, \tag{14}$$
$$p_i = \frac{\exp(\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_b x_{bi})}{1+\exp(\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_b x_{bi})}, \tag{15}$$
dengan γ_m adalah parameter ke-m untuk model *zero inflation*, $m = 0,1,2,\dots,b$, $i = 1,2,\dots,n$, $0 \le p_i \le 1$, dan b banyaknya parameter untuk model *zero inflation*.

Uji statistik model *Zero Inflated Negative Binomial* dilakukan menggunakan uji rasio *Likelihood* dan uji Wald. Uji rasio *Likelihood* dilakukan untuk mengetahui apakah model secara keseluruhan signifikan, yaitu apakah setidaknya terdapat satu variabel prediktor yang berpengaruh terhadap frekuensi klaim asuransi kendaraan dengan hipotesis sebagai berikut:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_a = 0$ dan $\gamma_1 = \gamma_2 = \cdots = \gamma_b = 0$,

H_1: Ada $\beta_j \neq 0$ atau $\gamma_m \neq 0$, dengan $j \in \{0,1,2,\dots,a\}$ dan $m \in \{0,1,2,\dots,b\}$.

dimana β_j adalah parameter ke-j dari model diskret dan γ_m adalah parameter ke-m untuk model *zero inflation*. Statistik uji:

$$G = -2\ln\left(\frac{L_0}{L_1}\right), \tag{16}$$

dengan L_0 adalah *likelihood* tanpa variabel prediktor dan L_1 adalah *likelihood* dengan variabel prediktor [7]. Tolak H_0 pada taraf signifikansi α jika $G > \chi^2_{(\alpha,2k)}$ dengan $k = a + b$.

Sedangkan uji Wald dilakukan untuk mengetahui pengaruh masing-masing variabel prediktor secara sekuensial terhadap variabel respon. Pada model ZINB, uji Wald dilakukan terpisah untuk model data diskret dan model *zero inflation.*

Pendugaan parameter pada model *Zero Inflated Negative Binomial* (ZINB) dilakukan menggunakan metode *Maximum Likelihood Estimation* (MLE). Metode ini bertujuan untuk memperoleh nilai parameter yang memaksimumkan fungsi *likelihood* berdasarkan data yang diamati. Parameter yang diestimasi dalam model ZINB meliputi parameter regresi pada model data diskret $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_a)$, parameter regresi pada model *zero inflation* $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_b)$, serta parameter dispersi θ dengan $\theta > 0$.
Didefinisikan *zero indicator* d_i sebagai berikut [11]:

$$d_i = \begin{cases} 1, y_i > 0 \\ 0, y_i = 0 \end{cases}, \tag{17}$$

sehingga fungsi massa peluang ZINB menjadi:

$$P(Y_i = y_i) = \left[p_i + (1 - p_i)(1 + \theta\mu_i)^{-\frac{1}{\theta}}\right]^{1-d} \times \left[(1 - p_i)\frac{\Gamma\left(y_i + \frac{1}{\theta}\right)(\theta\mu_i)^{y_i}}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\theta}\right)(1 + \theta\mu_i)^{y_i + \frac{1}{\theta}}}\right]^{d}. \tag{18}$$

Berdasarkan fungsi massa peluang di atas, fungsi *likelihood* model regresi ZINB dapat dituliskan sebagai:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta) = \prod_{i=1}^{n}\left[p_i + (1 - p_i)(1 + \theta\mu_i)^{-\frac{1}{\theta}}\right]^{1-d_i}\left[(1 - p_i)\frac{\Gamma\left(y_i + \frac{1}{\theta}\right)(\theta\mu_i)^{y_i}}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\theta}\right)(1 + \theta\mu_i)^{y + \frac{1}{\theta}}}\right]. \tag{19}$$

Selanjutnya, fungsi log-*likelihood* diperoleh sebagai berikut:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta) = \sum_{i=1}^{n}(1 - d_i)\ln\left[p_i + (1 - p_i)(1 + \theta\mu_i)^{-\frac{1}{\theta}}\right] + \sum_{i=1}^{n} d_i\left[\ln(1 - p_i) + \ln\frac{\Gamma\left(y_i + \frac{1}{\theta}\right)(\theta\mu_i)^{y_i}}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\theta}\right)(1 + \theta\mu_i)^{y + \frac{1}{\theta}}}\right], \tag{20}$$

dengan $\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_a x_{ai})$, $p_i = \frac{\exp(\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_b x_{bi})}{1 + \exp(\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_b x_{bi})}$, $i = 1,2, \dots, n$ dan θ merupakan parameter dispersi dengan $\theta > 0$. Nilai estimasi parameter $\beta_0, \beta_1, \dots, \beta_a, \gamma_0, \gamma_1, \dots, \gamma_b$, dan θ kemudian dapat diperoleh dengan memaksimumkan fungsi log-*likelihood* tersebut dengan a banyaknya parameter untuk model data diskret dan b banyaknya parameter untuk model *zero inflation*

2.5 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan, mulai dari analisis data awal hingga pemodelan dan evaluasi model. Adapun tahapan penelitian sebagai berikut.

1. Melakukan deskripsi data dengan statistik deskriptif. Selanjutnya dilakukan pemeriksaan asumsi GLM (*Generalized Linear Model*) yang terdiri dari uji multikolinearitas dan identifikasi sebaran variabel respon Y dengan melihat nilai rata-rata dan varians Y.
2. Jika variabel respon menunjukkan adanya overdispersi, maka selanjutnya digunakan model regresi binomial negatif dengan langkah-langkah meliputi estimasi parameter $\beta_0, \beta_1, \dots, \beta_k$ dan θ menggunakan *Maximum Likelihood Estimation* (MLE), serta uji statistik model menggunakan uji rasio *likelihood* dan dilanjutkan uji Wald.
3. Pemodelan *Zero Inflated Negative Binomial* (ZINB). Sebelum dilakukan pemodelan, dilakukan pemeriksaan adanya *zero inflation* pada variabel respon. Model ini digunakan untuk mengatasi data dengan overdispersi sekaligus *zero inflation*. Pada

tahap ini dilakukan estimasi parameter $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta)$ menggunakan *Maximum Likelihood Estimation* (MLE) serta uji statistik model menggunakan uji rasio *likelihood* pada persamaan (16) dan dilanjutkan uji Wald pada persamaan (17) dan (18). Sebelum masuk ke tahap pemodelan, variabel nilai kendaraan (X_4) diasumsikan dimasukkan ke dalam komponen *zero inflation* yang berarti bahwa nilai kendaraan dapat memengaruhi kemungkinan suatu polis termasuk dalam kelompok yang tidak pernah mengajukan klaim.

4. Melakukan perbandingan nilai AIC dan MAE antara model binomial negatif dan ZINB untuk menentukan model yang relatif lebih baik dalam memprediksi frekuensi klaim asuransi kendaraan, dengan rumus sebagai berikut.

$$AIC = -2 \ln L\left(\hat{\beta}\right) + 2k \tag{21}$$

$$MAE = \frac{\sum |Y' - Y|}{n} \tag{22}$$

dengan k adalah banyaknya parameter, $L\left(\hat{\beta}\right)$ adalah nilai *likelihood*, Y' adalah nilai prediksi dari model, Y merupakan nilai aktual, dan n adalah banyaknya observasi.

3 Hasil dan Pembahasan

3.1 Deskripsi Data dan Pemeriksaan Awal Data

Untuk menggambarkan karakteristik data, dilakukan analisis deskriptif pada seluruh variabel yang digunakan pada penelitian. Statistik yang digunakan meliputi nilai minimum, maksimum, dan rata-rata. Hasil analisis deskriptif disajikan pada Tabel 2.

Tabel 2. Statistik deskriptif

Variabel	Minimum	Maksimum	Rata-Rata
Y	0.00	14.00	0.66
X_1	18.68	75.87	44.44
X_2	0.01	55.77	21.55
X_3	1.00	61.00	11.07
X_4	480.80	143892.00	17636.20
X_5	49.00	5439.00	1590.00
X_6	0.00	476.00	90.80
X_7	40.29	2596.77	323.49

Dapat dilihat pada Tabel 2 bahwa rata-rata frekuensi klaim (Y) asuransi dari tahun 2016 sampai 2017 adalah 0.66 kali yang menunjukkan bahwa sebagian besar pemegang polis tidak melakukan klaim selama periode pertanggungan. Dari segi pemegang polis, yakni umur pemegang polis (X_1) dan lama pengalaman mengemudi (X_2), rata-rata berada pada usia yang memenuhi syarat legal untuk mengemudi dan masih berada pada usia yang relatif produktif serta merupakan pengemudi yang berpengalaman dan berpotensi memiliki risiko klaim yang lebih rendah dibanding pengemudi pemula. Variabel umur kendaraan (X_3) menunjukkan bahwa rata-rata kendaraan yang diasuransikan termasuk kendaraan lama. Variabel nilai kendaraan (X_4) memiliki rentang nilai yang cukup lebar yang menandakan adanya variasi nilai kendaraan yang cukup tinggi antar pemegang polis. Kapasitas mesin kendaraan (X_5) yang diasuransikan

cenderung berada pada nilai menengah dalam rentang data, meskipun terdapat beberapa kendaraan dengan kapasitas mesin yang sangat kecil maupun sangat besar. Rentang nilai yang cukup lebar ini mengindikasikan adanya keragaman kapasitas mesin yang diasuransikan oleh pemegang polis. Sedangkan daya kendaraan (X_6) menunjukkan bahwa sebagian besar kendaraan berada pada rentang menengah. Nilai minimum 0.00 tetap dipertahankan dalam analisis karena merupakan bagian dari dataset. Nilai tersebut dapat mengindikasikan daya kendaraan yang sangat rendah atau kondisi daya kendaraan yang tidak tercatat pada data sehingga statistik deskriptif merepresentasikan keseluruhan observasi. Sementara itu, variabel premi (X_7) memiliki rentang yang cukup lebar yang menunjukkan bahwa besaran premi bervariasi antar pemegang polis yang kemungkinan dipengaruhi oleh faktor risiko seperti umur pemegang polis, lama mengemudi, dan karakteristik kendaraan.

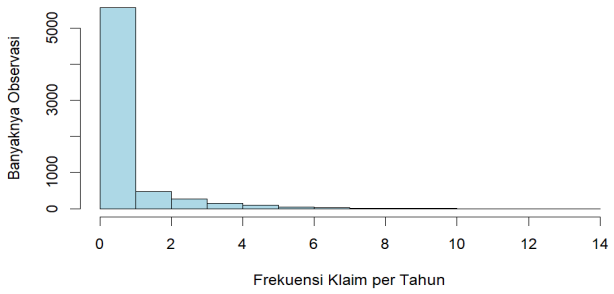
Salah satu asumsi yang harus dipenuhi dalam model GLM (*Generalized Linear Model*) adalah tidak adanya korelasi antar variabel prediktor. Dikarenakan penelitian ini akan menggunakan model GLM, maka dilakukan uji multikolinearitas dengan melihat nilai *Variance Inflation Factor* (VIF). Berikut disajikan nilai VIF dari masing-masing variabel prediktor.

Tabel 3. Nilai VIF masing-masing variabel prediktor

Notasi	Variabel	Nilai VIF
X_1	Umur pemegang polis	3.8648
X_2	Lama pengalaman mengemudi	3.8998
X_3	Umur kendaraan	1.1510
X_4	Nilai kendaraan	3.7574
X_5	Kapasitas mesin kendaraan	2.7875
X_6	Daya kendaraan	3.6598
X_7	Premi	1.5050

Dapat dilihat dari Tabel 3 bahwa nilai VIF yang dihasilkan oleh seluruh variabel prediktor kurang dari 10 yang artinya tidak teridentifikasi adanya multikolinearitas antar variabel sehingga seluruh variabel prediktor dapat digunakan dalam proses pemodelan.

Selanjutnya, eksplorasi awal kesesuaian distribusi Poisson pada variabel respon *Y* dilakukan menggunakan histogram sebagai berikut.



Gambar 1. Histogram distribusi frekuensi klaim (Y)

Dari Gambar 1 dapat dilihat bahwa data frekuensi klaim memiliki nilai non negatif sehingga masih memungkinkan untuk dimodelkan dengan distribusi Poisson. Namun, untuk memastikan apakah distribusi frekuensi klaim benar-benar berdistribusi Poisson, dilakukan pemeriksaan lebih lanjut terhadap rata-rata dan variansnya. Jika rata-rata sama dengan nilai variansnya maka asumsi distribusi Poisson terpenuhi. Diperoleh bahwa rata-rata (Y) sebesar 0.6598 dan varians (Y) sebesar 2.0843. Terlihat bahwa nilai varians lebih besar daripada rata-rata yang menunjukkan adanya pelanggaran asumsi dasar distribusi Poisson. Hal tersebut mengindikasikan bahwa variabel frekuensi klaim tidak berdistribusi Poisson dan terindikasi terjadi overdispersi pada data.

3.2 Model Regresi Binomial Negatif

Pada regresi binomial negatif, diasumsikan bahwa varians lebih besar daripada rata-rata. Model regresi ini dapat digunakan pada data yang mengalami overdispersi seperti pada penelitian ini. Dengan menggunakan *link function* diperoleh model regresi binomial negatif adalah sebagai berikut.

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i}\right)$$

dengan $i = 1, 2, \ldots, 6661$.

Setelah dilakukan pendugaan parameter model regresi binomial negatif, selanjutnya dilakukan uji rasio *likelihood*, dengan statistik uji pada persamaan (16). Uji ini bertujuan untuk mengetahui apakah model secara keseluruhan signifikan, yaitu minimal terdapat satu variabel prediktor yang berpengaruh terhadap frekuensi klaim asuransi kendaraan.

Hasil uji rasio *likelihood* pada taraf signifikansi $\alpha = 0.05$ menunjukkan bahwa nilai statistik uji G =140.74> $\chi^2_{(0.05,7)} = 14.07$ sehingga disimpulkan bahwa minimal ada satu variabel prediktor yang berpengaruh secara signifikan terhadap frekuensi klaim asuransi kendaraan. Selanjutnya dilakukan uji Wald, untuk menguji pengaruh masing-masing variabel prediktor secara sekuensial terhadap frekuensi klaim asuransi kendaraan dengan hipotesis uji H_0: $\beta_j = 0$ dan H_1: $\beta_j \neq 0$ dengan $j = 1,2, \ldots, k$ dengan statistik uji sebagai berikut [12].

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2 \sim \chi^2_{(\alpha,1)}. \tag{23}$$

Tolak H_0 pada taraf signifikansi α jika $W_j > \chi^2_{(\alpha,1)}$. Hasil uji Wald ditampilkan pada Tabel 4 berikut.

Tabel 4. Hasil uji Wald untuk model regresi binomial negatif

Parameter	Koefisien	Statistik uji Wald
$\hat{\beta}_0$	-1.83460	100.250
$\hat{\beta}_1$	0.01122	5.889**
$\hat{\beta}_2$	-0.00549	1.344
$\hat{\beta}_3$	-0.00898	2.680
$\hat{\beta}_4$	-0.00001	0.920
$\hat{\beta}_5$	0.00067	66.473**
$\hat{\beta}_6$	-0.00241	2.727
$\hat{\beta}_7$	0.00106	17.587**

Berdasarkan Tabel 4 di atas, pada taraf signifikansi $\alpha = 0.05$, variabel umur pemegang polis (X_1), kapasitas mesin kendaraan (X_5), dan premi (X_7) memiliki nilai statistik uji lebih dari $\chi^2_{(0.05,1)} = 3.841$ artinya variabel-variabel prediktor tersebut berpengaruh signifikan terhadap frekuensi klaim asuransi.

Selain itu, pada model regresi binomial negatif terdapat parameter dispersi, yaitu θ yang berfungsi untuk menangkap adanya overdispersi, yaitu kondisi ketika varians data lebih besar daripada rata-rata. Hasil estimasi menunjukkan nilai $\hat{\theta} = 0.2670$ dengan standar *error* sebesar 0.0106. Nilai parameter dispersi tersebut mengindikasikan adanya overdispersi pada data frekuensi klaim, sehingga penggunaan model binomial negatif lebih sesuai dibandingkan model regresi Poisson. Diperoleh model regresi binomial negatif sebagai berikut.

$$\mu_i = \exp(-1.83460 + 0.01122x_{1i} + 0.00067x_{5i} + 0.00106x_{7i}) \qquad (24)$$

dengan $i = 1, 2, \dots, 6661$

3.3 Model Regresi *ZINB*

Pada data cacah terdapat dua karakteristik yang sering terjadi yaitu overdispersi dan *zero inflation* [13]. Adanya *zero inflation* pada variabel respon jika proporsi frekuensi klaim nol yang terdapat pada data lebih besar daripada proporsi frekuensi klaim nol yang diprediksi oleh model regresi binomial negatif [5].

Proporsi frekuensi klaim yang dihasilkan pada data yakni sebesar 0.7286 dan proporsi frekuensi klaim yang diprediksi oleh model binomial negatif sebesar 0.7158. Dari hasil tersebut dapat disimpulkan bahwa terdapat indikasi masalah *zero inflation* pada data frekuensi klaim asuransi kendaraan.

Selanjutnya, pemodelan frekuensi klaim dilakukan menggunakan model *Zero Inflated Negative Binomial* (ZINB) karena teridentifikasi adanya *zero inflation* pada data frekuensi klaim. Model ZINB terdiri dari dua komponen, yaitu model data diskret pada persamaan (12) dan (13) dan model *zero inflation* pada persamaan (14) dan (15) dengan μ_i merupakan rata-rata dari frekuensi klaim dan p_i merupakan peluang suatu polis selalu berada pada keadaan nol klaim.

Pada model ini, variabel nilai kendaraan (X_4) ditempatkan di bagian model *zero inflation*. Nilai kendaraan diasumsikan dapat memengaruhi kemungkinan suatu polis tidak pernah mengajukan klaim. Setelah dilakukan estimasi parameter menggunakan *Maximum Likelihood Estimation* (MLE), dilakukan uji rasio *likelihood* pada persamaan (16), untuk mengetahui apakah model secara keseluruhan signifikan, yaitu apakah setidaknya terdapat satu variabel prediktor yang berpengaruh terhadap frekuensi klaim asuransi kendaraan.

Hasil pengujian pada uji rasio *likelihood* pada taraf signifikansi $\alpha = 0.05$ menunjukkan bahwa nilai statistik uji $G = 110.6629 > \chi^2_{(0.05,14)} = 23.68479$ sehingga dapat disimpulkan minimal ada satu variabel prediktor yang berpengaruh signifikan terhadap frekuensi klaim asuransi kendaraan. Berikutnya, dilanjutkan dengan uji Wald untuk menguji pengaruh masing-masing variabel prediktor secara sekuensial. Uji Wald dilakukan secara terpisah pada model data diskret (parameter $\boldsymbol{\beta}$) pada persamaan (13) dan model *zero inflation* (parameter $\boldsymbol{\gamma}$) pada persamaan (15), dengan statistik uji pada persamaan (23). Hasil uji Wald ditampilkan pada Tabel 5 berikut.

Tabel 5. Hasil uji Wald untuk model ZINB

Parameter	Koefisien	Statistik uji Wald
$\hat{\beta}_0$	-1.18954	21.676
$\hat{\beta}_1$	0.01110	5.574**
$\hat{\beta}_2$	-0.00582	1.473
$\hat{\beta}_3$	-0.00400	0.473
$\hat{\beta}_5$	0.00042	13.059**
$\hat{\beta}_6$	-0.00437	11.456**
$\hat{\beta}_7$	0.00083	9.116**
$\hat{\gamma}_0$	0.86292	3.596
$\hat{\gamma}_1$	-0.00020	169.209**

Model ZINB terdiri atas dua bagian, yaitu model data diskret dan model *zero inflation*. Model data diskret menjelaskan jumlah klaim yang diajukan oleh pemegang polis, sedangkan model *zero inflation* menjelaskan peluang suatu polis termasuk ke dalam kelompok yang tidak pernah mengajukan klaim.

Pada model data diskret, hasil uji Wald menunjukkan bahwa variabel umur pemegang polis (X_1), kapasitas mesin kendaraan (X_5), daya kendaraan (X_6), dan premi (X_7) memiliki nilai statistik uji Wald yang lebih besar dari $\chi^2_{(0.05,1)} = 3.841$. Hal ini menunjukkan bahwa variabel-variabel tersebut berpengaruh signifikan terhadap frekuensi klaim asuransi.

Sementara itu, pada bagian *zero inflation* model, hasil uji Wald menunjukkan bahwa variabel nilai kendaraan $(\hat{\gamma}_1)$memiliki nilai statistik uji Wald sebesar 169.209 yang juga lebih besar dari $\chi^2_{(0.05,1)} = 3.841$. Hal ini menunjukkan bahwa variabel nilai kendaraan berpengaruh signifikan terhadap peluang suatu polis termasuk ke dalam kelompok yang tidak pernah mengajukan klaim sama sekali. Pada model regresi binomial negatif, variabel nilai kendaraan tidak berpengaruh signifikan terhadap frekuensi klaim. Hal ini mengindikasikan bahwa nilai kendaraan tidak berperan dalam menentukan intensitas frekuensi klaim. Namun, pada model regresi ZINB, variabel nilai kendaraan terbukti signifikan pada model *zero inflation*. Hasil menunjukkan bahwa nilai kendaraan lebih berperan dalam menjelaskan peluang suatu polis masuk ke dalam kelompok yang tidak pernah mengajukan klaim. Nilai negatif pada $\hat{\gamma}_1$ mengindikasikan bahwa semakin tinggi nilai kendaraan, semakin kecil kemungkinan pemegang polis untuk tergolong dalam kelompok yang tidak pernah mengajukan klaim. Hal ini bermakna bahwa pemegang polis dengan nilai kendaraan yang lebih tinggi cenderung lebih mungkin untuk mengajukan klaim ketika terjadi kerusakan karena biaya perbaikan dan potensi kerugiannya juga lebih besar dibandingkan kendaraan dengan nilai yang lebih rendah.

Pada model *Zero Inflated Negative Binomial* (ZINB) terdapat parameter dispersi, yakni θ. Parameter ini digunakan untuk menangkap adanya overdispersi, yaitu kondisi ketika varians data lebih besar daripada rata-ratanya. Dalam pemodelan ini, nilai $\hat{\theta}$ yang dihasilkan sebesar 0.3126. Nilai $\hat{\theta}$ tersebut mengindikasikan bahwa data frekuensi klaim asuransi kendaraan mengalami overdispersi sehingga model ZINB cocok untuk digunakan dalam memodelkan frekuensi klaim asuransi kendaraan yang terjadi *zero inflation* dan overdispersi. Diperoleh model ZINB sebagai berikut:

1. Model data diskret

$$\ln(\mu_i) = -1.18954 + 0.01110x_{1i} - 0.00042x_{5i} - 0.00437x_{6i} + 0.00083x_{7i}. \quad (28)$$

2. Model *zero inflation*

$$\ln\left(\frac{p_i}{1-p_i}\right) = -0.00020x_{4i}. \quad (29)$$

dengan $i = 1,2, \ldots 6661$.

3.4 Perbandingan Kinerja Model

Nilai estimasi frekuensi klaim didapat dengan model regresi Binomial Negatif pada persamaan (24) dan dengan model regresi ZINB pada persamaan (28). Hasil estimasi tersebut dibandingkan dengan frekuensi klaim aktual menggunakan rumus MAE pada persamaan (22). Semakin kecil nilai MAE maka nilai estimasi semakin mendekati nilai aktual, artinya model semakin sesuai mengikuti karakteristik data. Estimasi parameter model regresi didapat dari memaksimumkan fungsi *likelihood*, sedangkan rumus AIC pada persamaan (21) menunjukkan negatif dari fungsi *likelihood*, berarti semakin kecil AIC maka model semakin sesuai dengan karakteristik data. Berikut hasil AIC dan MAE untuk kedua model.

Tabel 6. Nilai AIC dan MAE kedua model

Model	Nilai AIC	Nilai MAE
Regresi Binomial Negatif	13686.00	0.9426
Regresi *ZINB*	13651.15	0.9378

Berdasarkan Tabel 6, model regresi binomial negatif memiliki nilai AIC dan MAE yang lebih besar daripada model regresi ZINB yang mengindikasikan bahwa model regresi binomial negatif belum mampu mengatasi data dengan karakteristik adanya *zero inflation*.

Selain membandingkan nilai AIC secara langsung, pemilihan model juga dapat dilakukan menggunakan selisih nilai AIC pada model ke-i terhadap nilai AIC minimum (ΔAIC). Burnham dan Anderson [14] menyatakan jika $\Delta AIC > 10$, maka model dengan nilai AIC minimum adalah model terbaik. Nilai ΔAIC yang dihasilkan antara model regresi binomial negatif dengan ZINB yakni sebesar 34.85 yang menunjukkan bahwa model regresi ZINB merupakan model yang lebih baik. Keunggulan model ZINB ini disebabkan oleh kemampuannya dalam mengatasi dua karakteristik utama data frekuensi klaim asuransi kendaraan, yaitu *zero inflation* dan overdispersi.

Berdasarkan seluruh tahapan yang telah dilakukan, mulai dari pemeriksaan karakteristik data, pemodelan dan evaluasi kinerja model, diketahui bahwa data frekuensi klaim asuransi kendaraan mengandung overdispersi dan *zero inflation*. Model regresi binomial negatif mampu mengatasi overdispersi, tetapi belum bisa menangkap proporsi klaim nol yang tinggi. Selanjutnya, model regresi ZINB menunjukkan kecocokan yang lebih baik terhadap data aktual, baik dalam menangani overdispersi maupun *zero inflation*. Dengan demikian, model ZINB dinilai sebagai model yang relatif lebih sesuai untuk memodelkan frekuensi klaim asuransi kendaraan pada penelitian ini. Model ini

diharapkan dapat menjadi referensi dalam analisis frekuensi klaim serta memberikan kontribusi bagi perusahaan asuransi dalam menetapkan kebijakan yang lebih tepat terkait pengelolaan risiko dan penetapan premi.

4 Simpulan dan Saran

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, hasil pemodelan menunjukkan bahwa model ZINB dapat digunakan untuk menangkap karakteristik data frekuensi klaim yang mengalami overdispersi dan *zero inflation*. Berdasarkan nilai *Akaike Information Criterion* (AIC), *Mean Absolute Error* (MAE), dan nilai ΔAIC, model ZINB merupakan model yang lebih sesuai dalam memodelkan data frekuensi klaim asuransi kendaraan dibandingkan dengan model regresi binomial negatif. Hasil ini menunjukkan bahwa model regresi ZINB relatif lebih baik dalam memprediksi frekuensi klaim asuransi kendaraan.

Berdasarkan hasil uji signifikansi parameter pada model yang lebih sesuai, yaitu ZINB, diperoleh bahwa pada bagian model data diskret, faktor-faktor yang berpengaruh adalah umur pemegang polis, lama pengalaman mengemudi, umur kendaraan, kapasitas mesin kendaraan, daya kendaraan, dan premi, sedangkan pada bagian model *zero inflation*, variabel nilai kendaraan berpengaruh signifikan terhadap pemegang polis tidak mengajukan klaim.

Penelitian selanjutnya dapat mengembangkan pemodelan frekuensi klaim dengan pendekatan *hybrid*, yaitu mengombinasikan *Generalized Linear Model* (GLM) sebagai model distribusi dasar (Poisson atau binomial negatif) dengan metode *machine learning* untuk menangkap pola non-linear dan interaksi kompleks antar peubah penjelas. Pendekatan ini bukan bertujuan membandingkan metode, melainkan meningkatkan performa prediksi sekaligus mempertahankan interpretabilitas model.

Daftar Pustaka

[1] R. Manggala, J.A. J, D. Purwanto, dan A.K Indriastuti, "Studi kasus faktor penyebab kecelakaan lalu lintas pada tikungan tajam," *Jurnal Karya Teknik Sipil*, vol. 4, no. 4, pp. 462–470, 2015. http://ejournal-s1.undip.ac.id/index.php/jkts

[2] H. Wijaya, "Pengaruh service quality, word of mouth, dan brand awareness terhadap keputusan pembelian polis asuransi kendaraan di Jakarta," *Jurnal Manajemen Bisnis dan Kewirausahaan*, vol. 5, no. 5, pp. 518–523, 2020. https://doi.org/10.24912/jmbk.v5i5.13303

[3] J.P. Bouncher, M. Denuit, dan M. Guillen, "Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions," *Variance Actuarial Society*, vol. 2 no. 1, pp. 135–162, 2008. https://doi.org/10.66573/001c.142045

[4] R.N. Amalia, K. Sadik, dan K.A. Notodiputro, "A study of ZIP and ZINB regression modeling for count data with excess zeros," *J Phys Conf Ser*, vol. 1863, no. 1, pp. 1–12, 2021. https://doi.org/10.1088/1742-6596/1863/1/012022

[5] G. Alomair, "Predictive performance of count regression models versus machine learning techniques: a comparative analysis using an automobile insurance claims frequency dataset," *PLoS One*, vol. 19, no. 12, pp. 1–12, 2024. https://doi.org/10.1371/journal.pone.0314975

[6] D.I. Purnama, "Comparison of zero inflated poisson (ZIP) regression, zero inflated negative binomial regression (ZINB) and binomial negative hurdle regression (HNB) to model daily cigarette consumption data for adult population in Indonesia," *J Mat Stat dan Komputasi*., vol. 17, no.3, pp. 357–369, 2021. https://doi.org/10.20956/j.v17i3.12278

[7] A. Nuraeni, S. Martha, dan S. Aprizkiyandari, "Penerapan regresi zero-inflated negative binomial (ZINB) pada data kecelakaan lalu lintas di Kota Pontianak," *Bul Ilm Mat.stat dan Ter*, vol. 11, no. 1, pp. 89–96, 2022. https://doi.org/10.26418/bbimst.v11i1.51606

[8] J.W. Hardin dan J.M. Hilbe. *Generalized Linear Models and Extensions*, Texas: Stata Press, 2007. https://doi.org/10.1007/978-981-96-4726-2
[9] Y. Widyaningsih, G.P. Arum, K. Prawira, "Aplikasi k-fold cross validation dalam penentuan model regresi binomial negatif terbaik," *Barekeng*, vol. 15, no. 2, pp. 315–322. https://doi.org/10.30598/barekengvol15iss2pp315-322
[10] A.C. Cameron dan P.K. Trivedi, *Regression Analysis of Count Data, Second Edition*, Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9781139013567
[11] R.A. Dalimunthe dan I. Husein, "Zero inflated negative binomial regression in malaria cases in North Sumatera," *JISTech (Journal of Islamic Science and Technology)*, vol. 10, no. 1, pp.107–115, 2025. http://dx.doi.org/10.30829/jistech.v10i1.25658
[12] L.E. Afri, "Perbandingan regresi binomial negatif dan regresi Conway-Maxwell Poisson dalam mengatasi overdispersi pada regresi Poisson," *J Gantang*, vol. 2, no. 1, pp. 79–87, 2017. https://doi.org/10.31629/jg.v2i1.66
[13] H. Campbell., "The consequences of checking for zero-inflation and overdispersion in the analysis of count data," *Methods Ecol Evol*, vol. 12, no. 4, pp. 665–680, 2021. https://doi.org/10.1111/2041-210X.13559
[14] K.P. Burnham dan D.R. Anderson, *Model Selection and Multimodel Inference: Practical Information-Theoretic Approach, Second Edition*, New York (NY): Springer-Verlag, 2002. https://doi.org/10.1007/b97636