

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Analisis regresi merupakan bentuk analisis yang digunakan untuk mengetahui hubungan antara variabel prediktor dengan variabel respon. Secara umum, analisis regresi yang sering digunakan adalah regresi linier klasik dengan variabel responnya berupa data kontinu yang mengikuti distribusi normal. Namun, beberapa kasus sering dijumpai variabel respon berupa data diskrit dan tidak berdistribusi normal sehingga regresi linier klasik kurang tepat dalam memodelkan data tersebut. Dalam situasi seperti ini, *Generalized Linear Model* (GLM) dapat digunakan (Sinharay, 2010). GLM merupakan perluasan dari model regresi linier klasik dimana variabel responnya dapat berdistribusi selain normal tetapi termasuk dalam keluarga eksponensial. GLM memiliki tiga komponen yaitu komponen acak, komponen sistematis, dan fungsi penghubung. Salah satu distribusi dalam keluarga eksponensial adalah distribusi *Poisson* yang merupakan distribusi yang menggambarkan peristiwa yang probabilitas kejadiannya kecil yang bergantung pada interval waktu tertentu atau di suatu daerah tertentu (Agresti, 2007).

Pendekatan dengan model regresi *Poisson* merupakan model regresi non-linier yang menggambarkan adanya hubungan antara variabel respon diskrit dan berdistribusi *Poisson* dengan variabel prediktor (Agresti, 2007). Dalam regresi *Poisson*, salah satu asumsi yang harus dipenuhi adalah equidispersi, yaitu nilai rata-rata dan variansi dari variabel respon harus bernilai sama. Namun, dalam analisis data sering dijumpai data diskrit memperlihatkan variansi yang lebih besar dari rata-ratanya (*overdispersi*) atau variansi lebih kecil dari rata-ratanya (*underdispersi*). Kasus *overdispersi* lebih sering terjadi karena adanya keragaman dalam variabel respon serta adanya korelasi positif antar variabel respon yang terjadi pada analisis multivariat (Hilbe, 2011).

Menurut Cahyandari (2014), penyebab lain *overdispersi* yaitu banyaknya pengamatan bernilai nol, sumber keragaman antar individu tidak teramati, terdapat pengamatan yang hilang, adanya pencilan, atau kesalahan spesifikasi fungsi penghubung. *Overdispersi* dapat menyebabkan simpangan baku estimasi parameter *underestimate* dan uji signifikansi parameter *overestimate* sehingga kesimpulan yang diperoleh tidak valid (Ismail & Jemain, 2007). Untuk mengatasi masalah ini, para peneliti telah mengusulkan beberapa model dari perpaduan antara distribusi *Poisson* dengan distribusi kontinu yang disebut distribusi *mixed Poisson*. Distribusi ini digunakan untuk memodelkan data diskrit berdistribusi *Poisson* yang memiliki variansi yang sangat besar (Dean et al., 1989). Salah satu yang termasuk distribusi *mixed Poisson* adalah distribusi *Negative Binomial* yang merupakan campuran distribusi *Poisson* dengan distribusi *Gamma*.

Regresi *Negative Binomial* memiliki karakteristik yang sama dengan regresi *Poisson* hanya saja lebih fleksibel dibandingkan regresi *Poisson* karena model ini memiliki parameter dispersi yang membuat nilai variansi dapat bervariasi menjadi

lebih besar dari rata-rata (Coxe et al., 2009). Kinafa et al. (2020) telah melakukan penelitian dengan membandingkan model *Poisson* dengan *Negative Binomial* pada data kecelakaan lalu lintas di Nigeria. Selain itu, (Ulfa et al., 2021) telah melakukan penelitian yang sama pada data jumlah kasus baru kusta di pulau Jawa. Penelitian-penelitian tersebut menghasilkan kesimpulan bahwa model *Negative Binomial* lebih baik dibandingkan dengan model *Poisson* dalam mengatasi kasus overdispersi. Beberapa kasus dijumpai variabel respon mengandung banyak nilai nol (*excess zeros*) yang dapat menyebabkan overdispersi. Lambert (1992) mengatakan bahwa data dengan kasus *excess zeros* dapat dimodelkan dengan regresi *Zero Inflated Poisson* (ZIP) yang merupakan model campuran untuk data diskrit dengan banyak peristiwa bernilai nol. Penelitian terkait model regresi ZIP telah dilakukan oleh Hashim et al. (2021) pada data "*rainfall hours*" di Hilla dengan hasil bahwa model regresi ZIP lebih baik dibandingkan dengan model *Poisson* dalam memodelkan data diskrit yang mengandung *excess zeros* namun tidak dapat mengatasi masalah overdispersi. Oleh karena itu, untuk memodelkan data diskrit teroverdispersi yang mengandung *excess zeros* digunakan model *Zero Inflated Negative Binomial Regression* (ZINBR).

Secara umum, estimasi parameter model ZINBR menggunakan metode *maximum likelihood estimator* (MLE). Namun, alternatif lain untuk menduga parameter model ZINBR adalah menggunakan metode *Bayesian*. Berdasarkan penelitian Fikhri et al. (2014), metode *Bayesian* lebih konsisten dibandingkan dengan metode MLE dalam mengestimasi parameter pada regresi *Poisson*. Metode *Bayesian* memperhitungkan distribusi awal yang disebut distribusi prior, kemudian dikombinasikan dengan informasi data untuk mengestimasi parameter. Selanjutnya metode *Bayesian* menentukan distribusi posterior dari parameter, tetapi tidak dapat diselesaikan secara analitik dalam bentuk implisit. Oleh karena itu, diperlukan teknik simulasi untuk menyelesaikan estimasi parameter salah satunya dengan memanfaatkan metode *Markov Chain Monte Carlo* (MCMC) dengan algoritma *Gibbs Sampling* (Majumdar & Gries, 2010). *Gibbs sampling* merupakan algoritma dalam metode MCMC yang digunakan untuk pengambilan sampel distribusi kompleks berdimensi tinggi.

Salah satu data diskrit yang diasumsikan berdistribusi *Poisson* adalah data jumlah kematian akibat demam berdarah *dengue* (DBD). Sehubungan dengan tujuan ketiga dari *Sustainable Development Goals* (SDGs) yaitu kehidupan sehat dan sejahtera dengan salah satu indikatornya adalah mengurangi kematian akibat penyakit tidak menular termasuk DBD, maka pemerintah daerah seperti di Pulau Kalimantan diharapkan dapat mengupayakan percepatan pencapaian target tersebut (Bappenas, n.d.). Hal ini perlu pengamatan melalui analisis data untuk mengetahui faktor-faktor yang mempengaruhi kematian akibat DBD. Data jumlah kematian akibat DBD di Pulau Kalimantan tahun 2021 sendiri mengalami *excess zeros*. Oleh karena penulis tertarik untuk memodelkan data tersebut menggunakan ZINBR dengan menganalisis faktor-faktor yang diduga memengaruhi. Penulis menyusun penelitian ini dengan judul "*Pemodelan Zero Inflated Negative Binomial Regression* dengan

Estimasi *Markov Chain Monte Carlo* pada Data Jumlah Kasus Kematian Akibat Demam Berdarah *Dengue* di Pulau Kalimantan”.

## 1.2 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut.

1. Ruang lingkup penelitian adalah semua kabupaten/kota di Pulau Kalimantan.
2. Metode estimasi parameter model *Zero Inflated Negative Binomial Regression* diselesaikan menggunakan metode *Markov Chain Monte Carlo* dengan algoritma *Gibbs Sampling*.

## 1.3 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan penelitian adalah sebagai berikut.

1. Memperoleh estimasi parameter model *Zero Inflated Negative Binomial Regression* pada data jumlah kasus kematian akibat demam berdarah *dengue* di Pulau Kalimantan.
2. Memperoleh faktor-faktor yang mempengaruhi jumlah kasus demam berdarah *dengue* di Pulau Kalimantan menggunakan model *Zero Inflated Negative Binomial Regression*.

## 1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah dapat memperluas pemahaman teoritis dan praktis terkait model *Zero Inflated Negative Binomial Regression* terutama dalam estimasi parameter. Selain itu, penelitian ini dapat menjadi bahan informasi bagi masyarakat maupun instansi pemerintah terutama di Pulau Kalimantan dalam mengevaluasi upaya penurunan jumlah kasus demam berdarah *dengue*.

## 1.5 Teori

### 1.5.1 *Generalized Linear Model*

*Generalized Linear Models* (GLM) merupakan perluasan dari model regresi linier klasik yang mana variabel responnya termasuk salah satu anggota keluarga eksponensial (McCullagh & Nelder, 1989). Menurut (Agresti, 2007) terdapat tiga komponen utama pada GLM, sebagai berikut:

1. Komponen Acak

Komponen acak dari GLM terdiri dari variabel respon  $Y$  dari sebuah distribusi keluarga eksponensial. Komponen acak dari GLM mengidentifikasi variabel respon  $Y$  beserta distribusi peluangnya.

## 2. Komponen Sistematis

Komponen sistematis dari GLM adalah hubungan dari sebuah vektor prediktor linier untuk menjelaskan variabel-variabel yang berhubungan dalam sebuah model linier.

$$\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \beta_0 + \sum_{p=1}^k \beta_p x_{ip} \quad (1)$$

## 3. Fungsi Penghubung (*link function*)

Fungsi penghubung adalah fungsi yang menghubungkan ekspektasi variabel respon  $Y$  dengan variabel-variabel prediktor melalui persamaan linier. Misalkan  $\lambda_i = E(Y_i)$ , fungsi yang menghubungkan  $\lambda_i$  dengan prediktor linier adalah  $g(\cdot)$ . Fungsi penghubung  $g(\cdot)$  dapat dituliskan sebagai berikut.

$$g(\lambda_i) = \beta_0 + \sum_{p=1}^k \beta_p x_{ip} \quad (2)$$

### 1.5.2 Regresi Poisson

Regresi *Poisson* merupakan model regresi non-linear dengan variabel responnya berupa data hitung (diskrit) dan berdistribusi *Poisson*. Distribusi *Poisson* adalah suatu distribusi untuk peristiwa yang probabilitas kejadian kecil, dimana kejadian tersebut bergantung pada interval waktu atau di suatu daerah tertentu. Jika variabel acak diskrit ( $Y$ ) berdistribusi *Poisson* dengan parameter  $\lambda_i > 0$ , yang merupakan rata-rata suatu kejadian (Arkandi & Winahju, 2015). Oleh karena itu, maka fungsi massa peluang dinyatakan sebagai berikut (Walpole et al., 2011).

$$f(y_i) = e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!}; y_i = 0, 1, 2, \dots \quad (3)$$

Distribusi *Poisson* memiliki nilai mean dan variansi yang sama, yaitu  $E(Y) = Var(Y) = \lambda_i$ .

Salah satu metode pengujian kecocokan antara data dan distribusi yaitu *Kolmogorov Smirnov*. Menurut (Widarjono, 2010), uji *Kolmogorov Smirnov* digunakan untuk mengetahui apakah sampel berasal dari populasi dengan distribusi tertentu yakni distribusi *Poisson*. Uji *Kolmogorov Smirnov* ini didasarkan kecocokan antara distribusi kumulatif yang diamati (fungsi distribusi sampel) dan distribusi kumulatif dengan hipotesis dan statistik uji sebagai berikut.

$H_0 : F_1(y) = F_0(y)$  (variabel  $Y$  berdistribusi *Poisson*)

$H_1 : F_1(y) \neq F_0(y)$  (variabel  $Y$  tidak berdistribusi *Poisson*)

$$D = \sup |F_1(y) - F_0(y)| \quad (4)$$

dengan  $F_1(y)$  menyatakan fungsi distribusi sampel atau proporsi nilai-nilai pengamatan dalam sampel dan  $F_0(y)$  merupakan fungsi distribusi kumulatif. Kriteria pengambilan keputusan dalam uji *Kolmogorov Smirnov* adalah menolak  $H_0$  jika nilai  $D > D_{56(1-\frac{0.05}{2})}$  pada taraf signifikansi  $\alpha$  (Daniel, 1989).

Regresi *Poisson* menggunakan GLM agar modelnya dapat digunakan dalam data pengamatan, dimana variabel responnya tidak mengharuskan berdistribusi normal. Regresi *Poisson* biasanya menggunakan fungsi penghubung logaritma natural, karena rata-rata dari variabel responnya akan berbentuk fungsi eksponensial dan menjamin nilai variabel yang estimasinya dari variabel responnya akan bernilai non-negatif. Sehingga model regresi *Poisson* sebagai berikut (Cahyandari, 2014).

$$\begin{aligned} g(\lambda_i) &= \ln(\lambda_i) \\ \ln(\lambda_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \\ \lambda_i &= \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \end{aligned} \quad (5)$$

dengan

$\lambda_i$  menunjukkan rata-rata dari variabel respon  $Y$  yang berdistribusi *Poisson*.

$\mathbf{x}_i^T = [1 \ x_{i1} \ \dots \ x_{ip}]$  menunjukkan vektor yang berukuran  $1 \times (p + 1)$  dari variabel prediktor yang digunakan untuk memodelkan  $\lambda_i$ .

$\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$  menunjukkan parameter regresi yang berukuran  $(p + 1) \times 1$ .

### 1.5.3 Overdispersi

Salah satu asumsi yang harus dipenuhi dalam regresi *Poisson* adalah asumsi equidispersi, yaitu keadaan dengan *mean* dan variansi dari variabel responnya bernilai sama. Namun dalam aplikasinya asumsi tersebut kadang dilanggar, dimana nilai variansinya lebih besar daripada nilai mean yang disebut overdispersi atau nilai variansinya lebih kecil daripada nilai mean yang disebut underdispersi (Cameron & Trivedi, 1998). Overdispersi pada regresi *Poisson* menghasilkan simpangan baku dari estimasi parameter jauh lebih kecil daripada nilai sebenarnya (*underestimate*) dan uji signifikansi dari variabel prediktor jauh lebih besar daripada nilai sebenarnya (*overestimate*), sehingga kesimpulan yang dihasilkan menjadi tidak valid (Ismail & Jemain, 2007).

Dalam beberapa kasus, yang ditemukan adalah variansi data yang diamati lebih besar daripada *mean*-nya atau disebut *overdispersi*. Ada beberapa hal yang dapat menyebabkan terjadinya *overdispersi*, antara lain: banyaknya pengamatan yang bernilai nol (*excess zero*), adanya sumber keragaman antar individu yang tidak teramati (*unobserved heterogeneity*), adanya pengamatan yang hilang (*data missing*), adanya pencilan (*outlier*) pada data sehingga perlunya interaksi dalam model, atau kesalahan spesifikasi fungsi penghubung (Cahyandari, 2014).

Overdispersi dapat ditentukan dengan menggunakan nilai *deviance* maupun nilai *pearson chi-square* dengan hipotesis dan statistik uji sebagai berikut (McCullagh & Nelder, 1989).

$H_0: \phi = 1$  (variabel respon  $Y$  tidak mengalami overdispersi).

$H_1: \phi > 1$  (variabel respon  $Y$  mengalami overdispersi).

$$\phi_D = \frac{D(y; \lambda_i)}{n - k - 1} = \frac{\left[ 2 \sum_{i=1}^n \left( y_i \ln \left( \frac{y_i}{\lambda_i} \right) - (y_i - \lambda_i) \right) \right]}{n - k - 1} \quad (6)$$

$$\phi_{\chi^2} = \frac{\chi^2}{n - k - 1} = \frac{\sum_{i=1}^n \frac{(y_i - \lambda_i)^2}{\lambda_i}}{n - k - 1} \quad (7)$$

Kriteria pengambilan keputusan dalam uji overdispersi adalah menolak  $H_0$  jika nilai  $\hat{\phi} > 1$ .

#### 1.5.4 Regresi *Negative Binomial*

Regresi *Negative Binomial* adalah salah satu solusi untuk mengatasi masalah overdispersi berdasarkan model campuran *Poisson-Gamma*. Distribusi *Negative Binomial* memiliki fungsi kepadatan peluang sebagai berikut.

$$f(y_i; \lambda_i, \tau) = \frac{\Gamma(y_i + \tau)}{\Gamma(\tau)\Gamma(y_i + 1)} \left(\frac{\tau}{\tau + \lambda_i}\right)^\tau \left(\frac{\lambda_i}{\tau + \lambda_i}\right)^{y_i}; y_i = 0, 1, 2, \dots; \tau \geq 0 \quad (8)$$

dengan  $\Gamma(\cdot)$  adalah fungsi gamma dan  $\tau$  adalah parameter dispersi. Nilai *mean* dan variansi dari distribusi *Negative Binomial* adalah  $E(y_i) = \lambda_i$  dan  $Var(y_i) = \lambda_i + \tau\lambda_i^2$ .

Distribusi *Negative Binomial* bukan distribusi yang dapat dimodifikasi hanya dengan mengubah posisi, skala, atau kombinasi keduanya. Bentuk distribusi *Negative Binomial* tidak mengikuti pola distribusi yang dapat diubah lokasinya atau diubah skalanya seperti beberapa distribusi lainnya. Distribusi *Negative Binomial* dapat diuji menggunakan uji *Kolmogorov Smirnov* dengan hipotesis dan statistik uji sebagai berikut (Campbell & Oprian, 1979).

$H_0 : F_1(y) = P(y; \hat{\lambda})$  (variabel  $Y$  berdistribusi *Negative Binomial*)

$H_1 : F_1(y) \neq P(y; \hat{\lambda})$  (variabel  $Y$  tidak berdistribusi *Negative Binomial*)

$$D_{co} = \sup|\hat{F}_1(y) - P(y; \hat{\lambda})| \quad (9)$$

dengan  $P(y; \hat{\lambda})$  menyatakan fungsi distribusi kumulatif untuk distribusi *Poisson*( $\lambda$ ), dan menolak  $H_0$  jika  $D_{co} > D_{56(1-\frac{0.05}{2})}$  pada taraf signifikansi  $\alpha$ .

Model regresi *Negative Binomial* dapat dituliskan seperti berikut (Afri, 2013).

$$\begin{aligned} \ln(\lambda_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \\ \lambda_i &= \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \end{aligned} \quad (10)$$

dengan

$\lambda_i$  menunjukkan rata-rata dari variabel respon  $Y$  yang berdistribusi *Negative Binomial*.

$\mathbf{x}_i^T = [1 \ x_{i1} \ \dots \ x_{ip}]$  menunjukkan vektor yang berukuran  $1 \times (p + 1)$  dari variabel prediktor yang digunakan untuk memodelkan  $\lambda_i$ .

$\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$  menunjukkan parameter regresi yang berukuran  $(p + 1) \times 1$ .

#### 1.5.5 Excess Zeros

Salah satu permasalahan pada regresi *Poisson* adalah nilai nol yang berlebih yang dapat dilihat pada proporsi variabel respon yang bernilai nol lebih besar dari data diskrit lainnya. Menurut Winkelmann (2008), banyaknya proporsi data yang bernilai nol dari proporsi data lainnya ( $> 50\%$  atau  $\lambda < 0,69$ ) dapat berakibat pada ketepatan

(presisi) dalam pengambilan keputusan. Selain itu, regresi *Poisson* juga menjadi tidak tepat digunakan karena *excess zeros* merupakan salah satu penyebab terjadinya *overdispersi* (Hinde & Demétrio, 1998).

### 1.5.6 Zero Inflated Poisson Regression

Salah satu penyebab terjadinya *overdispersi* adalah lebih banyak pengamatan bernilai nol daripada yang diharapkan untuk model regresi *Poisson*. Salah satu metode analisis yang diusulkan untuk lebih banyak pengamatan bernilai nol adalah model *Zero Inflated Poisson Regression* (ZIPR) (Jansakul & Hinde, 2002).

Jika  $Y$  adalah variabel acak yang berdistribusi *Zero Inflated Poisson* (ZIP), maka nilai nol pada pengamatan diestimasikan muncul dalam dua cara yang sesuai dengan keadaan (*state*) yang terpisah. Keadaan pertama disebut *zero state* terjadi dengan probabilitas  $\pi$  dan menghasilkan hanya pengamatan yang bernilai nol. Sedangkan keadaan kedua disebut *Poisson state* terjadi dengan probabilitas  $(1 - \pi_i)$  dan berdistribusi *Poisson* dengan *mean*  $\lambda_i$  (Jansakul & Hinde, 2002).

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

$$f_{ZIP}(y_i) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\lambda_i}; & y_i = 0, 0 \leq \pi_i \leq 1, \lambda_i \geq 0 \\ \frac{(1 - \pi_i)e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}; & y_i > 0, 0 \leq \pi_i \leq 1, \lambda_i \geq 0 \end{cases} \quad (11)$$

Nilai ekpektasi dan variansi dari  $Y$  adalah  $E(Y) = \lambda_i(1 - \pi_i)$  dan  $Var(Y) = \lambda_i(1 - \pi_i)[1 + \lambda_i\pi_i]$ .

Lambert (1992) menyarankan model gabungan untuk  $\lambda_i$  dan  $\pi_i$ , yaitu sebagai berikut.

$$\begin{aligned} \text{Poisson state} & : & \ln(\lambda_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \\ & & \lambda_i &= \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \end{aligned} \quad (12)$$

$$\begin{aligned} \text{Zero state} & : & \text{logit}(\pi_i) &= \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \boldsymbol{\gamma} \\ & & \pi_i &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\gamma})} \end{aligned} \quad (13)$$

dengan

$\lambda_i$  menunjukkan rata-rata dari variabel respon  $Y$  yang berdistribusi *Poisson*.

$\pi_i$  menunjukkan probabilitas pengamatan ke- $i$

$\mathbf{x}_i^T = [1 \ x_{i1} \ \dots \ x_{ip}]$  menunjukkan vektor berukuran  $1 \times (p + 1)$  dari variabel prediktor yang digunakan untuk memodelkan  $\lambda_i$ .

$\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$  menunjukkan vektor yang berukuran  $(p + 1) \times 1$  dari parameter regresi model *Poisson state* yang tidak diketahui.

$\boldsymbol{\gamma} = [\gamma_0 \ \gamma_1 \ \dots \ \gamma_p]^T$  menunjukkan vektor yang berukuran  $(p + 1) \times 1$ . dari parameter regresi model *zero state* yang tidak diketahui.

### 1.5.7 Zero Inflated Negative Binomial Regression

Menurut Mwalili et al. (2008), distribusi *Zero Inflated Negative Binomial* (ZINB) merupakan hasil campuran antara distribusi Binomial Negatif dan distribusi pembentuk dari nilai nol. Model *Zero Inflated Negative Binomial Regression* (ZINBR) dapat digunakan untuk memodelkan data diskrit yang banyak memiliki nilai nol (*excess zeros*) pada peubah respon dan mengalami overdispersi (Garay et al., 2015).

Pada model ZINBR, peubah respon  $Y_i$  merupakan peubah acak bebas dengan  $i = 1, 2, 3, \dots, n$  yang mengikuti distribusi  $Y_i \sim \text{ZINB}(\lambda_i, \pi_i, \tau)$  dan dapat membentuk dalam dua keadaan yaitu *zero state* dan *Negative Binomial state*. Dengan demikian, model ZINBR terdiri dari model *excess zeros* dengan parameter  $\lambda_i$  dan  $r$  yang merupakan parameter dari distribusi *Negative Binomial*.

Model *excess zeros* terjadi dengan peluang  $\pi_i$  dan model *Negative Binomial* dengan peluang  $(1 - \pi_i)$ . Fungsi peluang peubah acak  $Y_i$  adalah sebagai berikut (Garay et al., 2015).

$$f_{\text{ZINB}}(y_i) = \begin{cases} \pi_i + (1 - \pi_i) \left( \frac{\tau}{\tau + \lambda_i} \right)^\tau & ; y_i = 0 \\ (1 - \pi_i) \frac{\Gamma(\tau + y_i)}{\Gamma(y_i + 1)\Gamma(\tau)} \left( \frac{\lambda_i}{\tau + \lambda_i} \right)^{y_i} \left( \frac{\tau}{\tau + \lambda_i} \right)^\tau & ; y_i > 0 \end{cases} \quad (14)$$

dengan  $0 \leq \pi_i \leq 1, \lambda_i \geq 0$  dan  $\tau$  merupakan parameter dispersi dengan  $\frac{1}{\tau} > 0$  dan  $\Gamma(\cdot)$  adalah fungsi gamma. Nilai *mean* dan variansi dari variabel respon  $Y_i$  yang membentuk model ZINBR dengan parameter  $\pi_i, \lambda_i$  dan  $\tau$  adalah  $E(Y_i) = (1 - \pi_i)\lambda_i$  dan  $\text{Var}(Y_i) = (1 - \pi_i)\lambda_i \left( 1 + \lambda_i \frac{1}{\tau} + \pi_i \lambda_i \right)$ . Ketika  $\pi_i = 0$ , peubah acak  $Y_i$  akan membentuk distribusi Binomial Negatif dengan rata-rata  $\lambda_i$  dan parameter dispersi  $\tau$ .

Pada GLM, fungsi penghubung logit digunakan ketika parameter model regresi bernilai diantara 0 dan 1 dan fungsi penghubung log digunakan jika parameter diharapkan bernilai positif. Oleh karena itu, fungsi penghubung yang tepat digunakan pada model ZINBR adalah fungsi penghubung logit untuk parameter  $\pi_i$  dan fungsi penghubung log untuk parameter  $\lambda_i$ . Menurut (Garay et al., 2011), pada praktiknya parameter  $\pi_i$  dan  $\lambda_i$  pada model ZINBR bergantung pada vektor peubah prediktor  $x_i$ , sehingga ZINBR terdiri dari dua model yaitu sebagai berikut.

$$\begin{aligned} \text{Negative Binomial state} : \quad & \ln(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta} \\ & \lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \end{aligned} \quad (15)$$

$$\begin{aligned} \text{Zero state} : \quad & \text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \boldsymbol{\gamma} \\ & \pi_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\gamma})} \end{aligned} \quad (16)$$

dengan

$\lambda_i$  menunjukkan rata-rata dari variabel respon  $Y$  yang berdistribusi *Negative Binomial*.  
 $\pi_i$  menunjukkan probabilitas pengamatan ke- $i$



$x_i^T = [1 \ x_{i1} \ \dots \ x_{ip}]$  menunjukkan vektor berukuran  $1 \times (p + 1)$  dari variabel prediktor yang digunakan untuk memodelkan  $\lambda_i$ .

$\beta = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$  menunjukkan vektor yang berukuran  $(p + 1) \times 1$  dari parameter regresi model *Negative Binomial state* yang tidak diketahui.

$\gamma = [\gamma_0 \ \gamma_1 \ \dots \ \gamma_p]^T$  menunjukkan vektor yang berukuran  $(p + 1) \times 1$  dari parameter regresi model *zero state* yang tidak diketahui.

### 1.5.8 Metode Bayesian

Metode Bayes didasarkan pada dua sumber informasi dalam melakukan pendugaan parameter untuk suatu model statistik. Sumber informasi pertama berasal dari data sampel yang diperoleh dari suatu fungsi *likelihood* dan sumber informasi kedua berasal dari opini subjektif ahli atau berasal dari penelitian terdahulu yang disebut dengan informasi prior. Berbagai ahli dapat memberikan informasi prior berbeda. Ketidaktentuan opini ahli ini diekspresikan melalui suatu distribusi peluang yang disebut distribusi prior (Bolstad, 2007). Pada metode ini untuk menduga parameter dilakukan pendekatan dengan memperlakukan semua parameter yang tidak diketahui sebagai peubah acak dan memiliki distribusi (Bolstad, 2007). Dari distribusi posterior diperoleh estimator bayesian yang merupakan *mean* dari distribusi posterior. Hubungan distribusi posterior dengan dsitribusi prior dan fungsi *likelihood* yaitu (Ntzaufas, 2009) :

$$\text{Distribusi posterior} \propto \text{likelihood} \times \text{distribusi prior}$$

Notasi  $\propto$  menyatakan suatu hubungan proporsionalitas dari disribusi posterior.

### 1.5.9 Distribusi Prior

Analisis bayesian pada suatu populasi mengikuti distribusi tertentu dengan suatu parameter didalamnya, parameter tersebut dapat mengikuti distribusi peluang tertentu yang disebut distribusi prior (Bain & Engelhardt, 1992). Distribusi prior merupakan distribusi awal yang memberikan informasi mengenai parameter. Distribusi prior mencerminkan kepercayaan subjektif parameter sebelum sampel diambil (Wijaya & Wulandari, 2016). Distribusi posterior akan bergantung pada pemilihan distribusi prior. Dengan demikian, permasalahan utama adalah memilih distribusi prior untuk parameter yang tidak diketahui namun relevan dengan permasalahan data penelitian. Pemilihan distribusi prior dapat didasarkan pada ruang parameternya (Shobri dkk., 2021). Pada dasarnya, distribusi prior adalah representasi subjektif peneliti terhadap suatu nilai parameter yang diduga.

1. Distribusi prior yang berkaitan dengan bentuk distribusi hasil identifikasi pola datanya dikelompokkan menjadi:

- a) Distribusi konjugat  
Distribusi prior konjugat mengacu pada acuan analisis model terutama dalam pembentukan fungsi *likelihood* sehingga dalam penentuan prior konjugat selalu dipikirkan mengenai pembentukan pola distribusi prior yang mempunyai fungsi kepekatan peluang pembangun *likelihood*.
  - b) Distribusi tidak konjugat  
Apabila pemberian distribusi prior pada suatu model tidak mempertimbangkan pola pembentuk fungsi *likelihood*.
2. Berdasarkan penentu parameter pada pola distribusinya, distribusi prior dikelompokkan menjadi:
- a) Distribusi informatif  
Distribusi prior informatif mengacu pada pemberian parameter dari distribusi prior yang telah dipilih, baik distribusi prior konjugat atau tidak konjugat. Pemberian nilai parameter pada distribusi prior ini akan sangat mempengaruhi bentuk distribusi posterior yang akan didapat pada informasi data yang diperoleh.
  - b) Distribusi non-informatif  
Prior non-informatif, apabila pemilihan distribusi priornya tidak didasarkan pada informasi yang ada sebelumnya. Apabila pengetahuan tentang priornya sangat lemah, maka bisa digunakan prior berdistribusi normal dengan mean nol dan varian besar. Efek dari penggunaan prior dengan mean nol adalah estimasi parameternya dihaluskan menuju nol. Tetapi, karena pemulusan ini dilakukan oleh varian, maka pemulusan tersebut bisa diturunkan dengan meningkatkan varian (Galindo-Garre dan Vermunt, 2004).

### 1.5.10 Distribusi Posterior

Distribusi posterior merupakan distribusi yang menyatukan informasi data sampel dengan informasi distribusi prior dengan teorema bayesian. Distribusi dinyatakan sebagai fungsi kepadatan peluang bersyarat dari  $\theta$  dengan pengamatan  $Y$ , dengan  $Y = y_1, y_2, \dots, y_n$  (Shobri dkk., 2021). Dalam estimasi bayesian, setelah informasi tentang sampel diperoleh dan prior dapat ditentukan maka distribusi posterior dapat dicari dengan cara mengalikan prior dengan informasi sampel yang diperoleh dari *likelihood* (Bolstad & Curran, 2007).

Box dan Tiao (1973) mendefinisikan bahwa distribusi posterior  $f(\theta_j|y)$  adalah fungsi kepadatan bersyarat parameter  $\beta$  jika diketahui nilai observasi  $y$ , dapat ditulis pada Persamaan (8).

$$f(\theta_j|y) = \frac{f(y, \theta_j)}{f(y)} \quad (8)$$

dengan  $f(y, \theta_j)$  adalah fungsi kepadatan bersama dari  $y$  dan  $\theta$  sedangkan  $f(y)$  merupakan distribusi marginal  $y$ . Fungsi kepadatan bersama  $f(y, \theta_j)$  merupakan perkalian dua fungsi kepadatan yaitu distribusi prior ( $f(\theta_j)$ ) dan distribusi data ( $f(y|\theta_j)$ ), yang ditulis pada Persamaan (9).

$$f(y, \theta_j) = f(y|\theta_j) f(\theta_j) \quad (9)$$

Sedangkan distribusi marginal  $y$  dapat dihitung dengan

$$f(y) = \begin{cases} \int f(y, \theta_j) d\theta = \int f(y|\theta_j) f(\theta_j) d\theta, & \theta \text{ kontinu} \\ \sum_{\theta} f(y, \theta_j) d\theta = \sum_{\theta} f(y|\theta_j) f(\theta_j), & \theta \text{ diskrit} \end{cases}$$

Distribusi parameter  $\theta$  yaitu  $f(\theta_j)$  disebut sebagai prior dan  $f(y|\theta_j)$  sebagai *likelihood* yang merupakan fungsi parameter dari  $\theta$ . Karena  $f(y)$  tidak bergantung pada  $\theta$ , maka  $\frac{1}{f(y)}$  dapat dianggap konstan, misalkan  $C$ . dengan kata lain,  $f(y)$  adalah konstanta yang disebut *normalized constant*. Sehingga posterior dapat ditulis pada Persamaan (10).

$$f(\theta_j|y) = L(y|\theta_j) f(\theta_j) \quad (10)$$

Persamaan (10) menunjukkan bahwa posterior adalah proporsional terhadap *likelihood* dikalikan dengan prior dari parameter model. Nilai tengah dari distribusi posterior yang akan digunakan untuk menentukan estimator dari parameter yang tidak diketahui. Fungsi kepadatan peluang dari  $\theta$  jika diketahui contoh pengamatan  $Y = y_1, y_2, \dots, y_n$  yang terdapat pada Persamaan (11).

$$f(\theta_j|Y, X) = \frac{L(Y, X; \theta) \times f(\theta_j)}{\int_{-\infty}^{\infty} L(Y, X; \theta) \times f(\theta_j) d\theta_j} \quad (11)$$

Adapun distribusi posterior dalam penyelesaiannya tidak dapat diselesaikan secara analitis sehingga dibutuhkan simulasi yaitu *markov chain monte carlo* yang memungkinkan pengambilan sampel numerik dari distribusi posterior yang mendasarinya dan digunakan untuk memperbarui parameter (Sumae dkk., 2022).

### 1.5.11 Markov Chain Monte Carlo

*Markov Chain Monte Carlo* (MCMC) adalah suatu metode simulasi yang membangkitkan sejumlah sampel dari distribusi tertentu yang cara kerjanya menggunakan sifat rantai *Markov*. Distribusi posterior yang rumit pada suatu metode *Bayesian* dapat diperoleh secara akurat menggunakan metode ini (Ntzoufras, 2009). Rantai *Markov* merupakan proses stokastik  $\{\theta^1, \theta^2, \dots, \theta^M\}$  sehingga dapat dinyatakan dengan persamaan berikut.

$$f(\theta^{m+1} | \theta^m \dots \theta^1) = f(\theta^{m+1} | \theta^m) \quad (17)$$

Menurut Ntzoufras (2009), tahapan membangkitkan  $f(\theta | y)$  berdasarkan metode MCMC adalah sebagai berikut:

1. Menentukan nilai awal  $\theta^0$ .
2. Membangkitkan  $M$  sampel hingga mencapai distribusi yang stasioner.
3. Memeriksa kekonvergenan dari algoritma dengan menggunakan uji diagnostik. Jika tidak konvergen, maka perlu menambah jumlah sampel.
4. Membuang  $B$  sampel pertama
5. Menganggap  $\{\theta^{B+1}, \theta^{B+2}, \dots, \theta^M\}$  sebagai sampel untuk analisis posterior di mana  $B$  adalah iterasi awal (*burn in*).
6. Membuat plot distribusi posterior.

7. Mendapatkan ringkasan dari distribusi posterior (*mean*, median, simpangan baku, *quantile*, korelasi).

*Output* dari proses MCMC akan membentuk sampel acak untuk iterasi pertama sampai iterasi  $M'$  (jumlah sampel yang dibangkitkan dipotong banyak sampel *burn in B*). Menurut Ntzoufras (2009), sampel acak yang dihasilkan adalah  $\theta^1, \theta^2, \dots, \theta^m, \dots, \theta^{M'}$ . Dari sampel acak tersebut, untuk setiap fungsi  $f(\theta; y)$  dari parameter  $\theta$  dapat diperoleh.

1. Sampel dari parameter yang diinginkan yaitu dengan mempertimbangkan.

$$f(\theta^{(1)}; y), f(\theta^{(2)}; y), \dots, f(\theta^{(t)}; y), \dots, f(\theta^{(T')}; y)$$

2. Ringkasan posterior  $f(\theta; y)$  dari sampel dengan pendugaan sampel sederhana. *Mean* dan standar deviasi dari posterior adalah sebagai berikut.

$$\hat{E}(f(\theta; y)) = \overline{f(\theta; y)} = \frac{1}{T'} \sum_{t=1}^{T'} f(\theta^t; y) \quad (18)$$

$$\widehat{SD}(f(\theta; y)) = \frac{1}{T' - 1} \sum_{t=1}^{T'} [f(\theta^t - \hat{E}(f(\theta; y)))]^2 \quad (19)$$

Skala pengukuran yang lain adalah median dan persentil (2.5% dan 97.5% sebagai *credible interval*).

3. Perhitungan dan pemeriksaan autokorelasi parameter. Autokorelasi dapat menggambarkan hubungan antara nilai parameter yang berurutan (hubungan antara parameter dengan dirinya sendiri pada *lag s*). Jika terdapat hubungan yang kuat maka hasil simulasi belum mencapai distribusi target. Menurut Hoff (2009), autokorelasi dapat diduga dengan rumus berikut.

$$acf_s(\theta) = \frac{\frac{1}{T-s} \sum_{t=1}^{T-1} (\theta^t - \bar{\theta})(\theta^{t+s} - \bar{\theta})}{\frac{1}{T-s} \sum_{t=1}^{T-1} (\theta^t - \bar{\theta})^2} \quad (20)$$

4. Plot dari distribusi marginal posterior adalah berupa histogram, plot kepadatan peluang, *boxplot*, dll. *Trace plot* pada MCMC adalah plot antara iterasi dengan nilai hasil bangkitan. Ketika *trace plot* menunjukkan pola tertentu atau periodisitas yang kuat maka diperlukan penambahan banyaknya iterasi.

Pada metode simulasi MCMC dua algoritma yang populer yaitu *Metropolis Hasting* dan *Gibbs Sampling*. Meskipun *Gibbs Sampling* merupakan kasus khusus dari algoritma *Metropolis Hasting*, namun algoritma ini sering digunakan karena terkenal dengan kemudahannya.

### 1.5.12 Algoritma Gibbs Sampling

Algoritma *Gibbs Sampling* menggunakan sampel sebelumnya untuk membangkitkan nilai sampel berikutnya secara acak dan menggunakan distribusi bersyarat penuh yang kemudian dihubungkan dengan distribusi posterior. Distribusi bersyarat tersebut memiliki bentuk yang telah dikenal, sehingga sejumlah nilai acak akan dengan mudah disimulasikan. Menurut Ntzoufras (2009), algoritma *Gibbs Sampling* selalu bergerak ke nilai-nilai baru tanpa memperhatikan spesifikasi dari distribusi-

distribusi yang diajukan, tetapi algoritma ini akan tidak efektif ketika terdapat korelasi yang tinggi antar parameter. Proses simulasi pada metode MCMC dengan menggunakan algoritma *Gibbs Sampling* dapat dilakukan dengan tahapan sebagai berikut.

1. Menentukan nilai awal dari  $\theta^{(0)} = \theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)}$  di mana  $\theta^{(0)}$  merupakan sembarang nilai sesuai dengan batas ketentuan masing-masing distribusi. Nilai ini digunakan untuk menduga masing-masing nilai yang lain.
2. Untuk  $t = 1, \dots, T$  Mengulangi langkah berikut.
  - a. Menentukan  $\theta = \theta^{t-1}$
  - b. Untuk  $p = 1, \dots, k$ , perbarui  $\theta_p = f(\theta_p; \theta_p, y)$
  - c. Menentukan  $\theta^t = \theta$  dan menyimpannya sebagai himpunan nilai-nilai yang dibangkitkan pada iterasi ke  $t + 1$  dari algoritma.
3. Berikut adalah proses simulasi menentukan parameter baru.
  - $\theta_1^t$  dari  $f(\theta_1; \theta_2^{t-1}, \theta_3^{t-1}, \dots, \theta_p^{t-1}, y)$
  - $\theta_2^t$  dari  $f(\theta_2; \theta_1^t, \theta_3^{t-1}, \dots, \theta_p^{t-1}, y)$
  - $\theta_3^t$  dari  $f(\theta_3; \theta_1^t, \theta_2^t, \dots, \theta_p^{t-1}, y)$
  - ⋮
  - $\theta_j^t$  dari  $f(\theta_j; \theta_1^t, \theta_2^t, \dots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \dots, \theta_p^{t-1}, y)$
  - ⋮
  - $\theta_p^t$  dari  $f(\theta_p; \theta_1^t, \theta_2^t, \dots, \theta_{p-1}^t, y)$
4. Amati konvergenitas data sampel. Jika kondisi konvergen belum tercapai, maka dibutuhkan sampel lebih banyak lagi dan ulangi langkah 2 sampai kondisi konvergen.
5. Melakukan proses *burn in* dengan membuang sebanyak C sampel pertama. Proses *burn in* adalah periode awal iterasi estimasi parameter dalam proses MCMC membuang sebanyak C iterasi pertama, guna menghilangkan pengaruh dari penggunaan nilai awal. Periode *burn in* dilakukan sampai kondisi stasioner tercapai. Dari proses MCMC akan diperoleh sampel sebanyak  $D = M - C$  untuk setiap parameter.
6. Menggunakan  $\{\theta^1, \theta^2, \dots, \theta^{(m)}\}$  sebagai sampel dalam analisis *posterior*.
7. Membuat plot distribusi *posterior*.
8. Membuat *summary* dari distribusi *posterior* (*mean*, standar deviasi, dll).

Prosedur diatas dinyatakan sebagai sifat *Markov* dan rangkaiannya dapat disebut juga sebagai rantai *Markov* (Hoff, 2009). Pada metode *Bayesian* nilai parameter ditunjukkan dengan nilai rata-rata hasil simulasi distribusi *posterior*. Proses simulasi pada *Gibbs Sampling* dilakukan sampai tercapai keadaan konvergen.

### 1.5.13 Pemeriksaan Konvergensi

Keadaan konvergensi pada MCMC adalah keadaan di mana algoritma telah mencapai keseimbangan dan sampel yang dibangkitkan telah sesuai dengan distribusi target. Metode yang sering digunakan sebagai kriteria pengujian konvergensi MCMC adalah:

### 1. Monte Carlo Error (MC Error)

*MC error* adalah sebuah skala pengukuran yang menggunakan variabilitas dari setiap pendugaan ketika simulasi. *MC error* harus bernilai kecil untuk menghitung parameter yang diinginkan dengan dengan peningkatan presisi. Salah satu cara menduga *MC error* adalah metode *batch mean*, sampel hasil *output* dibagi dalam  $k$  *batch*. Nilai dari  $K$  dan ukuran sampel masing-masing *batch*  $v = T'/K$  harus cukup besar untuk memperkirakan ragam secara konsisten dan juga menghilangkan efek autokorelasi.

Menurut (Ntzoufras, 2009), perhitungan rata-rata sampel setiap *batch* dan rata-rata sampel keseluruhan berturut-turut adalah sebagai berikut.

$$\overline{G(\theta)_b} = \frac{1}{v} \sum_{t=(b-1)v+1}^{bv} G(\theta^t) \quad (21)$$

$$\overline{G(\theta)} = \frac{1}{T'} \sum_{t=1}^{T'} G(\theta^t) = \frac{1}{K} \sum_{b=1}^K \overline{G(\theta)_b} \quad (22)$$

dengan  $b = 1, \dots, K$ . Sedangkan menduga *MC error* adalah dengan cara sebagai berikut.

$$\begin{aligned} MCE[G(\theta)] &= \widehat{SE}[\overline{G(\theta)}] \\ &= \sqrt{\frac{1}{K} \widehat{SD}[G(\theta)_b]} \\ &= \sqrt{\frac{1}{K(K-1)} \sum_{b=1}^K (\overline{G(\theta)_b} - \overline{G(\theta)})^2} \end{aligned} \quad (23)$$

dengan

$K$  : jumlah *batch*

$v$  : jumlah sampel setiap *batch*

$\overline{G(\theta)}$  : rata-rata sampel umum

$\overline{G(\theta)_b}$  : rata-rata setiap *batch*

$b$  : indeks untuk jumlah *batch*

$T'$  : jumlah sampel yang dibangkitkan setelah dipotong dengan sampel *burn in*

Jika *MC error* kurang 1% simpangan baku dari parameter maka dapat dinyatakan bahwa algoritma telah konvergen.

### 2. Trace Plot

*Trace Plot* pada MCMC adalah plot antara iterasi dengan nilai hasil bangkitan. Ketika *trace plot* menunjukkan pola tertentu atau periodisitas yang kuat maka dapat dinyatakan algoritma tidak konvergen sehingga diperlukan penambahan banyak iterasi.

### 3. Autokorelasi

Pada plot autokorelasi dinyatakan konvergen jika *lag* pertama mendekati satu dan *lag-lag* selanjutnya terus menurun menuju nol dengan perhitungan autokorelasi sesuai dengan persamaan. Pada periode bayesian nilai parameter ditunjukkan dengan nilai rata-rata hasil simulasi distribusi posterior.

#### 1.5.14 Uji Signifikansi

Menurut (Ntzoufras, 2009), pengujian parameter metode *Bayesian* dapat menggunakan *credible interval* yang ditunjukkan dengan batas bawah persentil 2.5% dan batas atas persentil 97.5% dari hasil nilai rata-rata setiap parameter yang merupakan hasil bangkitan distribusi posterior. Hipotesis yang digunakan adalah.

a. Parameter model *Negative Binomial state*.

$$H_0 : \beta_p = 0; p = 1, 2, \dots, k$$

$$H_1 : \beta_p \neq 0$$

b. Parameter model *zeros state*.

$$H_0 : \gamma_p = 0; p = 1, 2, \dots, k$$

$$H_1 : \gamma_p \neq 0$$

Kriteria pengambilan keputusan pada pengujian ini adalah dengan melihat adanya nilai nol pada *credible interval*. Jika pada selang antara batas bawah persentil 2.5% dan batas atas persentil 97.5% tidak mengandung nilai nol, maka  $H_0$  ditolak.

#### 1.5.15 Model Terbaik

*Akaike Information Criterion* (AIC) adalah salah satu metode dalam menentukan model yang terbaik dalam mengestimasi parameter. Nilai AIC adalah sebagai berikut (Burnham & Anderson, 2002).

$$AIC = 2k - 2 \ln(L(\hat{\theta})) \quad (24)$$

dengan  $L(\hat{\theta})$  adalah fungsi *likelihood* dan  $k$  adalah jumlah parameter. Kriteria lain yang merupakan ukuran relatif baiknya suatu model statistik adalah *Bayesian Information Criteria* (BIC). Nilai BIC adalah sebagai berikut (Schwarz, 1978).

$$BIC = k \ln(n) - 2 \ln(L(\hat{\theta})) \quad (25)$$

Model yang terbaik adalah model yang memiliki nilai AIC dan BIC yang terkecil.

#### 1.5.16 Multikolinearitas

Multikolinieritas adalah adanya hubungan antara variabel prediktor yang satu dengan variabel prediktor yang lain (Aulele, 2012). Permasalahan yang sering muncul pada multikolinieritas yaitu terjadinya korelasi yang cukup tinggi antara variabel-variabel prediktor. Pendeteksian multikolinieritas dapat dilakukan menggunakan nilai *Variance Inflation Factor* (VIF) dengan persamaan sebagai berikut.

$$VIF = \frac{1}{1 - R_p^2}; p = 1, 2, \dots, k \quad (26)$$

dengan  $R_p^2$  adalah nilai koefisien determinasi antara prediktor. Nilai  $VIF \geq 10$  menunjukkan bukti adanya multikolinieritas (Candraningtyas et al., 2013).

### 1.5.17 Demam Berdarah *Dengue*

Demam berdarah *dengue* (DBD) adalah penyakit yang ditularkan oleh gigitan nyamuk bernama *aedes aegypti*. Penyakit ini masih menjadi salah satu isu kesehatan masyarakat di Indonesia, dan tingkat penyebarannya di Indonesia termasuk yang tertinggi di antara negara-negara Asia Tenggara. Nyamuk *aedes aegypti* umumnya berukuran kecil dengan tubuh berwarna hitam pekat, memiliki dua garis vertikal putih di punggung dan garis-garis putih horizontal pada kaki. Nyamuk ini aktif terutama pada pagi hingga sore hari, meskipun kadang-kadang mereka juga menggigit pada malam hari.

Selain itu, memiliki riwayat terinfeksi virus *dengue* sebelumnya juga meningkatkan risiko mengalami gejala yang lebih parah ketika terkena DBD. Gejala utama penyakit DBD meliputi demam mendadak yang tinggi, mencapai suhu hingga 39 derajat *celsius*. Demam ini berlangsung terus-menerus selama 2-7 hari, kemudian turun dengan cepat. Gejala lain yang biasanya terjadi adalah nyeri kepala, menggigil, lemas, nyeri di belakang mata, otot, dan tulang, ruam kulit kemerahan, kesulitan menelan makanan dan minuman, mual, muntah, gusi berdarah, mimisan, timbul bintik-bintik merah pada kulit, muntah darah, dan buang air besar berwarna hitam.

Diagnosa kasus DBD ditegakkan dengan anamnesis perjalanan penyakit, gejala klinis, pemeriksaan fisik termasuk tanda vital dan tanda perdarahan, serta pemeriksaan penunjang konfirmasi diagnosis. Pasien terduga DBD akan menjalani pemeriksaan laboratorium yang mengindimekasikan penurunan trombosit < 100.000/mm<sup>3</sup> dan adanya kebocoran plasma yang ditandai dengan peningkatan hematokrit > 20%. Dalam pengendalian penyakit DBD, digunakan beberapa indikator untuk kegiatan pemantauan. Dua indikator utama yang digunakan adalah *Incidence Rate* per 100.000 penduduk dan *Case Fatality Rate* (Kemenkes RI, 2022).



## BAB II METODE PENELITIAN

### 2.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yang berasal dari Dinas Kesehatan dan Publikasi Badan Pusat statistik (BPS) Provinsi Kalimantan Utara, Kalimantan Timur, Kalimantan Selatan, Kalimantan Barat, dan Kalimantan Tengah tahun 2021. Jumlah unit data sebesar 56 wilayah yang terdiri dari 47 kabupaten dan 9 kota di pulau Kalimantan.

### 2.2 Identifikasi variabel

Variabel yang digunakan dalam penelitian ini terdiri dari variabel respon dan variabel prediktor yang diuraikan pada Tabel 1 sebagai berikut.

**Tabel 1.** Definisi operasional variabel penelitian

Jenis Variabel	Nama Variabel	Definisi Operasional	Satuan
Respon	Jumlah Kematian Akibat DBD ( $Y$ )	Jumlah kematian individu akibat penyakit DBD tahun 2021.	Kasus
Prediktor	Persentase Penduduk Miskin ( $X_1$ )	Jumlah penduduk yang memiliki rata-rata pengeluaran per kapita per bulan di bawah garis kemiskinan dibagi dengan jumlah penduduk dikali 100%.	Persen (%)
Prediktor	Persentase Rumah Tangga yang Memiliki Akses Terhadap Layanan Sanitasi Layak ( $X_2$ )	Jumlah rumah tangga yang dapat mengakses fasilitas sanitasi layak (fasilitas buang air besar sendiri/bersama, jenis kloset leher angsa, tempat pembuangan berupa septik tank/SPAL) dibagi dengan jumlah rumah tangga keseluruhan dikali 100%.	Persen (%)
Prediktor	Jumlah Fasilitas Kesehatan ( $X_3$ )	Jumlah unit fasilitas kesehatan berupa rumah sakit umum, rumah sakit khusus, puskesmas, dan klinik pratama.	Unit
Prediktor	Persentase Penduduk Yang Memiliki Jaminan Kesehatan Penerimaan	Jumlah peserta Jaminan Kesehatan Nasional (JKN) yang iurannya dibiayai oleh Pemerintah Daerah melalui Anggaran Pendapatan dan Belanja Negara (APBD) dibagi dengan jumlah penduduk dikali 100%.	Persen (%)

Bantuan Iuran (PBI)  
( $X_4$ )

**Lanjutan Tabel 1**

Jenis Variabel	Nama Variabel	Definisi Operasional	Satuan
Prediktor	Persentase Tempat-Tempat Umum (TTU) Memenuhi Syarat Kesehatan ( $X_5$ )	Jumlah unit fasilitas TTU memenuhi syarat kesehatan dibagi dengan jumlah TTU yang ada dikali 100%.	Persen (%)
Prediktor	Persentase Tempat Pengolahan Makanan (TPM) Memenuhi Syarat Kesehatan ( $X_6$ )	Jumlah unit fasilitas TPM memenuhi syarat kesehatan dibagi dengan jumlah TPM yang ada dikali 100%	Persen (%)
Prediktor	Persentase Rumah Tangga yang Memiliki Akses Terhadap Layanan Sumber Air Minum Layak ( $X_7$ )	Jumlah Rumah Tangga yang Memiliki Akses Terhadap Layanan Sumber Air Minum Layak dibagi dengan jumlah Rumah Tangga yang Memiliki Akses Layanan Sumber Air Minum dikali 100%	Persen (%)

### 2.3 Metode Analisis

Langkah-langkah yang dilakukan untuk menganalisis data dalam penelitian ini meliputi:

1) Estimasi Parameter Model ZINBR

Langkah-langkah estimasi parameter model ZINBR sebagai berikut:

- Menentukan fungsi kepadatan peluang  $Y \sim ZINB(\lambda, \pi, \tau)$  dengan mensubstitusi Persamaan (15) dan (16) ke Persamaan (14).
- Membentuk fungsi *likelihood* dari fungsi kepadatan peluang dari  $Y_i \sim ZINB(\lambda_i, \pi_i, \tau)$ .

$$L(y_i; \lambda_i, \pi_i, \tau) = \prod_{i: y_i=0} f(y_i = 0; \lambda_i, \pi_i, \tau) \times \prod_{i: y_i>0} f(y_i > 0; \lambda_i, \pi_i, \tau) \quad (27)$$

- Membentuk fungsi *ln likelihood*.

$$\ln L(y_i; \lambda_i, \pi_i, \tau) = \sum_{i: y_i=0} \ln(f(y_i = 0; \lambda_i, \pi_i, \tau)) + \sum_{i: y_i>0} \ln(f(y_i > 0; \lambda_i, \pi_i, \tau)) \quad (28)$$

- Menentukan fungsi distribusi prior dari parameter  $\beta$ ,  $\gamma$ , dan  $\tau$ .
- Menentukan fungsi distribusi posterior.
- Melakukan simulasi MCMC dengan algoritma *Gibbs Sampling*

2) Penerapan pada data jumlah kasus kematian akibat DBD di Pulau Kalimantan tahun 2021 sebagai berikut:

1. Menguji distribusi *Poisson* pada variabel respon menggunakan uji *Kolmogorov Smirnov*
2. Mendeteksi overdispersi pada variabel respon menggunakan nilai *Deviance*
3. Mendeteksi *excess zeros* pada variabel respon
4. Menguji asumsi non-multikolinieritas pada variabel prediktor menggunakan nilai *Variance Inflation Factor (VIF)*.
5. Mengestimasi parameter model ZINBR dengan metode MCMC dan algoritma *Gibbs Sampling*.
6. Menguji signifikansi parameter menggunakan *credible interval*
7. Menentukan model terbaik menggunakan nilai AIC
8. Menginterpretasikan model yang diperoleh

## 2.4 Diagram Alir

