ESTIMASI: Journal of Statistics and Its Application
Vol. 5, No. 1, Januari, 2024, Hal. 37-51
DOI: https://doi.org/10.20956/ejsa.v5i1.14552

Mengatasi Overdispersi Menggunakan Regresi Binomial Negatif dengan Penaksir Maksimum Likelihood pada Kasus Demam Berdarah di Kota Makassar

Muhammad Fadil[1*], Raupong[2], Nirwan Ilyas[3]
[123]Departemen Statistika, Fakultas MIPA, Universitas Hasanudin, Kota Makassar, 90245, Indonesia
***Corresponding author, email: mfadil.stat@gmail.com**

Abstract

The basic assumption in Poisson regression is that the mean value is the same as the variance value, which is called equidispersion. However, in some cases, this assumption is not met. A variance value that is greater than the average is called overdispersion and is called underdispersion if the variance value is smaller than the average value. So the Poisson regression model is no longer suitable for modeling this type of data because it will produce biased parameter estimates, therefore a negative binomial regression model is used. The research results show that estimating the parameters of the negative binomial regression model uses the maximum likelihood estimation method and then continues with the Newton-Raphson iteration method. The results obtained show that the negative binomial regression model overcomes the overdispersion that occurs in data on the number of dengue fever cases in Makassar City with the model $\hat{\mu}_i = \exp(7.01184 - 0.01356x_{i1} - 0.05200x_{i2})$ *and an AIC value of 236.06647. The negative binomial regression model produces many models and then the best model with the smallest AIC criteria is selected.*

***Keywords**: GLM, Poisson regression, Overdispersion, Negative Binomial Regression Model, AIC.*

Abstrak

Asumsi dasar dalam regresi Poisson yaitu nilai rataan sama dengan nilai variansinya yang disebut equidispersi. Namun, dalam beberapa kasus, asumsi tersebut tidak terpenuhi. Nilai variansi yang lebih besar daripada rataan disebut overdispersi dan disebut underdispersi jika nilai variansnya lebih kecil daripada nilai rataan. Sehingga model regresi Poisson tidak lagi cocok untuk memodelkan jenis data seperti ini karena akan mengakibatkan taksiran parameter menjadi bias, oleh karena itu digunakan model regresi binomial negatif. Hasil penelitian ini menunjukkan bahwa pendugaan parameter model regresi binomial negatif menggunakan metode estimasi maksimum likelihood kemudian dilanjutkan dengan menggunakan metode iterasi Newton-Raphson. Hasil yang diperoleh bahwa model regresi binomial negatif mengatasi overdispersi yang terjadi pada data jumlah kasus demam berdarah di Kota Makassar dengan model $\hat{\mu}_i = \exp(7.01184 - 0.01356x_{i1} - 0.05200x_{i2})$ dan nilai AIC yaitu 236.06647. Model regresi binomial negatif menghasilkan banyak model kemudian dipilih model terbaik dengan kriterial AIC terkecil.

Kata Kunci: GLM, regresi Poisson, overdispersi, model regresi binomial negatif, AIC.

1. Pendahuluan

Pada kasus tertentu, variabel penelitian mengandung excess zeros data. Excess zeros data adalah kondisi varibel respon bernilai nol yang disebabkan oleh data tidak terisi. Beberapa contoh data yang mengandung Excess zeros data diantaranya data jumlah kasus demam berdarah dengue (DBD), jumlah kematian ibu di suatu daerah, jumlah kecelakaan lalulintas, dan sebagainya. Pada penelitian ini data yang digunakan data jumlah kasus demam berdarah dengue (DBD).

Penyakit Demam Berdarah Dengue (DBD) adalah penyakit menular Aedes aegypti, disebabkan oleh virus dengue yang ditularkan melalui gigitan nyamuk. Data dari seluruh dunia menunjukkan Asia menempati urutan pertama dalam jumlah penderita DBD setiap tahunnya. Sehingga dapat dikatakan bahwa penyakit DBD merupakan salah satu permasalahan kesehatan di Asia terutama di Indonesia. Jumlah kabupaten/kota di Indonesia yang terjangkit DBD mengalami kenaikan, dari 434 (84.44%) pada tahun 2017 menjadi 440 (85.60%) pada tahun 2018. Terdapat 34 Provinsi di Indonesia dan semua terjangkit DBD terutama di Provinsi Sulawesi Selatan sebanyak 2.114 kasus dan jumlah meninggal sebanyak 19 kasus pada tahun 2018. Kota Makassar merupakan ibu kota daerah di Provinsi Sulawesi Selatan yang memiliki jumlah penduduknya cukup banyak dan berada pada ketinggian kurang dari 1000 meter. Sehingga, menjadikan Kota Makassar salah satu kota yang memiliki kasus DBD terbesar di Provinsi Sulawesi Selatan. Penyakit ini muncul sepanjang tahun dan dapat menyerang seluruh kelompok umur. Penyakit ini berkaitan dengan lingkungan sekitar dan perilaku masyarakat. Perbedaan perilaku masyarakat ini dapat mempengaruhi jumlah penderita DBD. Sehingga memungkinkan terjadinya overdispersi[1].

Beberapa peneliti telah menerapkan model regresi untuk data yang mengalami overdispersi, diantaranya Arwini (2018) telah menerapkan model Generalized Poisson Regression pada data yang mengalami overdispersi. Hasilnya menunjukkan bahwa overdispersi dapat diatasi dengan model Generalized Poisson Regression. Ade (2019) telah menerapkan model Regresi Hurdle Poisson pada data yang mengalami overdispersi. Hasilnya menunjukkan bahwa overdispersi dapat diatasi dengan membagi dua model. Pertama, model Zero Hurdle yang menginterpretasikan berapa besar peluang nilai variabel respon lebih besar dari nol (terjadi suatu event). Kedua, model Count yang menginterpretasikan berapa nilai rata–rata dari event tersebut. Pada penelitian ini, model regresi yang digunakan untuk mengatasi overdispersi adalah model regresi binomial negatif.

2. Material dan Metode

2.1 Regresi Linier Berganda

Dalam mengkaji hubungan antara beberapa variabel menggunakan analisis regresi, terlebih dahulu peneliti menentukan satu variabel yang disebut dengan variabel respon dan satu atau lebih variabel bebas. Jika ingin dikaji hubungan atau pengaruh satu variabel bebas terhadap variabel respon, maka model regresi yang digunakan adalah model regresi linier sederhana. Kemudian jika ingin dikaji hubungan atau pengaruh dua atau lebih variabel bebas terhadap variabel respon, maka model regresi yang digunakan adalah model regresi linier berganda (multiple linear regression model). Kemudian untuk mendapatkan model regresi linier sederhana maupun model regresi linier berganda dapat diperoleh dengan melakukan estimasi terhadap parameter-parameternya menggunakan metode tertentu[2].

Bentuk umum model regresi linier berganda dengan p variabel bebas adalah seperti pada persamaan (1) berikut[3].

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1,2,\cdots,n \tag{1}$$

dengan:

y_i = variabel respon pada pengamatan ke-i, untuk $i = 1,2,\dots,n$.
x_{ij} = variabel bebas pada pengamatan ke-i dan parameter ke-j.
β_j = parameter regresi yang tidak diketahui nilainya dan akan dicari nilai estimasinya.
ε_i = sisaan untuk pengamatan ke-i yang diasumsikan berdistribusi normalyang saling bebas dan identik dengan rata-rata 0 (nol) dan variansi σ^2 atau dituliskan sebagai $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$.

Dalam notasi matriks persamaan (1) dituliskan sebagai persamaan (2) berikut:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

dengan:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

2.2 Model Linier Umum

Model linier umum atau *Generalized linier Model* (GLM), termasuk salah satunya regresi linier klasik dapat dinyatakan dalam model (1). Variabel respon $\boldsymbol{Y}$ dalam model merupakan suatu fungsi linier dari koefisien model β_j, maka dari itu dinamakan model linier, sedangkan $\boldsymbol{\varepsilon}$ merupakan nilai sisaan. Sisaan dalam model diasumsikan berdistribusi normal.

Model (1) dapat digeneralisasikan untuk menangani sisaan yang tidak berdistribusi normal. Dalam GLM, variabel respon masih berhubungan dengan bebas melalui kombinasi linier $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$. Namun hubungan tersebut tidak bisa dilakukan secara langsung, perlu sebuah fungsi untuk menghubungkan keduanya. Fungsi tersebut dinamakan *link function*. Transformasi terhadap nilai rataan dari variabel respon memiliki hubungan linier dengan variabel-variabel bebasnya. Terdapat tiga komponen utama yang harus ada dalam suatu GLM, yaitu[4]:

1. Komponen Acak

Dalam sebuah GLM, harus terdapat $y_1,\ y_2, \dots, y_n$ yang merupakan variable respon berasal dari distribusi keluarga eksponensial.

2. Komponen Sistematis

   Terdapat variabel bebas yang linier $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ dari variabel respon.

3. Fungsi Penghubung

   Fungsi penghubung atau *link function* adalah suatu fungsi yang menghubungkan komponen acak dengan komponen sistematis.

$$g(\mu_i) = \boldsymbol{X\beta}$$

Fungsi transformasi g disebut dengan *link function*. Fungsi ini menentukan bagaimana nilai rataan berhubungan dengan variabel-variabel bebas. Untuk distribusi binomial negatif link function yang digunakan adalah fungsi logaritma.

2.3 Data Tercacah

Dalam statistik, data cacah mengacu pada pengamatan yang hanya memiliki nilai bilangan bulat non-negatif mulai dari nol hingga nilai yang lebih besar yang belum ditentukan. Secara teoritis, hitungan dapat berkisar dari nol hingga tak terbatas, tetapi mereka selalu terbatas pada beberapa nilai yang berbeda[5].

Model data cacah bertujuan untuk menjelaskan berapa cacah banyaknya suatu kejadian. Sebagai contoh, banyaknya kecelakaan yang terjadi per hari di suatu daerah, banyaknya pasien masuk per hari dalam sebuah rumah sakit.

2.4 Newton Raphson

Metode Newton-Raphson (NR) adalah salah satu metode untuk menyelesaikan persamaan $f(x) = 0$. Ciri-ciri Metode NR yaitu:

1. Memerlukan sebuah hampiran awal, dan
2. Memerlukan perhitungan turunan fungsi $f(x)$ dalam setiap iterasi.

Kedua ciri-ciri metode Newton menyatakan bahwa hampiran berikutnya diperoleh dengan cara menarik garis singgung kurva $y = f(x)$ pada titik yang mempunyai absis hampiran sebelumnya hingga memotong sumbu x. Titik potong garis singgung tersebut dengan sumbu x merupakan hampiran berikutnya. Proses berlanjut sampai hampiran yang diperoleh memenuhi syarat keakuratan yang ditentukan.

Misalkan g adalah suatu fungsi. Bilangan x pada domain g merupakan titik tetap g jika memenuhi $x = g(x)$. Iterasi

$$x_{n+1} = g(x), n = 0,1,2,3,\cdots \tag{3}$$

disebut iterasi titik tetap.

Misalkan fungsi f mempunyai turunan pertama f'. Barisan $x_0, x_1, x_2, \cdots$ yang diperoleh dari iterasi:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}, n = 0,1,2, \cdots \quad (4)$$

disebut barisan iterasi Newton. Fungsi g yang di definisikan sebagai:

$$g(x) = x_n - \frac{f'(x_n)}{f''(x_n)} \quad (5)$$

disebut fungsi iterasi Newton-Raphson[5].

2.5 Regresi Binomial Negatif

Regresi binomial negatif merupakan suatu model regresi yang digunakan untuk menganalisis hubungan antara suatu variabel respon yang berupa data cacah dengan satu atau lebih variabel bebas. Regresi binomial negatif dapat digunakan baik dalam keadaan equidispersi ataupun overdispersi.

Regresi binomial negatif merupakan salah satu model terapan dari GLM. Sebagai penerapan dari GLM maka distribusi binomial negatif memiliki tiga komponen yaitu:

1. Komponen Acak
   Pada model binomial negatif variabel respon diasumsikan berdistribusi binomial negatif yang dihasilkan dari distribusi campuran Poisson-gamma[6].
   Misalkan:
   $Y \sim$ Poisson(μ_i)
   $\mu_i \sim$ Gamma(α, β)
   Fungsi peluang campuran Poisson-gamma dapat diperoleh dengan cara:

$$\begin{aligned} \Pr(Y = y_i) &= \int_0^\infty \text{Poisson}(Y|\mu_i)\text{Gamma}(u_i|\alpha,\beta)d\mu_i \\ &= \frac{\Gamma\left(y_i + \frac{1}{\psi}\right)}{\Gamma\left(\frac{1}{\psi}\right)y!}\left(\frac{\mu\psi}{1+\mu\psi}\right)^y\left(\frac{1}{1+\mu\psi}\right)^{\frac{1}{\psi}} \end{aligned} \quad (6)$$

2. Komponen Sistematis
   Kontribusi variabel bebas dalam model regresi binomial negatif dinyatakan dalam bentuk kombinasi linier antara parameter μ_i dengan parameter regresi yang diestimasi yaitu:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (7)$$

   atau dituliskan dalam bentuk matriks:

$$\mu_i = \boldsymbol{X\beta} \quad (8)$$

dengan μ_i adalah vektor $n \times 1$ dari observasi, $\boldsymbol{X}$ adalah matriks $n \times (p+1)$ dari variabel bebas, dan $\boldsymbol{\beta}$ adalah matriks $(p+1) \times 1$ dari koefisien regresi.

3. Fungsi Penghubung
   Nilai rataan dari variabel acak adalah diskrit dan bernilai positif. Maka untuk mentransformasikan nilai (bilangan riil) ke rentang yang sesuai dengan rentang pada variabel respon diperlukan suatu fungsi *link* $g(\mu_i)$ yaitu:

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{X\beta}$$
$$\mu_i = e^{\boldsymbol{X\beta}} \tag{9}$$

Komponen GLM pada persamaan (9) dibagian ketiga (fungsi *link*), maka rataan dan variansi pada regresi binomial negatif dapat dituliskan kembali dalam bentuk sebagai berikut:

$$E[Y] = \mu$$
$$\mu_i = e^{\boldsymbol{X\beta}} \tag{10}$$

$$Var[Y] = \mu + \psi\mu^2$$
$$\mu_i = e^{(\boldsymbol{X\beta})} + \psi\left(e^{(\boldsymbol{X\beta})}\right)^2 \tag{11}$$

rataan dan variansi yang diperoleh dari persamaan (10) dan persamaan (11), maka persamaan (6) dapat dituliskan dalam bentuk sebagai berikut:

$$f(y) = \frac{\Gamma\left(y_i + \frac{1}{\psi}\right)}{\Gamma\left(\frac{1}{\psi}\right) y_i!} \left(\frac{e^{(\boldsymbol{X\beta})}\psi}{1 + e^{(\boldsymbol{X\beta})}\psi}\right)^{y_i} \left(\frac{1}{1 + e^{(\boldsymbol{X\beta})}\psi}\right)^{\frac{1}{\psi}} \tag{12}$$

Untuk mengestimasi parameter $\hat{\beta}$ dan $\hat{\psi}$ dalam regresi binomial negatif dapat digunakan metode penduga kemungkinan maksimum (MLE).

2.6 Overdispersi

Kegagalan asumsi Poisson dari equidispersi memiliki konsekuensi kualitatif serupa terhadap kegagalan asumsi homoskedastisitas dalam model regresi linier. Data overdispersi jika variansi bersyarat melebihi rataan bersyarat. Indikasi besarnya overdispersi atau underdispersi dapat diperoleh hanya dengan membandingkan rataan sampel dan variansi dari variabel respon. Selanjutnya, regresi Poisson menghasilkan variansi bersyarat dari variabel respon. Rata-rata dari rataan bersyarat tidak akan berubah, karena rata-rata dari *fittted* rataan sama dengan rataan sampel. Ini mengikuti karena sisaan Poisson berjumlah nol jika ada istilah konstan. Jika varians sampel kurang dari rataan sampel, data tentu bahkan lebih underdispersi setelah regresor disertakan. Jika varians sampel lebih dari dua kali rata-rata sampel, maka data kemungkinan akan tetap overdispersi setelah masuknya regressor. Hal ini terutama berlaku untuk *cross-section*, yang dimana regressor biasanya menjelaskan kurang dari setengah variasi dalam data[7].

Ada atau tidaknya overdispersi juga dapat dilihat dari nilai *Deviance* atau Pearson *Chi-square* yaitu sebagai beriku:

1. *Deviance*
   Nilai deviance adalah nilai logaritma dari uji rasio likelihood-nya[4]. Uji rasio likelihood-nya membandingkan current model-nya dan saturated model-nya yang di tuliskan sebagai berikut:
$$\phi_1 = \frac{D^2}{db}$$
$$D^2 = 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\widehat{y_i}}\right) - (y_i - \widehat{y_i})\right) \tag{13}$$
   dengan:
   y_i = Nilai variabel respon
   $\widehat{y_i}$ = Estimasi regresi Poisson
   dimana $db = n - p$ dengan p merupakan banyak parameter, n merupakan banyaknya pengamatan dan D^2 adalah nilai deviance.

2. *Pearson Chi-Square*
   Pengukuran lain yang digunakan untuk mendeteksi overdispersi yaitu statistic *person chi-square*[4], yang di definisikan sebagai:
$$\phi_2 = \frac{\chi^2}{db}$$
$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - \widehat{y_i})}{\widehat{y_i}} \tag{14}$$

Dimana $db = n - p$ dengan p merupakan banyaknya parameter, n merupakan banyaknya pengamatan dan χ^2 adalah nilai pearson *chi-square.* Jika nilai ϕ_1 dan ϕ_2 lebih besar dari 1, ini menunjukkan nilai variansi yang lebih besar daripada rataan, maka telah terjadi overdispersi[8].

2.7 Pemilihan Model Terbaik

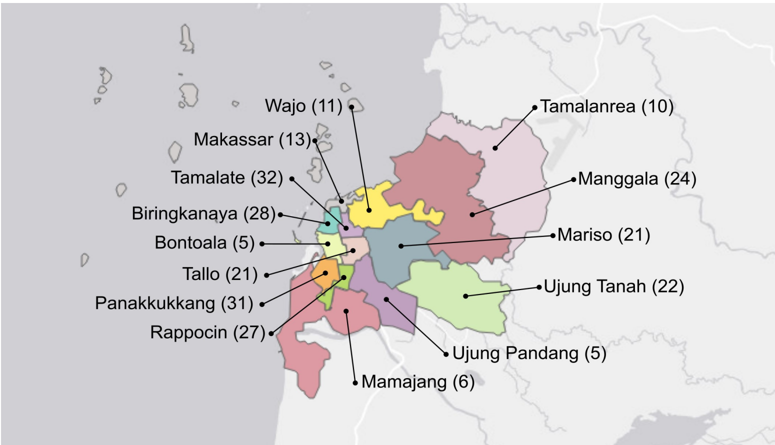
Pemilihan model regresi terbaik dapat dilihat dari perhitungan *nilai Akaike Information Criterion* (AIC). AIC model regresi binomial negatif yaitu:
$$AIC = -2[L - (p+1)] \tag{15}$$

dengan L adalah fungsi ln-likelihood dari hasil *estimasi maximum likelihood* dan p adalah jumlah parameter dalam model. Jika model memiliki nilai likelihood tertinggi atau nilai AIC minimum maka model tersebut adalah model terbaik[5].

3. Hasil dan Diskusi

3.1 Deskripsi Data

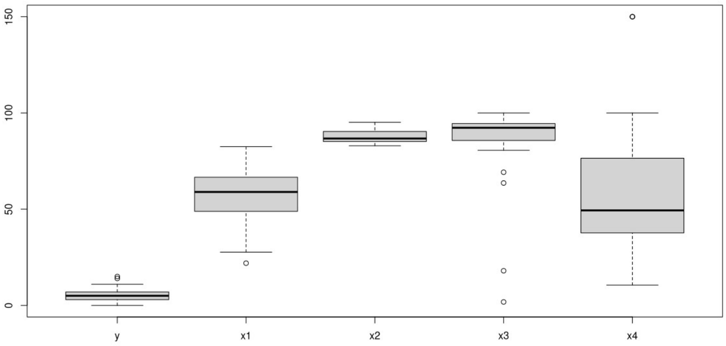
Data yang digunakan pada penelitian ini merupakan data sekunder yang diperoleh dari Profil Kesehatan Kota Makassar tahun 2018 yang terdiri atas 46 observasi berdasarkan jumlah puskesmas yang ada di kota Makassar untuk lebih lengkapnya, data jumlah kasus DBD Kota di Makassar tahun 2018 dapat dilihat pada lampiran 1. Daerah kota Makassar dapat dilihat pada peta berikut.



Gambar 1. Peta Kota Makassar dan daerah penderita DBD tahun 2018

Gambar 1 memperlihatkan penyebaran DBD yang beragam, hal ini disebabkan oleh perbedaan lingkungan. Terdapat 32 kasus pada Kec. Tamalate, 31 kasus pada Kec. Panakkukkang, 28 kasus pada Kec. Biringkanaya, 27 kasus pada Kec. Rappocin, 24 kasus pada Kec. Manggala, 22 kasus pada Kec. Ujung Tanah, 21 kasus pada Kec. Tallo Rappokalling, 21 kasus pada Kec. Mariso, 13 kasus pada Kec. Makassar, 11 kasus pada Kec. Wajo, 10 kasus pada Kec. Tamalanrea, 6 kasus pada Kec. Mamajang, 5 kasus pada Kec. Ujung Pandang, dan 5 kasus pada Kec. Bontoala. Maka jumlah kasus DBD di kota Makassar tahun 2018 sebanyak 256 kasus.

Perbedaan jumlah penderita DBD tidak menutup kemungkinan dipengaruhi oleh beberapa faktor misalnya persentase rumah yang memenuhi syarat kesehatan (x_{i1}), persentase rumah tangga berperilaku hidup bersih dan sehat (x_{i2}), persentase penduduk yang memiliki akses air minum layak (x_{i3}), dan persentase tempat-tempat umum yang memenuhi syarat kesehatan (x_{i4}). Gambar 2 menunjukkan boxplot dari mesing-masing variabel yang mungkin mempengaruhi jumlah penderita DBD di Kota Makassar tahun 2018.



Gambar 2. Boxplot masing-masing variabel data jumlah kasus DBD di Kota Makassar tahun 2018.

Gamber 2 memberikan informasi mengenai ada tidaknya *outliers*. Terlihat jelas bahwa variabel y_i mengandung *outliers*. Hal ini belum memberikan infomasi yang cukup untuk melihat krakteristik dari masing-masing variabel. krakteristik dari masing-masing variabel dapat diperoleh dari statistik deskriptif dari variabel respon y_i dan kelima variabel bebasnya. Statistik deskriptif masing-masing variabel data jumlah kasus demam berdarah di Kota Makassar tahun 2018 dapat dilihat pada **tabel 1** berikut:

Table 1. Statistik deskriptif masing-masing variable data jumlah kasus DBD di Kota Makassar tahun 2018

Variabel	N	Minimum	Maksimum	Rataan	Variansi	Std.Deviasi
y_i	46	0	15.0000	5.5652	11.1401	3.3377
$xi1$	46	21.9480	82.4880	57.0388	193.4870	13.9100
$xi2$	46	82.9530	95.1330	87.6863	11.4620	3.3856
$xi3$	46	1.8000	100.0000	86.7572	329.5232	18.1528
$xi4$	46	10.5300	150.0000	56.9389	895.0318	29.9171

Sumber: data olah 2021

Tabel 1 memberikan informasi mengenai statistik deskriptif masing-masing variabel data jumlah kasus demam berdarah di Kota Makassar. Terlihat jelas bahwa variabel y_i memiliki nilai variansi lebih besar dari nilai rataannya yang memungkinkan terjadi overdispersi. Untuk memastikan bahwa variabel respon y_i mengalami overdispersi maka dilakukan pengujian lanjutan

3.2 Pendugaan Parameter $\boldsymbol{\beta}$ dan $\boldsymbol{\psi}$ pada Model Regresi Binomial Negatif

Pendugaan parameter β pada model regresi binomial negatif dapat menggunakan metode estimasi maksimum likelihood. Fungsi likelihood untuk model regresi binomial negatif yaitu:

$$L(\boldsymbol{\beta},\psi)=\prod_{i=1}^{n}\big(f(y)\big)$$

$$=\prod_{i=1}^{n}\frac{\Gamma\left(y_i+\frac{1}{\psi}\right)}{\Gamma\left(\frac{1}{\psi}\right)y_i!}\prod_{i=1}^{n}\left(\frac{e^{(\boldsymbol{X\beta})}\psi}{1+e^{(\boldsymbol{X\beta})}\psi}\right)^{y_i}\prod_{i=1}^{n}\left(\frac{1}{1+e^{(\boldsymbol{X\beta})}\psi}\right)^{\frac{1}{\psi}} \quad (16)$$

Fungsi likelihood akan dilogaritmakan sehingga diperoleh fungsi log-likelihood regresi binomial negatif yaitu:

$$l(\boldsymbol{\beta},\psi)=\log\big(L(\beta,\psi)\big)$$

$$=\sum_{i=1}^{n}\left[\sum_{r=0}^{y_i-1}\log(1+r\psi)-y_i\log(\psi)-\log(y_i!)+y_i\log\big(e^{(\boldsymbol{X\beta})}\psi\big)-\left(y_i+\frac{1}{\psi}\right)\log\big(1+e^{(\boldsymbol{X\beta})}\psi\big)\right] \quad (17)$$

Fungsi log-likelihood model regresi binomial negatif diturunkan terhadap β kemudian disamakan dengan nol. Turunan pertama fungsi log-likelihood terhadap β merupakan matriks G yang dapat dituliskan dalam model sebagai berikut:

$$\boldsymbol{G}=\begin{bmatrix}\frac{\partial l(\beta,\psi)}{\partial\beta_0}\\ \frac{\partial l(\beta,\psi)}{\partial\beta_1}\\ \frac{\partial l(\beta,\psi)}{\partial\beta_2}\\ \vdots\\ \frac{\partial l(\beta,\psi)}{\partial\beta_p}\end{bmatrix} \quad (18)$$

secara umum:

$$\left.\frac{\partial l(\beta,\psi)}{\partial\beta_j}\right|_{\beta=\hat{\beta}}=\sum_{i=1}^{n}\left[\frac{\big(y_i-e^{(\boldsymbol{X}\hat{\boldsymbol{\beta}})}\big)x_{ij}}{\big(1+e^{(\boldsymbol{X}\hat{\boldsymbol{\beta}})}\psi\big)}\right]=0,\qquad j=0,1,2,\dots,p. \quad (19)$$

Turunan kedua fungsi log-likelihood disebut matiriks *Hessian* ($\boldsymbol{H}$) yang dituliskan dalam bentuk sebagai berikut:

$$H = \begin{bmatrix} \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial^2 \beta_0} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_0 \partial \beta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_0 \partial \beta_4} \\ \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial^2 \beta_1} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_1 \partial \beta_4} \\ \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_0 \partial \beta_2} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_1 \partial \beta_2} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial^2 \beta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_2 \partial \beta_4} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_0 \partial \beta_4} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_1 \partial \beta_4} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_2 \partial \beta_3} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial \beta_3 \partial \beta_4} & \frac{\partial^2 l(\boldsymbol{\beta},\psi)}{\partial^2 \beta_4} \end{bmatrix} \tag{20}$$

secara umum:

$$\left.\frac{\partial l(\boldsymbol{\beta},\psi)}{\partial \beta_j \partial \beta_k}\right|_{\beta=\hat{\beta}} = -\sum_{i=1}^{n} \left[\frac{x_{ij} x_{ik} e^{(X\hat{\beta})}(1+y_i\psi)}{\left(1+e^{(X\hat{\beta})}\psi\right)^2} \right], \; j,k = 0,1,\dots,p \tag{21}$$

Fungsi log-likelihood model regresi binomial negatif pada persamaan (17) terdapat parameter ψ yang tidak diketahui. Untuk mengestimasi parameter ψ, maka fungsi log-likelihood model regresi binomial negatif diturunkan terhadap ψ kemudian disamakan dengan nol. Turunan pertama fungsi log-likelihood terhadap ψ yang dapat dituliskan dalam bentuk sebagai berikut:

$$\left.\frac{\partial l(\beta,\psi)}{\partial \psi}\right|_{\psi=\hat{\psi}} = \sum_{i=1}^{n} \left[\sum_{r=0}^{y_i-1} \left(\frac{r}{1+e^{(X\beta)}\hat{\psi}} \right) + \frac{\log\left(1+e^{(X\beta)}\hat{\psi}\right)}{\hat{\psi}^2} - \frac{\left(y_i+\frac{1}{\hat{\psi}}\right)e^{(X\beta)}}{1+e^{(X\beta)}\hat{\psi}} \right] = 0 \tag{22}$$

Turunan kedua fungsi log-likelihood terhadap ψ yang diruliskan dalam bentuk sebagai berikut:

$$\left.\frac{\partial^2 l(\beta,\psi)}{\partial^2 \hat{\psi}}\right|_{\psi=\hat{\psi}} = \frac{\partial^2}{\partial^2 \hat{\psi}} \left[\sum_{i=1}^{n} \left[\sum_{r=0}^{y_i-1} \left(\frac{r}{1+e^{(X\beta)}\hat{\psi}} \right) + \frac{\ln\left(1+e^{(X\beta)}\hat{\psi}\right)}{\hat{\psi}^2} - \frac{\left(y_i+\frac{1}{\hat{\psi}}\right)e^{(X\beta)}}{1+e^{(X\beta)}\hat{\psi}} \right] \right] = 0 \tag{23}$$

Karena turunan fungsi log-likelihood model regresi binomial negatif terhadap β dan ψ tidak dapat diselesaikan secara analitik maka dilanjutkan dengan menggunakan metode iterasi Newton-Raphson dengan langkah-langkah sebagai berikut:

1. Menentukan nilai taksiran awal parameter $\hat{\psi}_0$
2. Menentukan nilai taksiran awal parameter $\hat{\beta}_0$ menggunakan metode OLS.
3. Dilakukan iterasi pada persamaan berikut:

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} - H_{(r)}^{-1} G_{(r)}$$

$\hat{\beta}_{(r)}$ merupakan nilai taksiran awal yang diperoleh dari metode OLS.

4. Dilakukan iterasi pada persamaan berikut:

$$\hat{\psi}_{(r+1)} = \hat{\psi}_{(r)} - \frac{\left(\partial l(\beta, \psi)\right)}{\left(\partial^2 l(\beta, \psi)\right)}$$

Jika belum mendapatkan penaksir parameter yang konvergen, maka di lanjutkan kembali langkah ke-3 dan berhenti apabila nilai $||\hat{\beta}_{(r+1)} - \hat{\beta}_{(r)}|| < \varepsilon$, ε adalah

3.3 Pengujian Overdispersi

Analisis regresi Poisson mengasumsikan data pada variabel respon berdistribusi Poisson dengan rataan dan variansi bernilai sama (equidispersi). Namun, pada kenyataanya asumsi bahwa data equidispersi seringkali tidak terpenuhi yaitu nilai variansi lebih besar dari nilai rataan (overdispersi) atau nilai variansi lebih kecil dari nilai rataan (undersipersi). Variabel respon yang digunakan pada penelitian ini adalah data diskrit yaitu jumlah kasus demam berdarah di Kota Makassar tahun 2018 sebanyak 46 data dengan nilai rataan sebesar 5.5652 dan variansi 11.1401 yang diperoleh dari **tabel 1**. Selain nilai rataan dan variansi, kasus overdispersi juga dapat dideteksi dengan memeriksa nilai *deviance* atau nilai *chi-square*.

1. **Uji *Deviance***
   a. Hipotesis:
      H_0: Terjadi overdispersi
      H_1: Tidak terjadi overdispersi
   b. Statistik Uji:
   $$\phi_1 = \frac{D^2}{db}$$
   $$D^2 = 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right)$$
   c. Kriteria Penolakan:
      Tolak H_0 jika $\phi_1 < 1$
   d. Keputusan:
   $$\phi_1 = \frac{2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right)}{db}$$
   $$= 8.13422$$
   Karena $\phi_1 = 8.13422$ tampak bahwa nilai $\phi_1 > 1$, maka terima H_0. Sehingga dapat dikatakan bahwa terjadi overdispersi.
2. **Uji *pearson Chi-square***
   a. Hipotesis:
      H_0: Terjadi overdispersi
      H_1: Tidak terjadi overdispersi

b. Statistik Uji:

$$\phi_2 = \frac{\chi^2}{db}$$

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

c. Kriteria Penolakan:
   Tolak H_0 jika $\phi_2 > 1$
d. Keputusan:

$$\begin{aligned} \phi_2 &= \frac{\chi^2}{db} \\ &= \frac{\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mu_i}}{db} \\ &= 3.3089 \end{aligned}$$

   Karena $\phi_2 = 3.3089$ tampak bahwa nilai $\phi_2 > 1$, maka terima H_0. Sehingga dapat dikatakan bahwa terjadi overdispersi.

Nilai ϕ_1 dan ϕ_2 lebih besar dari 1, hal tersebut menunjukkan nilai variansi yang lebih besar daripada rataan, maka dapat disimpulkan bahwa data variabel respon pada data jumlah khasus demam berdarah di Kota Makassar tahun 2018 terjadi overdispersi.

3.3 Estimasi nilai Parameter yang Signifikan Model Binomial Negatif

Estimasi nilai parameter regresi binomial negatif dengan variabel bebas yang signifikan, dengan bantuan *software* RStudio. Maka diperoleh estimasi parameter model regresi binomial negatif sebagai berikut:

Table 2. Estimasi nilai parameter regresi binomial negatif parameter signifikan

No	Beta	Estimasi	Est.Error	Z.Value	Wald
1.	β_0	7.01184	1.67392	4.18888	17.54675
2.	β_1	−0.01356	0.00467	−2.90205	8.42192
3.	β_2	−0.05200	0.01954	−2.66097	7.08075
4.	ψ	0.10740	5.84371	0.01837	0.00034
log − likelihood: − 1961.05615					

Sumber: data olah, 2021

Berdasarkan **tabel 2** diperoleh model regresi binomial negatif sebagai berikut:

$$\hat{\mu} = \exp(7.01184487 - 0.01355819x_{i1} - 0.05199649x_{i2})$$

Intepretasi:

1. Ketika nilai persentase rumah yang memenuhi syarat kesehatan (x_{i1}) bertambah 0.1 maka jumlah kasus demam berdarah berkurang sebesar $0.9865 \approx 1$ kasus, dengan menganggap variabel lainnya konstan.

2. Ketika nilai persentase rumah tangga berperilaku hidup bersih dan sehat (x_{i2}) bertambah 0.1 maka jumlah kasus demam berdarah berkurang sebesar $0.9493 \approx 1$ kasus, dengan menganggap variabel lainnya konstan.

3.4 Pemilihan Model Terbaik

Pemilihan model terbaik regresi binomial negatif dapat dilihat pada lampiran 6. Maka model yang memiliki nilai *Akaike Information Criterion* (AIC) minimum antar setiap model regresi binomial negatif yaitu:

$$\hat{\mu}_i = exp(7.01184487 - 0.01355819x_{i1} - 0.05199649x_{i2})$$

sehingga hal ini menunjukkan bahwa overdispersi dapat teratasi dengan menggunakan model $\hat{\mu}_i = exp(7.01184 - 0.01356x_{i1} - 0.05199x_{i2})$ dengan nilai AIC yang diperoleh yaitu 236.06647 yang menunjukkan bahwa rata-rata jumlah kasus demam berdarah pada tahun 2018 per Kecamatan adalah 5.563.

4. Kesimpulan

Berdasarkan hasil yang diperoleh dan diuraikan pada pembahasan sebelumnya, dapat ditarik kesimpulan bahwa penduga model regresi binomial negatif terbaik pada data jumlah kasus demam berdarah di Kota Makassar tahun 2018 dengan kriteria AIC terkecil 236.06647 adalah $\hat{\mu}_i = \exp(7.0118 - 0.0135x_{i1} - 0.05199x_{i2})$ yang menunjukkan bahwa penduga rata-rata jumlah kasus demam berdarah pada tahun 2018 per Kecamatan adalah 5.563.

Daftar Pustaka

[1] Badan Pusat Statistik. *Provinsi Sulawesi Selatan dalam Angka 2021*. Makassar: Badan Pusat Statistik Sulawesi Selatan. 2021.

[2] Kutner, M. H., Nachtsheim, C. J., & Neter, J. *Applied Linear Regression Models. 4th ed*. New York: McGraw-Hill Companies Inc. 2004.

[3] Hardle, W. *Applied Nonparametric Regression*. United State: Cambridge University Press. 1995.

[4] McCullagh, P., & Nelder, J. A. *Generalized Liner Model, second edition*. London: Chapman and Hall. 1989.

[5] Ismail, N., & Jemain, A. A. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. *Casualty Actuarial Society Forum*, 8, 367-394, 2007.

[6] Simarmata, R. T., & Ispriyanti, D. Penanganan Overdispersi pada Model Regresi Poisson Menggunakan Model Regresi Binomial Negatif. *Media Statistika*, 4(2), 95-104, 2011.

[7] Cameron, A. C., & Trivedi, P. K. *Regression Analysis of Count Data*. New York: Cambridge University Press. 1998.

[8] Hilbe, J. M. *Negative Binomial Regression (2nd ed.)*. New York, NY, US: Cambridge University Press. 2011.