MENGATASI OVERDISPERSI MENGGUNAKAN REGRESI BINOMIAL NEGATIF DENGAN PENAKSIR MAKSIMUM LIKELIHOOD PADA KASUS DEMAM BERDARAH DI KOTA MAKASSAR

MUHAMMAD FADIL

H 121 15 006

Pembimbing Utama	**:**	**Drs. Raupong, M.Si.**
Pembimbing Pendamping	**:**	**Dr. Nirwan Ilyas, M.Si.**
Penguji	**:**	**1. Dr. Nurtiti Sunusi, S.Si., M.Si.**
		2. Siswanto, S.Si., M.Si.

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2021

MENGATASI OVERDISPERSI MENGGUNAKAN REGRESI BINOMIAL NEGATIF DENGAN PENAKSIR MAKSIMUM LIKELIHOOD PADA KASUS DEMAM BERDARAH DI KOTA MAKASSAR

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin Makassar

MUHAMMAD FADIL

H 121 15 006

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2021

LEMBAR PERNYATAAN KEOTENTIKAN

Saya yang bertanda tangan di bawah ini menyatakan dengan sungguh-sungguh bahwa skripsi yang saya buat dengan judul:

MENGATASI OVERDISPERSI MENGGUNAKAN REGRESI BINOMIAL NEGATIF DENGAN PENAKSIR MAKSIMUM LIKELIHOOD PADA KASUS DEMAM BERDARAH DI KOTA MAKASSAR

adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.

Makassar, 30 April 2021

MUHAMAMD FADIL
NIM. H 121 15 006

iii

MENGATASI OVERDISPERSI MENGGUNAKAN REGRESI BINOMIAL NEGATIF DENGAN PENAKSIR MAKSIMUM LIKELIHOOD PADA KASUS DEMAM BERDARAH DI KOTA MAKASSAR

Disetujui oleh:

Pembimbing Utama	Pembimbing Pertama
Drs. Raupong, M.Si.	**Dr. Nirwan Ilyas, M.Si.**
NIP. 19621015 198810 1 001	**NIP. 19630306 198702 1 002**

Ketua Departemen Statistika

KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN UNIVERSITAS HASANUDDIN DEPARTEMEN STATISTIKA FAKULTAS MIPA

Dr. Nurtiti Sunusi, S.Si, M.Si.
NIP. 19720117 199703 200 2

Pada tanggal : 30 April 2021

iv

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh	:	
Nama	:	MUHAMMAD FADIL
NIM	:	H 121 15 006
Program Studi	:	STATISTIKA
Judul Skripsi	:	Mengatasi Overdispersi menggunakan Regresi Binomial Negatif dengan Penaksir Maksimum Likelihood pada Kasus Demam Berdarah di Kota Makassar

"Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin"

DEWAN PENGUJI

Tanda Tangan

1.	Ketua	:	Drs. Raupong, M.Si.	(..............................)
2.	Sekretaris	:	Dr. Nirwan Ilyas, M.Si.	(..............................)
3.	Anggota	:	Dr. Nurtiti Sunusi, S.Si., M.Si.	(..............................)
4.	Anggota	:	Siswanto, S.Si., M.Si.	(..............................)

Ditetapkan di : Makassar
Tanggal : 30 April 2021

v

KATA PENGANTAR

Segala puji syukur senantiasa penulis panjatkan kehadirat Allah Subhanahu Wa ta'ala mengiringi limpahan rahmat dan karunia-Nya yang tak terhingga sehingga penulis dapat menyelesaikan skripsi ini. Semua kemudahan yang penulis dapatkan tidak lepas dari pertolongan Allah Subhanahu Wa ta'ala sebagaimana Rasulullah shallallahu 'alayhi wa sallam bersabda "Jagalah Allah, pasti Dia menjagamu. Jagalah Allah, Dia senantiasa bersamamu. Jika kamu memohon sesuatu, mohonlah kepada-Nya. jika meminta pertolongan, mintalah tolong kepada-Nya. Ketahuilah seandainya semua umat manusia bersatu untuk memberikan suatu kebaikan kepadamu, mereka tidak akan mampu kecuali yang sudah ditetapkan Allah untukmu. Dan seandainya semua umat manusia bersatu untuk mencelakakanmu, mereka tidak mampu kecuali keburukan yang telah ditetapkan oleh Allah untukmu. Pena sudah diangkat dan tinta sudah kering." (HR Tirmidzi). Shalawat serta salam semoga tetap tercurahkan kepada nabi kita yang mulia, nabi Muhammad shallallahu 'alayhi wa sallam, beserta keluarga, para sahabat, dan pengikut-pengikutnya yang setia.

Alhamdulillahi rabbil'alamin Penulisan skripsi yang berjudul "Mengatasi Overdispersi menggunakan Regresi Binomial Negatif dengan Penaksir Maksimum Likelihood pada Kasus Demam Berdarah di Kota Makassar" yang disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin ini dapat dirampungkan.

Penyusunan skripsi ini tidak terlepas dari bantuan, pengarahan, dan bimbingan dari berbagai pihak. Karena itulah pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Keluarga penulis Almarhumah Ibunda **Dode** atas semua pengorbanan dan kasih sayang yang tulus kepada penulis serta Ayahanda **Bundu** atas semua nasihat, kerja

keras, pengorbanan, do'a dan kasih sayang yang tulus kepada penulis. Semoga Allah senantiasa memberkahi, menjaga, dan menyayangi kalian. Ketujuh saudara penulis **Pange, Kuti, Risna, Nadira, Ali, Nurmi, dan Nisma.** yang senantiasa membantu dan memberikan dukungan motivasi serta perhatian kepada penulis.

2. Ibu **Dr. Nurtiti, Sunusi, S.Si., M.Si.** selaku Ketua Departemen Statistika, sekaligus Penguji, yang telah membekali ilmu dan kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di jurusan Statistika.

3. Bapak **Drs. Raupong, M.Si.** sebagai dosen pembimbing utama sekaligus ketua tim penguji atas ilmu yang beliau berikan selama proses perkuliahan, dan kesediaan dalam membimbing, serta memotivasi penulis baik diluar maupun dalam penyusunan skripsi ini.

4. Bapak **Dr. Nirwan Ilyas, M.Si.** sebagai dosen pembimbing pertama sekaligus sekretaris tim penguji atas ilmu yang beliau berikan selama proses perkuliahan, dan bimbingan serta segala bentuk bantuan yang telah beliau berikan dalam penyusunan skripsi ini.

5. Bapak **Siswanto, S.Si., M.Si.** selaku penguji yang selama seminar telah banyak memberikan kritik dan saran yang sangat berharga dalam perbaikan skripsi ke arah yang lebih baik.

6. Bapak **Dr. Amran, S.Si., M.Si.** dan Bapak **Andi Kresna Jaya, S.Si., M.Si.** sebagai penasehat akademik, penulis mengucapkan terima kasih atas masukan dan ilmu yang diberikan kepada penulis selama menjalani pendidikan di Departemen Statistika.

7. **Segenap Dosen** dan seluruh **Staf Pegawai** Departemen Statistika Universitas Hasanuddin yang telah memberikan banyak ilmu serta bantuan kepada penulis

selama masa perkuliahan.

8. Kak **Rendi** dan Kak **Irfan Taufik** yang telah memberikan banyak ilmu, petunjuk, dan motivasi.
9. Sahabat-sahabatku **A.Muthiah Nur Angriany, Ade Kurniawan, Masjidil Aqsha, Hilmi Abyan, Ihza Kurniawan, Ilham, Irfan Apandi, Yoris Rombe, Muhammad Aris, Nurhidaya Rahim, Waode Rahmalia, Puji Puspasari, Trysha Aris, Sriwijayanti, Musaidah, Lisa Narisda (Almarhumah), dan Mirawati.** Terima kasih atas bantuan dan kebersamaannya selama ini.
10. **Keluarga Besar Statistika 2015**, terima kasih telah menjadi keluarga baru yang selalu ada.
11. Teman dan adik-adikku **Ahmad Nurhidayat, Ilham Halis, Rudi, Ainul Fajri, Arhan, Muhammad Yahya, dan Andi Riska Firiani** yang selalu membantuku dalam keadaan apapun.
12. *Software* **Arch-Linux, R dan RStudio, Latex dan Texmaker, Libreoffice, Inkscape,** Serta **Okular dan Chrome.** yang memberikan akses gratis kepada penulis.
13. Serta semua pihak yang telah membantu dalam penyelesaian skripsi ini yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa masih banyak kekurangan dalam tugas akhir ini, untuk itu dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga tulisan ini memberika manfaat untuk pembaca.

Makassar, 30 April 2021

Penulis

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama	:	Muhammad Fadil
NIM	:	H121 15 006
Program Studi	:	Statistika
Departemen	:	Statistika
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam
Jenis Karya	:	Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Prediktor Royalti Non-eksklusif (*Non-exclusive Royalty-Free Right*)** atas tugas akhir saya yang berjudul:

MENGATASI OVERDISPERSI MENGGUNAKAN REGRESI BINOMIAL NEGATIF DENGAN PENAKSIR MAKSIMUM LIKELIHOOD PADA KASUS DEMAM BERDARAH DI KOTA MAKASSAR

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.
Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar pada tanggal 30 April 2021
Yang menyatakan

(Muhammad Fadil)

ABSTRAK

Model regresi Poisson merupakan salah satu *Generalized Linear Model* (GLM) yang memodelkan data cacah (*Count data*). Asumsi dasar dalam regresi Poisson yaitu nilai rataan sama dengan nilai variansinya yang disebut equidispersi. Namun, dalam beberapa kasus, asumsi tersebut tidak terpenuhi. Misalnya terdapat nilai nol yang terlalu banyak pada variabel respon. Hal tersebut menyebabkan nilai rataan tidak lagi sama dengan variansi. 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.

ABSTRACT

The Poisson regression model is one of the Generalized Linear Models (GLM) that model data counts (data counts). The basic assumption in Poisson regression is that the rataan value is equal to the variance value called equidispersion. However, in some cases, these assumptions are not fulfilled. For example, there is too much zero in the response variable. That causes the rataan value is no longer the same as the variance. The variance value is greater than the rataan, called overdispersion. And called underdispersion if the variance value is smaller than the rataan value. So the Poisson regression model is no longer suitable for modeling data types like this because it will cause parameter estimates to be biased, therefore the Negative Binomial regression model is used. The results of this study indicate that the estimation of the negative binomial regression model parameters using the maximum likelihood estimation method is then continued by using the Newton-Raphson iteration method. The results obtained are that the negative binomial regression model overcomes the overdispersion that occurs in the 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 the AIC value is 236.06647. The negative binomial regression model produces many models then the best model is chosen with the minimum AIC criterion.

Keywords: GLM, Poisson regression, overdispersion, negative binomial regression, AIC.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang

Regresi merupakan suatu teknik yang digunakan untuk menganalisis hubungan antara variabel respon dengan satu atau lebih variabel bebas. Pada umumnya analisis regresi digunakan untuk menganalisis data variabel respon yang berupa data kontinu (Winkelmann, 2018). Salah satu model regresi yang dapat digunakan untuk menjelaskan hubungan antara variabel respon yang berupa data tercacah dengan variabel bebas yang berupa kontinu atau diskrit adalah regresi Poisson.

Analisi regresi Poisson adalah suatu model yang digunakan untuk menganalisis hubungan antara variabel respon yang berdistribusi Poisson dengan satu atau lebih variabel bebas. Pada model regresi Poisson terdapat asumsi yang harus dipenuhi yaitu nilai variansi yang diperoleh sama dengan nilai rataannya keadaan ini biasa disebut ekuidispersi. Namun pada kenyataannya asumsi ini sangat jarang terjadi. Biasanya data *count* memiliki variansi lebih besar dari rataan yang disebut overdispersi atau sebaliknya yaitu rataan lebih besar dari variansi atau biasa disebut undispersi.

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.

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.

1.2 Rumusan Masalah

Adapun rumusan masalah yang akan dikaji oleh penulis sesuai dengan latar belakang yang telah diuraikan yaitu bagaimana menduga model regresi binomial negatif terbaik menggunakan metode estimasi maksimum likelihood (MLE) melalui iterasi Newton-Raphson dengan kriteria AIC pada data jumlah kasus demam berdarah di Kota Makassar tahun 2018?

1.3 Batasan Masalah

Dalam penelitian ini, data yang digunakan adalah data jumlah kasus demam berdarah di Kota Makassar tahun 2018 yang mengalami overdispersi, kemudian diatasi dengan model binomial negatif dengan estimasi maksimum likelihood atau *Maximum Likelihood Estimation* (MLE) untuk menduga parameternya. Penulis tidak membahas tentang *Prior Natural Congjugate* pada regresi binomial negatif.

1.4 Tujuan Penelitian

Adapun tujuan penulisan berdasarkan masalah yang telah dirumuskan adalah untuk memperoleh penduga model regresi binomial negatif terbaik menggunakan metode estimasi maksimum likelihood (MLE) melalui iterasi Newton-Raphson dengan kriteria AIC pada data jumlah kasus demam berdarah di Kota Makassar tahun 2018.

BAB II

TINJAUAN PUSTAKA

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 (Kutner dkk., 2004).

Bentuk umum model regresi linier berganda dengan p variabel bebas adalah seperti pada persamaan (2.1) berikut (Hardle, 1995).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, ..., n \tag{2.1}$$

dengan:

y_i = variabel respon pada pengamatan ke-i, untuk $i = 1, 2, ..., n$.

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 normal yang 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 (2.1) dituliskan sebagai persamaan (2.2) berikut:

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

dengan:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Pada OLS tujuan utamanya adalah meminimumkan jumlah kuadrat sisaan.

$$\begin{aligned} S &= \sum_{j=1}^{p} \varepsilon_i^2 \\ &= \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_p^2 \\ &= \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_p \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \\ &= \boldsymbol{\varepsilon^t \varepsilon} \\ &= (\boldsymbol{Y} - \boldsymbol{X\beta})^t(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y}^t - 2\boldsymbol{X}^t\boldsymbol{Y}^t\boldsymbol{\beta} + \boldsymbol{X}^t\boldsymbol{\beta}^t\boldsymbol{X\beta} \end{aligned}$$

Untuk mengestimasi parameter regresi $(\boldsymbol{\beta})$ maka jumlah kuadrat sisaan harus diminimumkan (Supranto, 2009), hal ini dapat diproleh dari hasil turunan pertama terhadap $\boldsymbol{\beta}$ yang disamakan dengan $\mathbf{0}$.

$$\begin{aligned} \frac{\partial S}{\partial \beta|_{\beta=\hat{\beta}}} = -2\boldsymbol{X}^t\boldsymbol{Y} + 2\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \mathbf{0} \\ 2\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} &= 2\boldsymbol{X}^t\boldsymbol{Y} \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{Y}) \end{aligned}$$

Bukti bahwa $\hat{\boldsymbol{\beta}}$ merupakan estimasi linier tak bias dari $\boldsymbol{\beta}$

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{Y})] \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t E[\boldsymbol{Y}] \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Terbukti bahwa $\hat{\boldsymbol{\beta}}$ adalah estimasi tak bias dari $\boldsymbol{\beta}$ (Lains, 2003).

2.2 Model Linier Umum

Model linier umum atau *Generalized linier Model* (GLM), termasuk salah satunya regresi linier klasik dapat dinyatakan dalam model (2.1). 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 (2.1) 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:

1. Komponen Acak

   Dalam sebuah GLM, harus terdapat $y_1, y_2, \cdots, y_n$ yang merupakan variabel 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 (McCullagh and Nelder, 1989).

$$g(\mu_i) = \boldsymbol{X\beta} \tag{2.3}$$

   Fungsi transformasi g disebut dengan *link function*. Fungsi ini menentukan 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 (Cameron dan Trivedi, 1998).

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

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

disebut iterasi titik tetap.

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

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

disebut barisan iterasi Newton. Fungsi g yang di definisikan sebagai

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

disebut fungsi iterasi Newton-Raphson.

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 (Simarmata dan Ispriyanti, 2011)

Misalkan:

$\boldsymbol{Y} \sim$ Poisson(μ_i)

$\mu_i \sim$ Gamma(α, β)

Distirbusi peluang campuran Poisson-gamma dapat diperoleh dengan cara:

$$\begin{aligned} Pr(Y = y_i) &= \int_0^\infty \text{Poisson}(Y|\mu_i)\text{Gamma}(\mu_i|\alpha,\beta)d\mu_i \\ &= \int_0^\infty \frac{e^{\mu_i}\mu_i^{y_i}}{y_i!} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha}\mu_i^{\alpha-1}e^{-\frac{\mu_i}{\beta}}d\mu_i \\ &= \frac{\beta^{-\alpha}}{\Gamma(\alpha)y_i!}\int_0^\infty e^{-\mu_i(1+\frac{1}{\beta})}\mu_i^{y_i+\alpha-1}d\mu_i \end{aligned} \tag{2.4}$$

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:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{2.5}$$

atau dituliskan dalam bentuk matriks:

$$\mu_i = \boldsymbol{X\beta} \tag{2.6}$$

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 *Link*

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:

$$\begin{aligned} g(\mu_i) = \ln(\mu_i) &= \boldsymbol{X\beta} \\ \mu_i &= e^{(\boldsymbol{X\beta})} \end{aligned} \tag{2.7}$$

Misalkan Y adalah peubah acak yang bergantung pada parameter λ yang merupakan nilai dari suatu peubah acak dengan fungsi peluang $h(\lambda)$. Misal $Y|\lambda$ merupakan peubah acak yang berdistribusi Poisson dengan fungsi peluang $f(y|\lambda)$ maka:

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \tag{2.8}$$

Misalkan $h(\lambda)$ merupakan fungsi peluang dari λ yang berdistribusi gamma dengan parameter α dan β sehingga fungsi peluang sebagai berikut:

$$h(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}} \tag{2.9}$$

Jika $h(\lambda)$ berdistribusi gamma dengan parameter α dan β, Maka rataan dan variansi dari $h(\lambda)$ adalah:

$$\begin{aligned} E[h(\lambda)] &= \alpha\beta \\ Var[h(\lambda)] &= \alpha\beta^2 \end{aligned}$$

Pada umumnya rataan dinotasikan sebagai μ atau dapat ditusikan:

$$\begin{aligned} E[h(\lambda)] = \alpha\beta &= \mu \\ \beta &= \frac{\mu}{\alpha} \end{aligned}$$

Misalkan:

$$\alpha = \frac{1}{\psi}$$

Maka:

$$\beta = \mu\psi$$

dengan kata lain, $h(\lambda)$ berdistribusi gamma dengan parameter $\frac{1}{\psi}$ dan $\mu\psi$ dengan fungsi peluang:

$$h(\lambda) = \frac{1}{\Gamma\left(\frac{1}{\psi}\right)(\mu\psi)^{\frac{1}{\psi}}}\lambda^{\frac{1}{\psi}-1}e^{-\frac{\lambda}{\mu\psi}} \tag{2.10}$$

Fungsi peluang bersama antara $Y|\lambda$ dan $h(\lambda)$ adalah

$$
\begin{aligned}
f(y|\lambda, h(\lambda)) =& f(y|\lambda)h(\lambda) && (2.11)\\
=& \frac{e^{-\lambda}\lambda^{y}}{y!}\frac{1}{\Gamma\left(\frac{1}{\psi}\right)(\mu\psi)^{\frac{1}{\psi}}}\lambda^{\frac{1}{\psi}-1}e^{-\frac{\lambda}{\mu\psi}}\\
=& \frac{\lambda^{y+\frac{1}{\psi}-1}}{\Gamma\left(\frac{1}{\psi}\right)(\mu\psi)^{\frac{1}{\psi}}(y!)}e^{-y-\frac{\lambda}{\mu\psi}} && (2.12)
\end{aligned}
$$

Maka fungsi peluang marginal dari Y adalah

$$
\begin{aligned}
f(y) =& \int_{\lambda} f(y|\lambda)h(\lambda)d\lambda\\
Pr(Y=y) =& \frac{\Gamma\Big(y_i+\frac{1}{\psi}\Big)}{\Gamma(\frac{1}{\psi})y_i!}\left(\frac{\mu\psi}{1+\mu\psi}\right)^{y_i}\left(\frac{1}{1+\mu\psi}\right)^{\frac{1}{\psi}} && (2.13)
\end{aligned}
$$

Maka distribusi binomial negatif merupakan distribusi gabungan dari Poisson dan gamma dengan fungsi peluang pada persamaan (2.13).

Komponen GLM pada persamaan (2.4) dibagian pertama (komponen acak) diperolah model regresi binomial negatif sebgai berikut:

$$
\begin{aligned}
Pr(Y=y_i) =& \int_0^{\infty}\text{Poisson}(Y|\mu_i)\text{Gamma}(\mu_i|\alpha,\beta)d\mu_i\\
=& \frac{\Gamma\Big(y_i+\frac{1}{\psi}\Big)}{\Gamma(\frac{1}{\psi})y_i!}\left(\frac{\mu\psi}{1+\mu\psi}\right)^{y_i}\left(\frac{1}{1+\mu\psi}\right)^{\frac{1}{\psi}} && (2.14)
\end{aligned}
$$

dengan rataan $=\mu$ dan variansi $=\mu+\psi\mu^2$ (Loar, 2014).

Komponen GLM pada persamaan (2.7) dibagian ketiga (fungsi *link*), maka rataan dan variansi pada regresi binomial negatif dapat dituliskan kembali dalam bentuk sebagai berikut:

$$
\begin{aligned}
E(Y) &= \mu_i \\
&= e^{(\boldsymbol{X\beta})} \qquad (2.15) \\
Var(Y) &= \mu + \psi\mu^2 \\
&= e^{(\boldsymbol{X\beta})} + \psi\left(e^{(\boldsymbol{X\beta})}\right)^2 \qquad (2.16)
\end{aligned}
$$

rataan dan variansi yang diperoleh dari persamaan (2.15) dan persamaan (2.16), maka persamaan (2.14) dapat dituliskan dalam bentuk sebagai berikut:

$$
f(y) = \frac{\Gamma\left(y_i + \frac{1}{\psi}\right)}{\Gamma(\frac{1}{\psi})y_i!}\left(\frac{e^{(\boldsymbol{X\beta})}\psi}{1+e^{(\boldsymbol{X\beta})}\psi}\right)^{y_i}\left(\frac{1}{1+e^{(\boldsymbol{X\beta})}\psi}\right)^{\frac{1}{\psi}} \qquad (2.17)
$$

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

2.6 Uji Kecocokan *Chi-Square*

Variabel respon yang menyatakan banyaknya hasil dalam suatu percobaan Poisson disebut variabel acak Poisson dan distribusi peluangnya disebut distribusi Poisson. Uji kecocokan *chi-square* untuk distribusi Poisson dilakukan untuk mengetahui suatu data berdistribusi Poisson atau tidak

Pengujian hipotesis untuk uji kecocokan distribusi Poisson dilakukan sebagai berikut:

a. Hipotesis:

   H_0 : data berditribusi Poisson

   H_1 : data tidak beditribusi Poisson

b. Statistika Uji:

$$
\chi^2 = \sum_{i=1}^{n} \frac{(fo_i - fe_i)^2}{fe_i} \qquad (2.18)
$$

   dengan:

$fo_i \quad = \quad$ banyaknya frekuensi data pada kategori ke-i

$fe_i \quad = \quad$ peluang pengamatan untuk kategori ke-i

$n \quad = \quad$ banyaknya kategori

c. Kriteria Penolakan:

Tolak H_0 apabila $\chi^2 > \chi^2_{(\alpha,k-1)}$

Jika nilai $\chi^2 \leq \chi^2_{(\alpha,k-1)}$, maka data tersebut berdistribusi Poisson, begitupun dengan sebaliknya. Jika $\chi^2 > \chi^2_{(\alpha,k-1)}$, maka data tersebut tidak berdistribusi Poisson (Mutiu dkk., 2016).

2.7 Uji Linieritas

Uji linieritas adalah suatu prosedur yang digunakan untuk mengetahui linier atau tidaknya suatu distribusi. Adapun tabel anavanya sebagai berikut:

Tabel 2.1 Anava untuk uji linearitas regresi binomial negatif

Source Of Varian	**df**	**SS**	**MS**	**F**
Regression	$p-1$	$SSR = b\left[\sum \boldsymbol{XY} - \frac{\sum \boldsymbol{X} \sum \boldsymbol{Y}}{n}\right]$	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$
Error	$n-p$	$SSE = \sum \boldsymbol{Y}^2 - \frac{(\sum \boldsymbol{Y})^2}{n} - SSR$	$MSE = \frac{SSE}{n-p}$	
Lack of fit	$k-2$	$SSLF = SSE - SSPE$	$MSLF = \frac{SSLF}{k-2}$	$\frac{MSLF}{MSPE}$
Pure error	$n-k$	$SSPE = \sum_k \left[\sum Y - \frac{(\sum Y)}{n}\right]$	$MSPE = \frac{SSPE}{n-k}$	
Total	$n-1$	$SSTO = SSR + SSE$		

Adapun langkah-langkah hipotesis uji linieritas sebagai berikut:

a. Hipotesis:

H_0 : ada hubungan linier antara variabel y_i dengan $x_{ij}, j = 1, 2, \cdots p$

H_1 : tidak ada hubungan linier antara variabel y_i dengan $x_{ij}, j = 1, 2, \cdots p$

b. Statistika Uji:

$$F = \frac{MSLF}{MSPE} \tag{2.19}$$

c. Kriteria Penolakan:

   Tolak H_0 jika $F_{hitung} > F_{tabel(k-2,n-k)}$

Kelinieran pada data dipenuhi jika $F_{hitung} \leq F_{tabel}$ begitupun sebaliknya. Jika $F_{hitung} > F_{tabel}$ menunjukkan kelinieran tidak dipenuhi (Kutner dkk., 2004).

2.8 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 (Cameron dan Trivedi, 1998).

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 (McCullagh dan Nelder, 1989). Uji rasio likelihood-nya membandingkan *current model*-nya dan *saturated model*-nya yang di tuliskan sebagai berikut:

$$\begin{aligned}\phi_1 =& \frac{D^2}{db} \\ D^2 =& 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right)\end{aligned} \tag{2.20}$$

dengan:

y_i	=	Nilai variabel respon
$\hat{y}_i$	=	Estimasi regresi Poisson

dengan $db = n - p$ 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 statistik *person chi-square* (McCullagh dan Neider, 1989), yang didefinisikan sebagai:

$$\begin{aligned}\phi_2 =& \frac{\chi^2}{db} \\ \chi^2 =& \sum_{i=1}^{n}\frac{(y_i - \hat{y})^2}{\hat{y}_i}\end{aligned} \tag{2.21}$$

   dengan $db = n - p$ dengan p merupakan banyak 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 (Hilbe, 2011).

2.9 Pengujian Parameter

Persamaan yang mengandung beberapa variabel bebas dan berpengaruh terhadap variabel respon dapat dilakukan pengujian dengan *likelihood ratio test* (Greene, 2008). *Likelihood ratio test* digunakan untuk menguji estimasi parameter secara serentak, sedangkan uji wald digunakan untuk pengujian secara individu, adapun pengujian yakni sebagai berikut:

1. Uji *Likelihood Ratio*

   Uji *likelihood ratio* digunakan untuk menguji parameter secara bersama-sama. Hipotesis yang digunakan adalah sebagai berikut :

   a. Hipotesis:

      $H_0 : \beta_1 = \beta_1 = \cdots = \beta_p = 0$

      H_1 : Minimal ada satu $\beta_j \neq 0,\ j = 1, 2, \cdots, p$

   b. Statistik Uji:

      $$T = 2(l_1 - l_0)$$

      dengan

      l_o : nilai model maksimum likelihood tanpa variabel bebas tertentu

      l_1 : nilai model maksimum likelihood dengan variabel bebas tertentu

   c. Kriteria Penolakan:

      Tolak H_0 jika $T > \chi^2_{(1-2\alpha,1)}$

   T merupakan asimtotik dengan distribusi probabilitas *one-half* pada distribusi nol dan *one-half-chi-squares* dengan derajat bebasnya 1 (Cameron dan Trivedi, 1986). α adalah tingkat signifikan dan p adalah jumlah paremeter atau jika $T > \chi^2_{(1-2\alpha,p)}$ yang berarti ada salah satu atau lebih β_j yang berpengaruh pada model.

2. Uji Parsial Satu-Satu

   Uji parsial digunakan untuk pengujian individu yang menunjukkan suatu variabel bebas signifikan atau layak untuk masuk model. Pengujian masing-masing parameter yang digunakan adalah uji Wald (Cantoni dan Zedini, 2011).

   a. Hipotesis:

      $H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0,\ j = 1, 2, \cdots, p$

b. Statistik Uji:

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)^2 \tag{2.22}$$

c. Kriteria Penolakan:

Tolak H_0 jika $W_j < \chi^2_{(\alpha,p)}$

W_j distribusi *chi-square*. Jika $W_j \geq \chi^2_{\alpha,1}$ yang berarti bahwa variabel mempengaruhi model secara signifikan.

2.10 Pemilihan Model Terbaik

Pemilihan model regresi terbaik dapat dilihat dari perhitungan nilai *Akaike Information Criterion* (AIC). AIC model regresi binomial negatif (Micheal, 2013), yaitu:

$$AIC = -2[l - (p + 1)] \tag{2.23}$$

Dengan l adalah fungsi $\log$-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 (Jersey dkk., 2015).