BAB II

TINJAUAN PUSTAKA

2.1 HIV/AIDS

Human Immunodeficiency Virus (HIV) merupakan penyebab penyakit *Acquired Immunodeficiency Syndrome* (AIDS) dengan cara menyerang sel darah putih sehingga dapat merusak sistem kekebalan tubuh manusia. Kasus HIV/AIDS merupakan fenomena gunung es, dengan jumlah orang yang dilaporkan jauh lebih sedikit dibandingkan dengan yang sebenarnya. Hal ini terlihat dari jumlah kasus AIDS yang dilaporkan setiap tahunnya sangat meningkat secara signifikan. Di seluruh dunia, setiap hari diperkirakan sekitar 2000 anak di bawah 15 tahun tertular HIV dan sekitar 1400 anak di bawah usia 15 tahun meninggal dunia, serta menginfeksi lebih dari 6000 orang berusia produktif (Purwaningsih, 2008).

Virus HIV masuk ke dalam tubuh manusia melalui perantara darah, semen dan sekret vagina. *Human Immunodeficiency Virus* tergolong retrovirus yang mempunyai materi genetik RNA yang mampu menginfeksi limfosit CD4 *(Cluster Differential Four)*, dengan melakukan perubahan sesuai dengan DNA inangnya. Sindrom ini diikuti oleh penurunan jumlah CD4 dan peningkatan kadar RNA HIV dalam plasma. CD4 secara perlahan akan menurun dalam beberapa tahun dengan laju penurunan CD4 yang lebih cepat pada 1,5 – 2,5 tahun sebelum pasien jatuh dalam keadaan AIDS. *Viral load* (jumlah virus HIV dalam darah) akan cepat meningkat pada awal infeksi dan pada fase akhir penyakit akan ditemukan jumlah CD4 < 200/mm3 kemudian diikuti timbulnya infeksi oportunistik, berat badan turun secara cepat dan muncul komplikasi neurulogis. Pada pasien tanpa

pengobatan ARV, rata-rata kemampuan bertahan setelah CD4 turun < 200/mm3 adalah 3,7 tahun (Pinsky L, 2009).

Penularan HIV/AIDS akibat melalui cairan tubuh yang mengandung virus HIV yaitu melalui hubungan seksual, baik homoseksual maupun heteroseksual, jarum suntik pada pengguna narkotika, transfusi komponen darah dari ibu yang terinfeksi HIV ke bayi yang dilahirkannya (Alwi I, dkk, 2015).

2.2 Model Linier

Model yang sederhana namun efektif pada abad terakhir adalah model regresi linier**.** Model regresi linier dimulai dengan adanya kajian oleh Galton (1822-1911) yang membahas tentang hubungan tinggi badan ayah dan anaknya, dilanjutkan dengan perkembangan analisis regresi pada abad ke-19 oleh Pearson, dilanjutkan dengan perkembangan korelasi setelah itu. Teori regresi ini yang menjadi dasar perkembangan teori model linier (Tirta, 2009).

Bentuk umum regresi linier dapat di tuliskan sebagai berikut (Tirta, 2009):

$$Y = X\beta + \varepsilon \tag{2.1}$$

dengan:

Di mana Y adalah peubah tetap yang bukan acak, β merupakan parameter yang menentukan peubah tetap tadi, dan ε merupakan kesalahan atau galat yang diasumsikan merupakan peubah acak yang berasal dari suatu distribusi tertentu. Jika galat yang diasumsikan peubah acak berdistribusi Normal maka model linier (2.1) disebut model linier normal (Normal Linier Model). Pada model linier normal variabel respon hanya diasumsikan berdistribusi normal, namun dalam kenyataan variabel respon tidak hanya berdistribusi normal saja melainkan juga

berdistribusi eksponensial yang dikenal dengan *Generalized Linear Model* (GLM).

Nelder dan Wedderburn mengembangkan model linier yang dikenal dengan *Generalized Linear Model* (GLM) atau model linier tergeneralisir. Model ini mempunyai cakupan distribusi yang lebih luas, yaitu menggunakan asumsi bahwa respon berdistribusi keluarga eksponensial di mana distribusi normal termasuk di dalamnya.

Menurut Nelder dan Wedderburn (1972) GLM terdiri dari 3 komponen, yaitu:

1. Komponen acak, menentukan distribusi bersyarat dari variabel respon, di mana Y_i saling bebas atau independen. Distribusi dari Y_i adalah anggota dari keluarga eksponensial, seperti Distribusi Normal, Binomial, Poisson, Gamma, atau distribusi dari keluarga Invers-Gaussian.
2. Prediktor linier, yaitu sebuah fungsi dari $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$ di mana x_i adalah variabel prediktor untuk unit sebanyak j dengan efek acak β.
3. Fungsi *link*, yaitu fungsi yang mentransformasikan ekspektasi dari variabel respon, $\mu_i = E(Y_i)$ dengan prediktor linier:

   $$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \quad (2.2)$$

Berdasarkan uraian di atas, GLM memiliki keterbatasan yaitu hanya mampu mencocokkan model dengan variabel respon berdistribusi eksponensial dengan prediktor linier. Untuk data yang tidak linier, misalnya mengandung pencilan, GLM tidak tepat lagi diaplikasikan pada data tersebut. Oleh karena itu,

dibutuhkan model baru yang fungsi dari variabel-variabelnya tidak harus linier, melainkan mencakup pemulusan (*smoothing*) yang kemudian disebut model aditif.

2.3 *Generalized Additive Model* (GAM)

GAM adalah suatu generalisasi dari model aditif dan diperkenalkan untuk menyelesaikan masalah yang tidak dapat diselesaikan oleh model aditif. Menurut Hastie dan Tibhsirani (1986), GAM adalah suatu regresi semi parametrik karena dapat dimodelkan linier, polynomial, dan non parametrik. Model dari GAM dituliskan dalam persamaan 2.3

$$g(\mu_i) = f(X_1, X_2, \dots, X_p) = s_0 + s_1(X_1) + s_2(X_2) + \cdots + s_p(X_p) \quad (2.3)$$

Dengan $s_j(X), j = 0,1, \dots, p$ adalah fungsi *smoothing* (fungsi penghalus). Kekurangan dari GAM sendiri, yaitu GAM hanya mampu mengakomodasi distribusi keluarga eksponensial serta tidak dapat memodelkan *skewness* dan kurtosis secara langsung tetapi hanya bisa dilakukan dengan cara melalui ketergantungan *skewness* dan kurtosis terhadap μ. Oleh karena itu dibutuhkan sebuah model baru yang mampu memodelkan keempat parameter distribusi (termasuk *skewness* dan kurtosis) dan mencakup distribusi yang lebih luas dari keluarga eksponensial. (Fauziah, 2015).

2.3.1 *Skewness*

Skewness merupakan statistik yang digunakan dalam memberikan gambaran distribusi data apakah miring ke kiri atau ke kanan, atau simetris. Untuk mengukur derajat kemencengan suatu distribusi dinyatakan dengan koefisien kemencengan (koefisien *skewness*). Menurut Ramachandran dan Tsokos (2009)

skewness didefinisikan sebagai momen ke-3 standar terhadap *mean* dan dapat di lihat pada persamaan 2.4.

$$v_i = \frac{E[(X-\mu)]^3}{\sigma^3} \quad (2.4)$$

Ukuran kemiringan (*skewness*) atau ukuran ketidaksimetrisan suatu distribusi data dibagi ke dalam tiga jenis, yaitu:

a. Simetris: menunjukkan letak nilai rata-rata, median, dan modus berimpit. Salah satu contoh distribusi yang simetris adalah dstribusi normal, sehingga nilai *skewness*nya sama dengan nol, dengan *mean* = median = modus atau pada saat $v_i = 0$.

b. Menceng ke kanan: *skewness* bernilai positif di mana ujung dari kecondongan menjulur ke arah positif (ekor kurva sebelah kanan lebih panjang), di mana modus < median < *mean* atau pada saat $v_i > 0$.

c. Menceng ke kiri: *skewness* bernilai negatif di mana ujung dari kecondongan menjulur ke arah negatif (ekor kurva sebelah kiri lebih panjang), di mana *mean* < median< modus atau pada saat $v_i < 0$.

2.3.2 Kurtosis

Menurut Ramachandran dan Tsokos (2009), kurtosis adalah ukuran untuk menggambarkan keruncingan (*peakness*) atau kerataan (*flatness*) suatu distribusi data. Terdapat tiga jenis kurtosis, yaitu *leptokurtic*, *mesokurtic,* dan *platikurtic*. *Leptokurtic* yaitu bagian tengah distribusi data yang memiliki puncak yang lebih runcing (nilai keruncingan lebih dari 3), *platikurtic* bagian tengah distribusi data yang memiliki puncak yang lebih datar (nilai keruncingan kurang dari 3), dan *mesokurtic* yaitu bagian tengah distribusi data yang memiliki puncak di antara

leptokurtic dan *platikurtic*. Distribusi normal sendiri memiliki bentuk *mesokurtic* dengan nilai koefisien kurtosis sama dengan 3.

Kurtosis dimodelkan dengan momen keempat standar terhadap *mean* yang dimodelkan pada persamaan 2.5

$$\tau = \frac{E[(X-\mu)]^4}{\sigma^4} \tag{2.5}$$

Adapun 3 jenis kurtosis yang dapat diklasifikasikan sebagai berikut:

a. *Leptokurtic*: bagian tengah distribusi data memiliki puncak yang lebih runcing dengan $\tau > 3$
b. *Platikurtic*: bagian tengah distribusi data memiliki puncak yang lebih runcing dengan $\tau < 3$
c. *Mesokurtic*: bagian tengah distribusi data memiliki puncak di antara *leptokurtic* dan *platikurtic* jika $\tau = 3$.

2.4 *Generalized Additive Model for Location and Shape* (GAMLSS)

GAMLSS adalah sebuah kelas dalam model statistik yang dikembangkan oleh Rigby & Stasinopoulos yang menyediakan perluasan kemampuan dari GLM dan GAM. Pada model yang lebih sederhana menyediakan parameter lokasi yang hanya mendeskripsikan aspek yang terbatas dari distribusi dari variabel respon. Pendekatan melalui GAMLSS menyediakan parameter lain dari distribusi yang memiliki hubungan dengan variabel prediktor, dimana parameter lain tersebut diinterpretasikan sebagai parameter skala (*scale*) dan bentuk (*shape*) dari distribusi, dari variabel respon (y) yang berupa fungsi linier maupun nonlinier,

bersifat parametrik maupun non-parametrik aditif dari variabel prediktor dan efek acak.

Pada GAMLSS variabel respon berasal dari distribusi keluarga eksponensial dan tambahan distribusi-distribusi lain termasuk untuk distribusi diskrit dan kontinu dengan *highly skewed* dan kurtosis. Untuk jenis respon cacahan, metode ini cocok untuk data yang mengalami overdispersi dengan menggunakan distribusi overdispersi untuk data diskrit.

2.4.1 Bentuk dan Asumsi GAMLSS

GAMLSS mengasumsikan variabel respon y_i untuk $i = 1,2,\dots,n$ dengan fungsi kepadatan peluang $f\langle y_i|\theta^i\rangle$ dengan $\theta^i = \theta_{i1}, \theta_{i2}, \theta_{i3}, \dots, \theta_{ip}$. θ^i merupakan vektor dari 4 parameter distribusi yaitu μ, σ, v, τ yang dapat disebut sebagai fungsi variabel prediktor. Parameter μ dan σ dikarakteristikkan sebagai parameter lokasi *(location)* dan skala *(scale)*, sedangkan dua parameter lainnya yaitu v dan τ disebut sebagai parameter *skewness* (v) dan kurtosis (τ) yang tergabung dalam parameter ukuran *(shape)*.

Rigby dan Stasinopoulos (2007) mendefinisikan model dari GAMLSS sebagai berikut. Misalkan $y^T = y_1, y_2, y_3, \dots, y_n$ dengan n adalah panjang vektor dari variabel respon, k = 1,2,3,4, dan $g_k(.)$ diketahui sebagai fungsi link monotonik yang menghubungkan antara parameter distribusi dengan variabel prediktor, maka

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{Jk} Z_{jk}\gamma_{jk} \qquad (2.6)$$

Jika Z_{jk} =I_n, dengan I_n adalah matriks identitas berukuran $n \times n$ dan $\gamma_{jk} = h_{jk} = h_{jk}(x_{jk})$ untuk semua kombinasi dari j dan k, maka didapat bentuk lain dari GAMLSS yang dapat di tuliskan sebagai berikut:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{Jk} h_{jk}x_{jk} \quad (2.7)$$

$$g_i(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}x_{j1}$$

$$g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}x_{j2}$$

$$g_3(v) = \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} h_{j3}x_{j3}$$

$$g_4(\tau) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} h_{j4}x_{j4}$$

Keterangan:

$\mu, \sigma, v, \tau, \eta_k$ = vektor dengan panjang n

β_k^T = $\beta_{1k}, \beta_{2k}, \beta_{3k}, \ldots, \beta_{J'_k k}$ adalah sebuah vektor parameter

X_k = matriks berukuran $n \times J'_k$

h_{jk} = fungsi *smooth* non-parametrik dari variabel prediktor x_{jk}, $J = 1,2, \ldots, j_k$ dan $k = 1,2,3,4$.

di mana x_{jk} untuk $J = 1,2, \ldots, j_k$ juga vektor dengan panjang n. Fungsi h_{jk} adalah fungsi tak diketahui dari variabel prediktor X_k dan h_{jk} adalah sebuah vektor yang mengevaluasi fungsi h_{jk} pada (x_{jk}).

2.4.2 Algoritma Rigby dan Stasinopoulos (RS)

Ada 3 algoritma dalam GAMLSS, yaitu algoritma Rigby dan Stasinopoulos (RS), algoritma Cole dan Green (CG), dan algoritma *mixed* atau perpaduan antara RS dan CG. Sebagai algoritma dasar, RS mempunyai kelebihan dibanding dua pilihan algoritma yang lain. Selain proses perhitungannya hanya membutuhkan

waktu yang relatif singkat, algoritma ini lebih cocok untuk pengepasan semua distribusi, baik distribusi diskrit maupun kontinu (Rigby, 2005).

Rigby (2005) dalam Fauziah (2015) mendefinisikan algoritma dasar dalam GAMLSS yaitu Algoritma Rigby dan Stasinopoulos (RS) adalah sebagai berikut. Misalkan $u_k = \frac{\partial_l}{\partial_{\eta_k}}$ merupakan fungsi nilai $Z_k = \eta_k + W_{kk}^{-1}\eta_k$ dengan variabel prediktor yang dapat disesuaikan dengan W_{ks} matriks diagonal hasil dari iterasi bobot untuk $k = 1,2,\dots,p$ dan $s = 1,2,\dots,p$. Algoritma ini memiliki *outer cycle* yang dapat memaksimalkan *penalized likelihood* dengan keterkaitan β_k dan γ_{jk} untuk $J = 1,2,\dots,j_k$ dalam model berturut-turut untuk θ_k dengan $k = 1,2,\dots,p$. Setiap kalkulasi nilai yang didapatkan nilai kuantitas yang akan selalu digunakan pada setiap iterasi. Algoritma RS bukan bentuk khusus dari algoritma Cole dan Green (CG) karena dalam algoritma RS diagonal matriks berbobot W_{kk} dievaluasi (di-*update*) dalam pencocokan setiap parameter θ_k, sedangkan pada algoritma CG semua diagonal matriks berbobot W_{ks} untuk $k = 1,2,\dots,p$ dan $s = 1,2,\dots,p$.

Misalkan r adalah indeks iterasi dari *outer cycle*, k parameter indeks, I adalah indeks iterasi dari *outer cycle*, m indeks algoritma *backfitting*, dan j *random effect* (atau *nonparametric*). Misalkan $\gamma_{jk}^{r,i,m}$ merupakan nilai terbaru yang didapat dari γ_{jk} pada saat ke-r (indeks *outer cycle*), ke-i (indeks *outer cycle*), dan algoritma *backfitting* ke-m dan misalkan $\gamma_{jk}^{r,i}$ menyatakan nilai dari γ_{jk} pada saat nilai *backfitting* konvergen untuk saat ke-i dan dan ke-r dengan $j = 1,2,\dots,j_k$ dan $k = 1,2,\dots,p$.

Langkah-langkah pada Algoritma RS adalah sebagai berikut:

1. Memberikan nilai awal *fitted value* $\theta_k^{(1,1)}$ dan *random effect* $\gamma_{jk}^{1,1,1}$ untuk $j = 1,2,\dots,j_k$ dan $k = 1,2,\dots,p$.
2. Memasukkan nilai r (indeks *outer cycle*) dengan $r = 1,2,\dots$ hingga konvergen untuk $k = 1,2,\dots,p$.
   a. Memberikan nilai awal *inner cycle* $i = 1,2,\dots$ hingga konvergen.
      i. Evaluasi nilai terbaru $u_k^{(r,i)}$, $W_{kk}^{r,i}$, dan $Z_k^{r,i}$
      ii. Mulai pemberian nilai awal algoritma *backfitting* dengan $m = 1,2,\dots$ hingga konvergen
      iii. Meregresi nilai residual terbaru secara parsial dari $\varepsilon_{0k}^{r,i,m} = Z_k^{r,i} - \sum_{j=1}^{Jk} Z_{jk}\gamma_{jk}^{r,i,m}$ yang merupakan iterasi berbobot $W_{kk}^{r,i}$ untuk mendapatkan parameter estimasi terbaru $\beta_k^{r,i,m+1}$
      iv. Untuk $j = 1,2,\dots,j_k$ pemulusan parsial residual $\varepsilon_{0k}^{r,i,m} = Z_k^{r,i} - X_k\beta_k^{r,i,m+1} - \sum_{t=1,t\neq j}^{Jk} Z_{tk}\gamma_{tk}^{r,i,m}$ menggunakan *shrinking* (pemulusan) matriks S_{jk} diberikan oleh persamaan $S_{jk} = Z_{jk}^T\left(W_{kk}Z_{jk} + G_{jk}\right)^{-1}Z_{jk}^T W_{kk}$ untuk mendapatkan prediktor aditif terbaru $Z_{jk}\gamma_{jk}^{r,i,m}$
      v. *Backfitting* berakhir ketika didapat nilai yang konvergen dari $\beta_k^{r,i}$ dan $Z_{jk}\gamma_{jk}^{r,i}$ dengan $\beta_k^{r,i+1} = \beta_k^{r,i}$ dan $\gamma_{jk}^{r,i+1} = \gamma_{jk}^{r,i}$ untuk $j = 1,2,\dots,j_k$. Jika tidak, *update m* dan kembali mengulang *backfitting*
      vi. Kalkulasi $\gamma_{jk}^{r,i+1}$dan $\theta_k^{r,i+1}$ terkini.

b. *Inner cycle* berakhir dengan didapat β_k^r yang konvergen dan prediktor aditif $Z_{jk}\gamma_{jk}^r$ dengan $\beta_k^{r+i,1} = \beta_k^{r,i}, \gamma_{jk}^{r+1,i} = \gamma_{jk}^{r,.}$, dan $\theta_{jk}^{r+1,i} = \theta_k^r$ untuk $j = 1,2, \dots, j_k$.

3. *Update* nilai *k*

4. *Outer cycle* berakhir jika didapatkan (*penalized*) likelihood yang cukup kecil. Jika tidak, *update r* dan ulangi kembali ke *outer cycle*.

2.5 Metode Pemulusan *Locally Estimated Scatterplot Smoothing* (LOESS)

LOESS merupakan akronim dari *local regression*, yaitu suatu strategi untuk pemulusan kurva dari data empiris dan menyediakan bentuk rangkuman grafis hubungan antara variabel respon dan variabel prediktor (Jacoby, 2000). LOESS menambahkan *scatterplot* agar dengan mudah melihat hubungan prosedur model data statistik. Proses pemulusan dikatakan lokal karena setiap nilai yang dimuluskan ditentukan oleh titik data (x_i, y_i) yang berdekatan dalam suatu *span* (rentang), dengan i didefinisikan $i = 1, \dots, n$ dan n merupakan banyaknya data.

Cleveland (1979) menyatakan ada empat pokok yang menjadi dasar untuk melakukan pemulusan LOESS, yaitu pemilihan f $(span), W$ (fungsi bobot), d (derajat polynomial), dan t (iterasi). Parameter f merupakan *span* (rentang) untuk menentukan jumlah pemulusan yang memberikan takaran dari sebuah observasi yang digunakan setiap daerah regresi. Nilai *span* di spesifikasi antara 0 sampai 1, yang mana nilai f tersebut mempengaruhi pemulusan terhadap kurva, semakin besar nilai f, makan *fitting* kurva semakin mulus,dan berlaku untuk sebaliknya. W

merupakan fungsi bobot yang terkandung di dalam *span*, berikut fungsi bobot *tricube.*

$$W(x) = \begin{cases} (1-x^3)^3, |x| < 0 \\ 0\ , |x| \geq 0 \end{cases} \quad (2.8)$$

untuk memperoleh *local regression* yaitu dengan menghitung estimasi $\hat{\beta}_m(x_i)$ yang merupakan nilai dari $\hat{\beta}_m$ dengan meminimalkan persamaan berikut.

$$\hat{\beta} = \sum_{k=1}^{n} w_k\left(y_k - \beta_0 - \beta_1 x_k -, \ldots, \beta_d x_k^d\right)^2 \quad (2.9)$$

dengan $\hat{\beta}_m$ merupakan parameter pada regresi polynomial berderajat d dari y_k pada x_k, yang mana pencocokan dengan menggunakan bobot kuadrat terkecil dengan bobot $w_k(x_i)$ untuk (x_k, y_k) dan d merupakan parameter *degree* polinomial untuk pencocokan lokal di setiap titik pada *scatterplot,* jika yang di masukkan $d = 1$, maka polinomial yang digunakan berbentuk persamaan linier, namun jika yang di masukkan $d = 2$ polinomial yang digunakan berbentuk persamaan kuadratik. Pemulusan titik pada x_i dengan menggunakan *locally weighted regression* berderajat d adalah $(x_i, \hat{y}_i)$ dengan $\hat{y}_i$ merupakan *fitted value* dari regresi pada x_i. Sehingga dapat ditulis persamaannya sebagai berikut.

$$\hat{y}_i = \sum_{m=0}^{d} \hat{\beta}(x_i) x_i^m = \sum_{k=1}^{n} r_k\,(x_i) y_k \quad (2.10)$$

dengan r_k tidak bergantung pada y_j. Misalkan $e_i = y_i - \hat{y}_i$ dan s median dari $|e_i|$ dengan e_i merupakan *residuals* dari *fitted value*, sehingga didefinisikan *robustness weight* $\delta_k = B\left(\frac{e_k}{6s}\right)$. Perhitungan $\hat{y}_i$ baru untuk setiap i diperoleh dari *fitting* polinomial berderajat d menggunakann bobot kuadrat terkecil dengan bobot $\delta_k w_k(x_i)$ pada titik (x_k, y_k).

Proses pemulusan akan berhenti dengan beberapa iterasi t jika proses tesebut mendapatkan hasil maksimal. Jumlah iterasi yang diindikasikan memadai pada sebuah percobaan yaitu dengan dua iterasi untuk semua situasi.

2.6 Distribusi Data Cacahan pada GAMLSS

Beberapa distribusi data berpotensi untuk data cacahan atau diskrit pada GAMLSS adalah sebagai berikut:

a. Distribusi *Negative Binomial* I

Dalam Distribusi *Negative Binomial Type I* (NBI) diketahui memiliki dua parameter, yaitu μ dan σ. Di mana μ sebagai *mean* dan σ sebagai parameter dispersi bentuk umum dari fungsi kepadatan peluang dari $Y \sim NB(\mu, \alpha)$ diberikan oleh:

$$f(y|\mu,\sigma) = \frac{\Gamma(y+1/\sigma)\alpha^y}{\Gamma(y+1)\Gamma(1/\sigma)}\left[\frac{(\mu\sigma)^y}{(\mu\sigma+1)}\right]^{y+\frac{1}{\sigma}}, \text{untuk } y = 0, 1, \dots, \infty, \qquad (2.11)$$

dengan *mean* $E(y) = \mu$ dan varian $\sigma = \mu(1 + \alpha)$, sehingga saat $\alpha > 0$, varian akan melebihi *mean* dan terjadi overdispersi.

b. Distribusi *Negative Binomial* II

Dalam Distribusi *Negative Binomial Type II* (NBII) atau juga bisa disebut Binomial Negatif Kuadratik, *Negative Binomial* Kuadratik diketahui memiliki dua parameter, yaitu μ dan σ. Menurut Stasinopoulus dan Rigby (2014) fungsi kepadatan peluang dari NBII dapat dituliskan sebagai berikut:

$$f(y|\mu,\sigma) = \frac{\Gamma(y+(\mu/\sigma))\sigma^y}{\Gamma(\mu/\sigma)\Gamma(y+1)(1/\sigma)^{y+(\mu/\sigma)}}, \text{untuk } y = 0, 1, \dots, \infty, \qquad (2.12)$$

dengan $E(y) = \mu$ dan varian $\sigma = \mu + \alpha\mu^2$, sehingga saat $\alpha > 0$, maka akan terjadi overdispersi karena varian lebih besar dari *mean.*

c. Distribusi Poisson Invers Gaussian

Distribusi poisson invers gaussian merupakan salah satu distribusi *mixed poisson*. Bentuk dari distribusi *mixed poisson* tergantung pada distribusi pada random efek (v). Misalkan $g(v)$ adalah fungsi kepadatan peluang dari v dan distribusi marjinal untuk Y diperoleh dengan integral v_i:

$$P(Y = y|\mu) = \int f(y|\mu, v)g(v)dv. \quad (2.13)$$

Untuk distribusi Poisson Invers Gaussian, v diasumsikan mengikuti distribusi invers gaussian dan memiliki fungsi kepadatan peluang yang dapat ditulis sebagai berikut.

$$g(v) = (2\pi\tau v^3)^{-0,5} e^{-(v-1)^2/2\tau v}, v > 0, \quad (2.14)$$

di mana

$\tau = Var(V), E(V) = 1$

Pada akhirnya distribusi PIG dilambangkan dengan PIG(μ, τ) diberikan oleh :

$$P(y|\mu, \tau) = \left(\frac{2z}{\pi}\right)^{\frac{1}{2}} \frac{\mu^y e^{\frac{1}{\tau}} K_s(z)}{(z\tau)^y y!}, \quad (2.15)$$

di mana

$s = y - \frac{1}{2}$ dan $z = \sqrt{\frac{1}{\tau^2} + \frac{2\mu}{\tau}}$,

$K_s(z) = K_{y-\frac{1}{2}}\left(\frac{1}{\tau}\sqrt{(2\mu\tau + 1)}\right)$ adalah fungsi Bassel modifikasi jenis ketiga (Wilmott,1987).

2.7 *Akaike's Information Criterion* (AIC)

Akaike's Information Criterion (AIC) adalah metode yang berguna untuk mendapatkan model terbaik yang ditemukan oleh Akaike. AIC memperkirakan

kualitas masing-masing model relative terhadap model lain. Misalkan L adalah nilai maksimum dari fungsi *likelihood* suatu model, dan k adalah jumlah parameter yang diestimasi dalam model tersebut, maka nilai AIC dari model tersebut adalah sebagai berikut.

$$AIC = -2\ln(L(\hat{\theta})) + 2k \tag{2.16}$$

Apabila diberikan beberapa model untuk sebuah set data, maka model yang lebih baik adalah model dengan AIC terkecil (Akaike, 1978).