

Pendekatan Hibrid *Principal Component Analysis* dan Regresi Binomial Negatif untuk Pemodelan Data Cacahan: Studi Kasus Tuberkulosis di Indonesia

Nurul Azizah Muzakir^{1*}, Astri², Mukminin B³, Irwan Usman⁴

^{1,2,3} Program Studi Ilmu Aktuaria, Universitas Sulawesi Barat, Majene 91412, Indonesia

⁴ Program Studi Statistika, Universitas Sulawesi Barat, Majene 91412, Indonesia

Corresponding Email*: nurulazizahmuzakir@unsulbar.ac.id

Abstrak

Tingginya jumlah kasus Tuberkulosis (TBC) di Indonesia dan kompleksitas faktor yang memengaruhinya menuntut metode analisis yang mampu menangani hubungan antarvariabel prediktor dan karakteristik data cacahan. Penelitian ini bertujuan untuk mengatasi multikolinieritas pada variabel prediktor kasus TBC, menyusun model regresi binomial negatif berbasis komponen utama, serta mengidentifikasi faktor-faktor yang berpengaruh signifikan terhadap jumlah kasus TBC di Indonesia. Data yang digunakan berupa jumlah kasus TBC tahun 2023 di 34 provinsi Indonesia dengan 10 variabel prediktor terkait faktor kesehatan, sosial, dan demografis. Metode *Principal component analysis* (PCA) diterapkan untuk mengatasi masalah multikolinieritas variabel prediktor dan regresi binomial negatif digunakan untuk memodelkan jumlah kasus TBC. PCA menghasilkan tiga komponen utama yang menjelaskan 78% keragaman data. Ketiga komponen kemudian digunakan dalam regresi binomial negatif sebagai variabel prediktor. Hasil analisis menunjukkan bahwa model yang dibangun signifikan dan layak digunakan, dengan deviansi yang mendekati nilai ekspektasi. Secara parsial, PC1 yang merepresentasikan faktor infrastruktur dan kebijakan kesehatan berpengaruh negatif dan signifikan terhadap jumlah kasus TBC, sedangkan PC2 yang menggambarkan faktor sosial-kependudukan berpengaruh positif dan signifikan. PC3 yang mewakili faktor sosial-demografis dan layanan kesehatan lanjut usia tidak berpengaruh signifikan. Temuan ini menegaskan pentingnya peningkatan infrastruktur kesehatan dan penguatan kebijakan dalam menekan kasus TBC, serta perlunya perhatian khusus pada wilayah berpenduduk padat dengan masalah sosial tinggi. Pendekatan hibrid *PCA*–regresi binomial negatif terbukti efektif dalam mengatasi multikolinieritas dan overdispersi, sekaligus memberikan interpretasi epidemiologis yang lebih representatif.

Kata Kunci: multikolinieritas, overdispersi, *PCA*, regresi binomial negatif, tuberkulosis

Abstract

The high number of tuberculosis (TB) cases in Indonesia and the complexity of the factors that influence them require analytical methods capable of handling the relationships between predictor variables and count data characteristics. This study aims to overcome multicollinearity in TB case predictor variables, construct a negative binomial regression model based on principal components, and identify factors that significantly influence the number of TB cases in Indonesia. The data used consists of the number of TB cases in 34 provinces in Indonesia with 10 predictor variables related to health, social, and demographic factors. Principal component analysis (PCA) was used in this study to address the problem of multicollinearity of predictor variables, and a negative binomial regression model was used to model the number of TB cases. PCA produced three principal components that explained 78% of the data variance. The three components were then used in negative binomial regression as predictor variables. The results showed that the model built was significant and feasible for use, with deviations close to the expected values. Partially, PC1, which represents infrastructure and health policy factors, had a negative and significant effect on the number of TB cases, while PC2, which describes social and population density factors, had a positive and significant effect. PC3 had no significant effect. These findings emphasize the importance of improving health infrastructure and strengthening policies to reduce TB cases, as well as the need for special attention to densely populated areas with high social problems. The *PCA*–Negative Binomial hybrid approach has proven effective in overcoming multicollinearity and overdispersion, while providing a more representative epidemiological interpretation.

Keywords: multicollinearity, negative binomial regression, overdispersion, *PCA*, tuberculosis

Received :05-02-2026 Revised :12-03-2026 Accepted :03-04-2026 Published :30-04-2026

1. Pendahuluan

Penyakit menular masih menjadi salah satu tantangan utama dalam kesehatan masyarakat global, terkhususnya di negara berkembang di mana perbedaan faktor sosial, ekonomi, dan lingkungan memengaruhi penyebaran penyakit. Salah satu penyakit menular yang masih menjadi perhatian serius adalah tuberkulosis (TBC), yaitu penyakit infeksi yang menyerang paru-paru dan ditularkan melalui udara oleh *Mycobacterium tuberculosis*. Indonesia termasuk negara dengan angka TBC nasional yang tinggi, menurut Organisasi Kesehatan Dunia [1]. Dengan perkiraan ratusan ribu kasus baru setiap tahunnya, Indonesia merupakan salah satu dari tiga negara dengan jumlah kasus tuberkulosis terbesar di dunia, bersama Tiongkok dan India. Selain tingkat insiden yang tinggi, Indonesia juga memiliki masalah signifikan berupa kasus tuberkulosis laten yang tidak terdiagnosis, meningkatnya resistensi obat, terbatasnya akses terhadap layanan kesehatan, dan rendahnya tingkat keberhasilan pengobatan di beberapa daerah. Meskipun telah dilakukan berbagai upaya pengendalian, jumlah kasus TBC di Indonesia masih signifikan, sehingga membutuhkan strategi berbasis komunitas untuk menjangkau semua lapisan masyarakat [2]. Kondisi ini menyoroti pentingnya metode analisis yang lebih akurat untuk mengidentifikasi faktor-faktor yang berdampak pada fluktuasi jumlah kasus.

Secara statistik, jumlah kasus penyakit menular merupakan bilangan bulat dan non-negatif, data ini diklasifikasikan sebagai data hitungan dalam statistik. Regresi Poisson, yang mengasumsikan bahwa rerata dan variansinya sama, merupakan model yang populer untuk menganalisis data jenis ini. Namun, karena variansi data seringkali lebih tinggi daripada rerata dalam epidemiologi, asumsi ini sering kali tidak terpenuhi [3]. Kondisi variansi data cacahan melebihi nilai reratanya dikenal sebagai overdispersi. Overdispersi dapat mengakibatkan estimasi parameter tidak efisien dan kesalahan baku menjadi bias [4]. Regresi binomial negatif memberikan solusi yang lebih adaptif untuk masalah overdispersi. Model ini lebih cocok untuk data jumlah yang terdispersi berlebih karena memiliki parameter dispersi yang memungkinkan variansi lebih besar daripada rerata. Hasil dari beberapa penelitian mengungkapkan bahwa regresi binomial negatif dapat menawarkan pemodelan yang lebih stabil dan tepat dalam konteks data kesehatan masyarakat [5], [6]. Syafiqoh *et al.* [7] memodelkan faktor risiko TBC di Provinsi Jawa Barat dengan menggunakan regresi binomial negatif dan mendapati bahwa model tersebut lebih cocok daripada regresi Poisson karena adanya overdispersi.

Selain overdispersi, isu lain yang kerap timbul dalam analisis regresi adalah multikolinieritas, yang terjadi ketika variabel prediktor saling berkorelasi kuat antara satu sama lain. Fenomena ini cukup umum terjadi pada data yang berkaitan dengan kondisi sosial, ekonomi, atau lingkungan karena banyak variabel yang saling berkaitan erat. Keberadaan multikolinieritas dapat membuat estimasi koefisien tidak stabil dan sulit diinterpretasikan. Menurut Navelski dan Odongo [8], penggunaan *Principal Component Analysis* (PCA) merupakan salah satu cara yang efisien untuk menyelesaikan masalah multikolinieritas dengan mentransformasi variabel asli menjadi komponen-komponen baru yang umumnya tidak saling berkaitan. Muhaa *et al.* [9] menggunakan PCA dalam tren global dalam proses klasifikasi tuberkulosis dan menemukan bahwa metode reduksi dimensi ini dapat mentransformasi variabel epidemiologi menjadi komponen yang lebih informatif.

Penggunaan PCA dan regresi binomial negatif secara terpisah telah banyak diterapkan dalam analisis data kesehatan [10]-[13]. Namun, kajian yang mengintegrasikan kedua pendekatan tersebut untuk memodelkan jumlah kasus TBC, terutama terkait dengan data TBC di Indonesia, masih sangat minim. Secara metodologis, integrasi ini memiliki keunggulan karena regresi binomial negatif mampu menangani permasalahan overdispersi, sementara PCA berperan dalam mengatasi multikolinieritas melalui transformasi variabel prediktor menjadi komponen yang saling bebas. Dengan demikian, pendekatan gabungan PCA dan regresi binomial negatif berpotensi menghasilkan model yang lebih stabil, andal, dan akurat dalam memodelkan jumlah kasus TBC.

Dengan memperhatikan hal tersebut, studi ini bertujuan untuk mengatasi masalah multikolinieritas pada data TBC di Indonesia menggunakan PCA, membangun model regresi binomial negatif untuk jumlah kasus TBC di Indonesia menggunakan komponen utama sebagai variabel prediktor, dan mengidentifikasi

Muzakir, dkk (2026)

faktor-faktor yang berpengaruh signifikan terhadap jumlah kasus TBC berdasarkan pendekatan integrasi PCA dan regresi binomial negatif. Pendekatan ini diharapkan dapat menghasilkan model yang lebih *robust* melalui penanganan masalah overdispersi dan multikolinieritas secara bersamaan. Temuan dari studi ini diharapkan mampu memberikan sumbangsih pada kemajuan metode pemodelan statistik untuk data cacahan dalam bidang epidemiologi, serta menyediakan dasar ilmiah yang kokoh untuk pengembangan kebijakan penanganan penyakit yang lebih efisien dan didasarkan pada data.

2. Metode

2.1 Sumber Data

Penelitian ini menggunakan data sekunder berupa jumlah kasus terkonfirmasi TBC di setiap provinsi di Indonesia tahun 2023 serta beberapa variabel yang diduga berpengaruh terhadap angka kasus TBC. Data didapatkan dari sumber resmi, yaitu Badan Pusat Statistik (BPS) dan publikasi Profil Kesehatan Indonesia oleh Kementerian Kesehatan tahun 2023. Adapun analisis data dilakukan menggunakan *software R*.

2.2 Variabel Penelitian

Data yang digunakan bersifat *cross-sectional* dengan cakupan 34 provinsi di Indonesia. Variabel respons dalam penelitian ini adalah jumlah kasus terkonfirmasi Tuberkulosis (TBC) pada tahun 2023 di setiap provinsi dan terdapat 10 variabel prediktor yang diduga memengaruhi jumlah kasus TBC, yang mencakup aspek demografi, sosial, ekonomi, dan kesehatan. Tabel 1 menampilkan variabel yang digunakan.

Tabel 1. Variabel penelitian

Variabel	Nama Variabel	Definisi	Satuan
Y	Jumlah kasus	Jumlah kasus terkonfirmasi TBC setiap provinsi	Kasus
X_1	Persentase perokok	Hasil bagi antara penduduk berumur 15 tahun ke atas yang merokok tembakau dengan jumlah penduduk di setiap provinsi	Persen
X_2	Jumlah kabupaten/kota yang menerapkan kebijakan germas	Jumlah kabupaten/kota yang menerapkan kebijakan gerakan masyarakat hidup sehat (germas)	Jumlah
X_3	Jumlah kasus HIV	Jumlah kasus positive HIV di setiap provinsi	kasus
X_4	Jumlah kabupaten/kota yang menerapkan KTR	Jumlah kabupaten/kota yang menerapkan kebijakan kawasan tanpa rokok (KTR)	Jumlah
X_5	Jumlah penyalahgunaan NAPZA	Jumlah penduduk yang menyalahgunakan NAPZA	Jumlah
X_6	Persentase pelayanan kesehatan lanjut usia	Hasil bagi antara jumlah lanjut usia yang mendapatkan pelayanan kesehatan dengan penduduk usia 60 ke atas	Persen
X_7	Jumlah puskesmas	Jumlah puskesmas di setiap provinsi	Unit
X_8	Kepadatan penduduk	Hasil bagi antara jumlah penduduk dengan luas wilayah di setiap provinsi	Jiwa/ km^2
X_9	Persentase penduduk lansia	Hasil bagi antara jumlah penduduk berusia 60 tahun ke atas dengan jumlah penduduk di setiap provinsi	persen
X_{10}	Jumlah kabupaten/kota sehat	Jumlah kabupaten/kota sehat setiap provinsi	jumlah

2.3 Koefisien Korelasi Rank Spearman

Korelasi Rank Spearman merupakan metode nonparametrik yang digunakan untuk mengukur kekuatan hubungan antara dua variabel berdasarkan peringkat data, bukan nilai aslinya. Oleh karena itu, setiap pengamatan perlu diubah menjadi bentuk peringkat sebelum dilakukan perhitungan korelasi. Metode ini tidak mensyaratkan asumsi distribusi tertentu dan tepat digunakan ketika data tidak menyebar normal atau mengandung pencilan [14]. Secara matematis, koefisien korelasi sampel Rank Spearman, dinotasikan dengan r_s , didefinisikan sebagai korelasi Pearson antara dua peringkat dari dua variabel. Untuk data tanpa peringkat yang sama (*ties*), koefisien Spearman dapat dihitung menggunakan formula berikut:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \quad (1)$$

di mana d_i adalah selisih peringkat antara pasangan pengamatan ke- i dan n adalah jumlah pengamatan. Nilai koefisien Spearman berada pada interval $-1 \leq r_s \leq 1$, di mana nilai mendekati 1 mencerminkan adanya hubungan monoton positif yang sangat kuat, nilai yang mendekati -1 mencerminkan adanya hubungan monoton negatif yang sangat kuat, sedangkan nilai yang mendekati 0 mencerminkan bahwa tidak ada hubungan monoton yang signifikan [15].

2.4 Principal Component Analysis (PCA)

Metode PCA digunakan untuk menyederhanakan sejumlah variabel yang diamati melalui proses reduksi dimensi. Proses tersebut dilakukan dengan mentransformasi variabel asal menjadi sejumlah variabel baru yang tidak saling berkorelasi, sehingga hubungan antarvariabel dapat dihilangkan. Setelah didapatkan beberapa komponen utama yang bebas dari masalah multikolinearitas, komponen-komponen tersebut selanjutnya digunakan sebagai variabel prediktor dalam analisis regresi untuk mengkaji pengaruhnya terhadap variabel respons (Y). Salah satu kelebihan metode PCA adalah kemampuannya dalam mengatasi korelasi antarvariabel tanpa perlu mengeliminasi variabel awal [16]. Tahapan penerapan metode PCA dijabarkan sebagai berikut.

a) Uji Kaiser-Meyer-Olkin (KMO) dan Barlett's test of sphericity

Dalam PCA, uji Kaiser-Meyer-Olkin (KMO) dan Barlett's test of sphericity digunakan untuk mengevaluasi kelayakan data sebelum proses reduksi dimensi dilakukan. Kedua pengujian ini bertujuan untuk memastikan bahwa struktur korelasi antarvariabel cukup memadai hingga PCA dapat menghasilkan komponen yang bermakna. KMO mengukur kelayakan sampel (*sampling adequacy*) dengan membandingkan nilai koefisien korelasi parsial terhadap koefisien korelasi total. Nilai statistik KMO berada pada interval 0 hingga 1. Nilai KMO yang mendekati 1 mengindikasikan bahwa pola hubungan antarvariabel sudah memadai, sehingga PCA bisa diterapkan. Secara umum, nilai KMO $> 0,5$ dianggap sebagai batas minimum kelayakan. Statistik KMO diberikan sebagai [17]:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2} \quad (2)$$

dengan r_{ij} menyatakan koefisien korelasi sederhana antara variabel ke- i dan ke- j , dan p_{ij} merupakan koefisien korelasi parsial antara kedua variabel tersebut.

Sementara itu, Barlett's test of sphericity digunakan untuk menilai apakah matriks korelasi berbeda secara signifikan dari matriks identitas. Hipotesis yang diuji dalam uji Barlett adalah:

$H_0: \rho = I$ (matriks korelasi identik dengan matriks identitas)

$H_1: \rho \neq I$ (matriks korelasi tidak sama dengan matriks identitas)

Adapun statistik uji Bartlett dirumuskan sebagai [18]:

$$\chi^2 = - \left(n - 1 - \frac{2p + 5}{6} \right) \ln [\det(R)] \quad (3)$$

dengan n menyatakan banyaknya pengamatan, p adalah banyaknya variabel, dan R adalah matriks korelasi berukuran $p \times p$. Statistik uji ini diaproksimasikan berdistribusi *chi-square* dengan derajat bebas $df = \frac{p(p-1)}{2}$. Dengan taraf signifikansi α , keputusan pengujian diambil dengan menolak H_0 apabila nilai $\chi^2 > \chi_{\alpha, df}^2$ atau jika $p - value < \alpha$. Penolakan H_0 menunjukkan bahwa terdapat korelasi yang signifikan antarvariabel, sehingga data dinyatakan memenuhi syarat untuk dianalisis menggunakan PCA.

b) Standarisasi data

Jika variabel memiliki satuan yang berbeda, data perlu distandarisasi agar setiap variabel memiliki kontribusi yang sebanding. Standarisasi dilakukan dengan mentransformasi data awal menjadi data baru dengan:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (4)$$

di mana x_{ij} adalah nilai pengamatan ke- i pada variabel ke- j , \bar{x}_j adalah rerata variabel ke- j , dan s_j adalah simpangan baku variabel ke- j .

c) Pembentukan matriks korelasi

Setelah data distandarisasi, dibentuk matriks korelasi. Untuk data yang telah distandarisasi, matriks korelasi didefinisikan sebagai [18]:

$$R = \frac{1}{n-1} Z^T Z \quad (5)$$

dengan Z adalah matriks data terstandarisasi dan n adalah jumlah pengamatan.

d) Dekomposisi eigen

Langkah selanjutnya adalah melakukan dekomposisi eigen terhadap matriks korelasi untuk memperoleh nilai eigen (λ_k) dan vektor eigen (\mathbf{e}_k) sebagai berikut [18]:

$$R\mathbf{e}_k = \lambda_k \mathbf{e}_k \quad (6)$$

Nilai eigen menyatakan besar ragam yang dijelaskan oleh masing-masing komponen utama, sedangkan vektor eigen menentukan arah komponen tersebut.

e) Pembentukan komponen utama

Komponen utama ke- k (PC_k) diperoleh sebagai kombinasi linier dari variabel asal:

$$PC_k = Z\mathbf{e}_k = e_{1k}z_1 + e_{2k}z_2 + \dots + e_{pk}z_p \quad (7)$$

dengan p menyatakan banyaknya variabel dan e_{jk} menyatakan elemen ke- j dari vektor eigen ke- k .

f) Penentuan jumlah komponen utama

Dalam PCA, penentuan banyaknya komponen utama yang dipertahankan umumnya didasarkan pada besarnya proporsi ragam kumulatif. Secara matematis, misalkan diperoleh sejumlah p nilai eigen dari matriks korelasi, yaitu $\lambda_1, \lambda_2, \dots, \lambda_p$, yang telah diurutkan dari terbesar ke terkecil. Besarnya kontribusi ragam yang dapat dijelaskan oleh komponen utama ke- k didefinisikan sebagai berikut.

$$Proporsi\ ragam_k = \frac{\lambda_k}{\sum_{j=1}^p \lambda_j} \quad (8)$$

Jumlah komponen utama dipilih sedemikian rupa sehingga proporsi ragam kumulatif mencapai atau melampaui ambang batas tertentu. Proporsi ragam kumulatif di atas 70% umumnya dianggap memadai untuk analisis eksploratif [19].

2.5 Model Regresi Binomial Negatif

Regresi binomial negatif merupakan salah satu bentuk penerapan dari Generalisasi Model Linear (GML), mengingat distribusi binomial negatif termasuk dalam keluarga distribusi eksponensial. Model ini secara khusus digunakan untuk memodelkan data dengan variabel respons yang berbentuk data cacahan. Jika model regresi Poisson menunjukkan adanya overdispersi, maka regresi binomial negatif dapat dipilih sebagai pendekatan yang lebih sesuai. Variabel respons pada regresi binomial negatif diasumsikan mengikuti distribusi binomial negatif $Y_i \sim BN(\mu_i, k)$. Fungsi *link* pada regresi binomial negatif dinyatakan sebagai $\mu(X_i, \beta) = \exp\{\mu(X_i^T \beta)\}$, di mana X_i merupakan matriks variabel prediktor berukuran $n \times (p+1)$, β merupakan vektor parameter regresi berdimensi $(p+1) \times 1$, dan $X_i \beta$ merupakan fungsi linear. Dengan demikian, penduga bagi μ_i , yaitu $\hat{\mu}_i$ adalah:

$$\hat{\mu}_i = \hat{\mu}(X_i, \beta) = \exp\{X_i^T \beta\} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}\} \quad (9)$$

di mana $\hat{\beta}$ merupakan penduga kemungkinan maksimumkan bagi β [20]. Berdasarkan persamaan, maka model umum regresi binomial negatif dinyatakan sebagai:

$$\hat{y}_i = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}\} \quad (10)$$

2.6 Estimasi Parameter Model Regresi Binomial Negatif

Metode estimasi *Maximum Likelihood* (MLE) adalah metode yang sering diterapkan dalam pemodelan regresi binomial negatif karena sifatnya yang efisien dan konsisten. Misalkan y_i menyatakan jumlah kejadian pada pengamatan ke- i dengan nilai rerata μ_i . Fungsi *likelihood* dari sebaran binomial negatif dengan parameter dispersi ϕ diberikan sebagai berikut [21]:

$$L(\beta, \phi | y, x) = \prod_{i=1}^n \left\{ \frac{\Gamma(y_i + \phi^{-1})}{\Gamma(\phi^{-1}) y_i!} \left(\frac{\phi \mu_i}{1 + \phi \mu_i} \right)^{y_i} \left(\frac{1}{1 + \phi \mu_i} \right)^{\phi^{-1}} \right\} \quad (11)$$

Selanjutnya, fungsi *log-likelihood* dari model regresi binomial negatif diperoleh dengan mengambil logaritma natural dari fungsi *likelihood*, sehingga diperoleh:

$$\ln L(\beta, \phi | y, x) = \sum_{i=1}^n \left\{ \sum_{r=1}^{y_i-1} \ln(1 + \phi r) \right\} - y_i \ln(\phi) - \ln(y_i!) + y_i \ln(\phi \mu_i) - (y_i + \phi^{-1}) \ln(1 + \phi \mu_i) \quad (12)$$

Parameter ϕ merupakan parameter dispersi yang berperan dalam mengakomodasi overdispersi pada data cacahan. Variansi dari peubah acak Y_i dalam model binomial negatif dinyatakan sebagai:

$$\text{Var}(Y_i) = \mu_i + \phi \mu_i^2 \quad (13)$$

yang menunjukkan bahwa variansi meningkat lebih cepat dibandingkan rerata, sehingga model ini lebih fleksibel dibandingkan regresi Poisson. Estimasi parameter β dan ϕ dilakukan dengan memaksimalkan fungsi *log-likelihood* menggunakan metode numerik iteratif, karena solusi analitik tertutup tidak tersedia. Prosedur optimasi yang umum digunakan antara lain algoritma *Newton-Raphson* atau *Fisher Scoring*. Ketepatan estimasi parameter selanjutnya dievaluasi melalui simpangan baku yang diperoleh dari invers matriks informasi Fisher.

2.7 Uji Signifikansi Parameter Secara Parsial

Setelah parameter dari model regresi binomial negative diestimasi, tahap berikutnya adalah menguji signifikansi tiap parameter. Untuk pengujian secara parsial digunakan uji Wald. Hipotesis yang dirumuskan sebagai berikut:

Muzakir, dkk (2026)

$H_0: \beta_j = 0$ (variabel prediktor tidak memberikan pengaruh signifikan terhadap variabel respons)

$H_1: \beta_j \neq 0$ (variabel prediktor memberikan pengaruh signifikan terhadap variabel respons)

Statistik uji Wald dapat dihitung menggunakan formula berikut:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (14)$$

dengan $\hat{\beta}_j$ merupakan penduga untuk parameter β_j dan $SE(\hat{\beta}_j)$ merupakan penduga *standard error* $\hat{\beta}_j$. Statistik uji W berdistribusi *chi-square*, sehingga nilainya dibandingkan dengan $\chi_{\alpha,df=1}^2$. Kriteria pengambilan keputusan adalah H_0 ditolak jika nilai $W > \chi_{\alpha,df=1}^2$ dengan α menyatakan taraf signifikansi dan derajat bebas yang digunakan adalah satu [21].

2.8 Uji Signifikansi Parameter Secara Simultan

Untuk mengetahui apakah parameter regresi memiliki pengaruh secara keseluruhan, dilakukan pengujian simultan terhadap seluruh koefisien regresi. Pengujian ini dapat menggunakan uji *Likelihood Ratio*. Hipotesis yang diuji adalah:

$H_0: \beta_1 = \beta_2 = \dots = \beta_j = 0$

H_1 : terdapat minimal satu $\beta_j \neq 0$

Statistik uji yang digunakan dirumuskan sebagai berikut:

$$G = -2 \ln \left[\frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \right] = -2 [\ln L(\hat{\theta}_0) - \ln L(\hat{\theta})] \quad (15)$$

di mana $\ln L(\hat{\theta}_0)$ merepresentasikan *log-likelihood* dari model yang hanya memuat intersep (tanpa variabel predictor), sedangkan $\ln L(\hat{\theta})$ merepresentasikan *log-likelihood* dari model lengkap yang mencakup intersep dan seluruh variabel predictor. Untuk kriteria pengambilan keputusan, jika $G \geq \chi_{\alpha,df}^2$, maka H_0 ditolak. Keputusan ini menunjukkan bahwa setidaknya terdapat satu koefisien β_j yang bernilai tidak nol, sehingga dapat disimpulkan bahwa variabel predictor X_j memiliki pengaruh signifikan terhadap variabel respons [22].

2.9 Uji Kelayakan Model

Uji kelayakan model dilakukan untuk menilai kecocokan model regresi terhadap data pengamatan. Salah satu ukuran yang umum digunakan adalah *deviance*, yaitu ukuran jarak antara model yang diusulkan (*current model*) dan model jenuh (*saturated model*). *Deviance* didefinisikan sebagai:

$$D = 2 \{ (\log L(y; \alpha) - \log L(\mu; \alpha)) \} \quad (16)$$

Nilai *deviance* yang kecil berarti bahwa model yang diusulkan memiliki kesesuaian yang baik terhadap data. Pada regresi binomial negatif, statistik *deviance* dapat dinyatakan sebagai [23]:

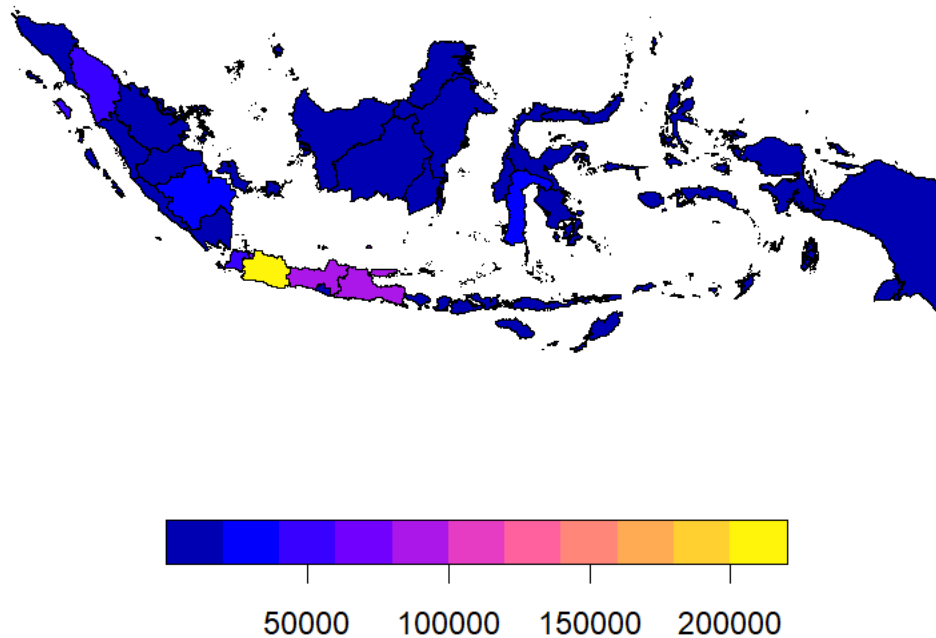
$$D = 2 \sum_{i=1}^n \left\{ y_i \log \left(\frac{y_i}{\mu_i} \right) - (y_i + a^{-1}) \log \left(\frac{1 + ay_i}{1 + a\mu_i} \right) \right\} \quad (17)$$

Secara asimtotik, *deviance* mengikuti distribusi *chi-square* dengan derajat bebas sebesar $n - p$. Model dinyatakan layak apabila nilai *deviance* tidak signifikan secara statistik.

3. Hasil dan Pembahasan

3.1 Deskripsi Data

Data yang digunakan dalam penelitian ini mencakup 34 provinsi di Indonesia dengan variabel respons berupa jumlah kasus terkonfirmasi Tuberkulosis (TBC) di setiap provinsi di Indonesia tahun 2023. Terdapat 10 variabel prediktor yang diduga memengaruhi jumlah kasus TBC. Peta sebaran kasus TBC di Indonesia tahun 2023 ditampilkan pada Gambar 1.



Gambar 1. Peta sebaran jumlah kasus TBC di Indonesia

Pada Gambar 1, warna pada peta menunjukkan intensitas kasus dengan gradasi dari biru (rendah) hingga kuning (tinggi). Semakin terang (menuju kuning) warnanya, semakin tinggi jumlah kasus TBC yang terkonfirmasi pada provinsi tersebut. Berdasarkan peta, terlihat bahwa Provinsi Jawa Barat memiliki intensitas warna kuning paling terang, menandakan jumlah kasus TBC tertinggi nasional, yaitu 212.316 kasus. Jawa Timur dan Jawa Tengah menunjukkan warna ungu yang menandakan jumlah kasus juga relatif tinggi, yaitu masing-masing sebesar 86.869 dan 85.099 kasus. Hal ini sejalan dengan kepadatan penduduk di Pulau Jawa yang merupakan wilayah dengan populasi terbesar dan risiko penularan TBC lebih tinggi akibat kepadatan hunian dan aktivitas sosial-ekonomi yang intens. Beberapa provinsi seperti DKI Jakarta, Banten, Sumatera Utara, Sulawesi Selatan, dan Sumatera Selatan menunjukkan warna biru keunguan yang mengindikasikan jumlah kasus TBC menengah (sekitar 20.000–70.000 kasus). Sebaliknya, Sebagian besar provinsi di Kalimantan, Sulawesi bagian tengah dan utara, Maluku, serta Papua dan Papua Barat menunjukkan warna biru tua, menandakan jumlah kasus relatif rendah (<10.000 kasus). Statistik deskriptif dari variabel respons dan prediktor disajikan pada Tabel 2.

Tabel 2. Statistik deskriptif jumlah kasus TBC di Indonesia

Variabel	Minimum	Maksimum	Rata-rata	Variansi
Y	1982	212613	23753	1610239676
X_1	18,90	34,08	27,24	11,46
X_2	1,00	38,00	11,18	83,06
X_3	75,0	9500	1610,8	4984688
X_4	3,00	35,00	13,24	62,85
X_5	0	1844	396,2	247607,9
X_6	42,00	93,00	73,03	176,27
X_7	44,00	1100	286,1	65102,01
X_8	8,00	17153	783	8527684
X_9	6,20	16,02	10,25	5,60
X_{10}	0	38,00	11,12	76,96

Berdasarkan Tabel 2, jumlah kasus TBC antarprovinsi (Y) memiliki keragaman yang sangat tinggi

dengan nilai minimum 1.982 kasus dan maksimum 212.613 kasus, serta rata-rata 23.753 kasus. Variabel persentase perokok (X_1) menunjukkan rata-rata sebesar 27,24% dengan variasi yang relatif kecil, menandakan perilaku merokok cukup merata di seluruh provinsi. Sebaliknya, jumlah kabupaten/kota yang menerapkan kebijakan GERMAS (X_2) dan kawasan tanpa rokok (X_4) memperlihatkan variasi yang cukup besar. Hal ini mencerminkan perbedaan tingkat implementasi kebijakan kesehatan antarprovinsi. Jumlah kasus HIV (X_3) juga sangat bervariasi. Selain itu, jumlah penyalahgunaan NAPZA (X_5) menunjukkan rentang yang lebar antarprovinsi. Dari sisi pelayanan kesehatan, persentase pelayanan kesehatan lansia (X_6) relatif tinggi dan merata, sementara jumlah puskesmas (X_7) menunjukkan variasi besar sesuai dengan luas dan jumlah penduduk provinsi. Variabel kepadatan penduduk (X_8) memiliki perbedaan ekstrem antara provinsi padat seperti DKI Jakarta dan provinsi dengan kepadatan rendah seperti Papua Barat, yang berpotensi memengaruhi tingkat penularan TBC. Adapun persentase penduduk lansia (X_9) relatif stabil di seluruh provinsi, sedangkan jumlah kabupaten/kota sehat (X_{10}) masih bervariasi, menandakan bahwa upaya menciptakan lingkungan sehat belum merata di Indonesia.

3.2 Analisis Korelasi Antarvariabel Prediktor

Analisis korelasi dilakukan untuk mengetahui tingkat kekuatan hubungan antarvariabel prediktor. Hubungan antarvariabel pada penelitian ini dianalisis menggunakan koefisien korelasi Rank Spearman, karena sebagian besar variabel prediktor tidak berdistribusi normal sehingga metode nonparametrik ini lebih sesuai digunakan. Dua variabel dianggap memiliki hubungan yang kuat apabila nilai koefisien korelasi $|r_s| \geq 0,6$. Matriks korelasi divisualisasikan melalui Gambar 2 berikut.



Gambar 2. Matriks korelasi antarvariabel prediktor

Hasil analisis korelasi Rank Spearman menunjukkan bahwa terdapat beberapa variabel prediktor memiliki hubungan kuat, di antaranya $X_2 \leftrightarrow X_4$ ($r_s = 0,78$), $X_2 \leftrightarrow X_7$ ($r_s = 0,74$), $X_2 \leftrightarrow X_{10}$ ($r_s = 0,76$), $X_4 \leftrightarrow X_7$ ($r_s = 0,92$), $X_4 \leftrightarrow X_{10}$ ($r_s = 0,8$), $X_7 \leftrightarrow X_{10}$ ($r_s = 0,73$), dan $X_8 \leftrightarrow X_9$ ($r_s = 0,72$). Kondisi ini menandakan adanya multikolinearitas, sehingga diperlukan reduksi dimensi menggunakan analisis komponen utama (PCA) sebelum membangun model regresi.

3.3 Principal Component Analysis

Untuk memastikan bahwa PCA dapat diterapkan, perlu dilakukan dua uji statistik, yaitu uji KMO dan uji Bartlett. Kelayakan penerapan PCA dievaluasi menggunakan uji KMO dengan melihat kecukupan pola korelasi antarvariabel. Di sisi lain, uji Bartlett digunakan untuk menilai apakah matriks korelasi secara

signifikan berbeda dari matriks identitas yang menunjukkan tidak adanya hubungan antarvariabel. Hasil dari kedua pengujian tersebut ditampilkan pada Tabel 3.

Tabel 3. Hasil uji KMO dan uji Bartlett

Pengujian	Statistik Uji	Nilai	Derajat Bebas	<i>p-value</i>	Kriteria Kelayakan
Kaiser-Meyer-Olkin (KMO)	KMO	0,65	-	-	$KMO > 0,5$
Bartlett's test of sphericity	Chi-Square	226,85	45	$4,25 \times 10^{-26}$	$p < 0,05$

Hasil perhitungan menunjukkan nilai statistik KMO adalah 0,65, yang mengindikasikan bahwa tingkat kecukupan sampel tergolong memadai untuk penerapan PCA. Dengan demikian, korelasi antarvariabel prediktor dinilai cukup kuat untuk diekstraksi menjadi komponen utama. Selanjutnya, uji Bartlett menghasilkan nilai *chi-square* sebesar 226,85 dengan *p-value* $< 0,05$, sehingga hipotesis nol yang menyatakan bahwa matriks korelasi adalah matriks identitas ditolak. Hal ini menandakan adanya hubungan yang signifikan antarvariabel prediktor, sehingga data layak untuk dianalisis menggunakan PCA.

PCA dilakukan terhadap 10 variabel prediktor yang telah distandarisasi. Hasil perhitungan nilai eigen dan proporsi variansi (keragaman) yang dijelaskan disajikan pada Tabel 4.

Tabel 4. Nilai eigen dan proporsi keragaman yang dijelaskan oleh komponen utama

Komponen	Nilai eigen	Proporsi variansi (%)	Kumulatif (%)
PC1	4,999	49,9	49,9
PC2	1,727	17,3	67,2
PC3	1,083	10,8	78
PC4	0,794	8	86
PC5–PC10	$< 0,7$	< 7	100

Berdasarkan Tabel 4, dipilih komponen utama yang memiliki nilai eigen > 1 dan proporsi variansi kumulatif minimal 70%. Berdasarkan kriteria tersebut diperoleh tiga komponen utama, yaitu PC1, PC2, dan PC3, yang secara kumulatif menjelaskan 78% keragaman total data. Selanjutnya, menghitung matriks *loading* yang merupakan korelasi antar setiap variabel dengan masing-masing komponen utama. Matriks *loading* disajikan melalui Tabel 5.

Tabel 5. Matriks *loading*

Variabel	PC1	PC2	PC3
X_1	-0,340	-0,411	0,745
X_2	-0,904	-0,215	-0,112
X_3	-0,795	0,330	-0,271
X_4	-0,917	-0,214	-0,084
X_5	-0,421	0,613	0,154
X_6	-0,542	0,286	0,609
X_7	-0,921	-0,207	-0,150
X_8	-0,080	0,915	0,064
X_9	-0,624	0,100	-0,103
X_{10}	-0,938	-0,097	-0,061

Variabel dikatakan memiliki hubungan yang kuat dengan komponen utama apabila nilai *loading factor* > 0,5 (dalam nilai absolut). Berdasarkan matriks *loading*, komponen utama pertama (PC1) memiliki *loading* tinggi pada variabel jumlah kabupaten/kota yang menerapkan kebijakan germas (X_2), jumlah kasus HIV (X_3), jumlah kabupaten/kota yang menerapkan KTR (X_4), jumlah puskesmas (X_7), persentase penduduk lansia (X_9), dan jumlah kabupaten/kota sehat (X_{10}). Variabel-variabel tersebut berkontribusi kuat terhadap PC1 dan arahnya negatif. Artinya, jika nilai PC1 meningkat, maka nilai variabel-variabel tersebut cenderung menurun. Komponen utama pertama (PC1) dapat diartikan sebagai komponen yang merepresentasikan faktor infrastruktur dan kebijakan kesehatan.

Variabel-variabel yang memiliki *loading* tinggi pada komponen utama kedua (PC2) adalah penyalahgunaan NAPZA (X_5) dan kepadatan penduduk (X_8). Variabel-variabel tersebut berkontribusi kuat terhadap PC2 dan arahnya positif. Komponen utama kedua (PC2) menggambarkan karakteristik sosial-kependudukan, yaitu daerah dengan kepadatan tinggi dan masalah sosial, seperti penyalahgunaan NAPZA. Dengan demikian, PC2 diinterpretasikan sebagai faktor sosial-kependudukan.

Komponen utama ketiga (PC3) memiliki *loading* tinggi pada variabel persentase perokok (X_1) dan persentase pelayanan kesehatan lanjut usia (X_6). PC3 menggambarkan komponen faktor sosial-demografis dan layanan kesehatan lanjut usia. Selanjutnya, skor ketiga komponen utama digunakan sebagai variabel prediktor dalam model regresi binomial negatif.

3.4 Pemodelan Regresi Binomial Negatif

Variabel respons pada penelitian ini adalah jumlah kasus terkonfirmasi TBC di Indonesia yang memiliki nilai rerata sebesar 23.753 dan variansi sebesar 1.610.239.676. Nilai variansi yang jauh lebih besar dari rerata menunjukkan adanya indikasi overdispersi pada data. Oleh karena itu, model regresi Poisson kurang tepat digunakan, sehingga diperlukan model alternatif yang mampu menangani overdispersi, yaitu model regresi binomial negatif. Model regresi binomial negatif digunakan untuk memodelkan jumlah kasus TBC (Y) terhadap skor komponen utama (PC1, PC2, PC3). Persamaan model dapat dituliskan sebagai berikut.

$$\log(\mu_i) = \beta_0 + \beta_1 PC1_i + \beta_2 PC2_i + \beta_3 PC3_i + \varepsilon_i, i = 1, 2, \dots, 34$$

Parameter model regresi binomial negatif diestimasi menggunakan metode MLE. Hasil estimasi parameter disajikan pada Tabel 6.

Tabel 6. Hasil estimasi parameter model binomial negatif

Parameter	Estimasi	Standard error	z-value	p-value
β_0	9,516	0,097	97,915	$< 2 \times 10^{-16}$
β_1	-0,421	0,044	-9,540	$< 2 \times 10^{-16}$
β_2	0,205	0,075	2,732	0,006
β_3	0,041	0,095	0,433	0,665

Berdasarkan hasil estimasi parameter, diperoleh model sebagai berikut.

$$\log(\mu) = 9,516 - 0,421PC1 + 0,205PC2 + 0,041PC3$$

Untuk memudahkan interpretasi, model dapat dituliskan dalam bentuk berikut.

$$\mu = e^{9,516-0,421PC1+0,205PC2+0,041PC3}$$

Koefisien untuk PC1 adalah -0,421. Pada skala multiplikatif, satu satuan kenaikan pada PC1 mengalikan nilai harapan (rerata) jumlah kasus TBC sebesar $\exp(-0,421) \approx 0,656$. Artinya, untuk setiap peningkatan satu unit pada PC1, rata-rata jumlah kasus TBC diestimasi menurun sekitar 34,4% ($1 - 0,656 = 0,344$). PC1 mewakili faktor infrastruktur dan kebijakan kesehatan, sehingga semakin baik kesiapan infrastruktur/kebijakan kesehatan di suatu provinsi, semakin rendah jumlah kasus TBC di provinsi tersebut.

Koefisien untuk PC2 adalah 0,205, jika dieksponenkan menjadi sekitar 1,228. Artinya, untuk setiap peningkatan satu unit pada PC2, rata-rata jumlah kasus TBC diestimasi meningkat sekitar 22,8%. PC2 berkaitan dengan faktor sosial (penyalahgunaan NAPZA) dan kepadatan penduduk. Sehingga jika semakin besar kedua faktor tersebut, semakin meningkat pula jumlah kasus TBC di suatu provinsi.

Koefisien untuk PC3 adalah 0,041 yang jika dieksponenkan menjadi sekitar 1,042. Artinya, untuk setiap peningkatan satu unit pada PC3, rata-rata jumlah kasus TBC diestimasi meningkat sekitar 4,2%. PC3 berkaitan dengan faktor sosial-demografis dan layanan kesehatan lanjut usia. Sehingga, semakin besar persentase perokok dan persentase pelayanan kesehatan lanjut usia, semakin meningkat jumlah kasus TBC di suatu provinsi.

Selanjutnya menguji signifikansi parameter model regresi binomial negatif secara simultan menggunakan uji *likelihood ratio* (LR). Pada uji LR, dilakukan perhitungan nilai log likelihood model penuh (memuat semua variabel prediktor) dan model konstan (hanya memuat *intercept*). Hasil uji LR disajikan pada Tabel 7.

Tabel 7. Hasil uji likelihood ratio model binomial negatif

Model	Log Likelihood	df	Likelihood Ratio	p-value
Model konstan	-752,103	33		
Model penuh	-697,095	30	55,008	$6,84 \times 10^{-12}$

Berdasarkan hasil uji LR, diperoleh nilai statistic $LR = 55,008$ dengan $p\text{-value} = 6,84 \times 10^{-12}$, sehingga H_0 ditolak. Dengan semikian, model penuh yang memuat variabel PC1, PC2, dan PC3 secara simultan berpengaruh signifikan terhadap jumlah kasus TBC. Setelah uji signifikansi parameter secara simultan, dilakukan pengujian signifikansi parameter secara parsial menggunakan uji Wald. Merujuk pada Tabel 6, parameter yang berpengaruh signifikan adalah β_1 dan β_2 karena memiliki $p\text{-value} < \alpha = 0,05$. Sehingga, dapat disimpulkan bahwa faktor infrastruktur dan kebijakan kesehatan (PC1) berpengaruh negatif dan signifikan terhadap jumlah kasus TBC, serta faktor sosial dan kepadatan penduduk (PC2) berpengaruh positif dan signifikan terhadap jumlah kasus TBC di Indonesia.

Uji kelayakan model (*goodness of fit*) bertujuan untuk mengevaluasi apakah model binomial negatif yang dibentuk cocok dengan data TBC. Kelayakan model pada penelitian ini menggunakan uji deviansi dengan H_0 menyatakan model regresi binomial negatif layak (*fit*). Hasil uji kelayakan model disajikan pada Tabel 8.

Tabel 8. Hasil uji kelayakan model regresi binomial negatif

Statistik uji	Derajat bebas	Rasio	p-value
35,798	30	1,19	0,215

Berdasarkan Tabel 8, diperoleh nilai rasio deviansi sebesar 1,19 yang menunjukkan bahwa nilainya mendekati 1. Selain itu, diperoleh $p\text{-value} > 0,05$, berarti bahwa tidak ada perbedaan signifikan antara nilai yang diamati dan nilai yang diprediksi oleh model. Dengan demikian, model regresi binomial negatif dinyatakan layak untuk digunakan pada data jumlah kasus TBC di Indonesia.

4. Kesimpulan

Penelitian ini bertujuan untuk mengatasi permasalahan multikolinieritas dan overdispersi dalam pemodelan jumlah kasus TBC di Indonesia melalui pendekatan hibrid PCA dan regresi binomial negatif. Analisis korelasi menunjukkan adanya hubungan kuat antarvariabel prediktor, sehingga PCA digunakan untuk mereduksi sepuluh variabel menjadi tiga komponen utama yang mampu menjelaskan sekitar 78% keragaman total data tanpa kehilangan informasi penting. Selanjutnya, model regresi binomial negatif

yang dibangun menggunakan ketiga komponen utama tersebut terbukti signifikan secara simultan dan memenuhi kriteria kelayakan model. Hasil uji Wald menunjukkan bahwa dua komponen utama berpengaruh signifikan terhadap jumlah kasus TBC, yaitu komponen yang merepresentasikan faktor infrastruktur dan kebijakan kesehatan yang berpengaruh negatif, serta komponen yang mencerminkan faktor sosial dan kepadatan penduduk yang berpengaruh positif. Temuan ini menegaskan bahwa peningkatan kualitas layanan dan kebijakan kesehatan berpotensi menurunkan kasus TBC, sementara kondisi sosial dan kepadatan penduduk berkontribusi terhadap peningkatan risiko penularan. Untuk penelitian selanjutnya, disarankan untuk memasukkan aspek spasial dalam pemodelan, mengingat penyebaran TBC antarprovinsi berpotensi saling memengaruhi melalui mobilitas penduduk dan keterhubungan wilayah. Pendekatan seperti *spatial negative binomial regression* atau model autoregresif spasial dapat memberikan gambaran yang lebih komprehensif mengenai pola penyebaran penyakit. Selain itu, penelitian dapat memperluas variabel prediktor dengan memasukkan faktor lingkungan dan sosial-ekonomi yang lebih rinci, seperti kualitas udara, tingkat kemiskinan, atau kepadatan hunian, untuk meningkatkan kemampuan model dalam menjelaskan variasi kasus TBC.

Referensi

- [1] WHO, "Global Tuberculosis Report 2024." Accessed: Nov. 27, 2025. [Online]. Available: <https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024>
- [2] A. Pratama, M. H. Nugraha, N. A. Naofal, N. Anggara, P. Lumbantobing, and E. Sulastri, "Community-based strategies for tuberculosis prevention and control: synergistic approach in reaching all levels of society," *J. Public Health*, vol. 46, no. 4, pp. e693–e694, 2024, doi: <https://doi.org/10.1093/pubmed/fdae155>.
- [3] K. H. Lee, C. Pedroza, E. B. C. Avritscher, R. A. Mosquera, and J. E. Tyson, "Evaluation of negative binomial and zero-inflated negative binomial models for the analysis of zero-inflated count data: application to the telemedicine for children with medical complexity trial," *Trials*, vol. 24, no. 1, p. 613, Sep. 2023, doi: [10.1186/s13063-023-07648-8](https://doi.org/10.1186/s13063-023-07648-8).
- [4] V. Eminita, A. Kurnia, and K. Sadik, "Penanganan Overdispersi pada Pemodelan Data Cacah dengan Respon Nol Berlebih (Zero-Inflated)," *FIBONACCI J. Pendidik. Mat. Dan Stat.*, vol. 5, no. 1, pp. 71–80, Jul. 2019, doi: [10.24853/fbc.5.1.71-80](https://doi.org/10.24853/fbc.5.1.71-80).
- [5] T. Koç and H. Koç, "A New Effective Jackknifing Estimator in the Negative Binomial Regression Model," *Symmetry*, vol. 15, no. 12, p. 2107, Dec. 2023, doi: [10.3390/sym15122107](https://doi.org/10.3390/sym15122107).
- [6] C. Tapan, A. Na-udom, and J. Rungrattanaubol, "A Comparison of Models for Count Data with an Application to Over-Dispersion Data," *J. Appl. Sci. Emerg. Technol.*, vol. 22, no. 2, pp. 1–9, 2023.
- [7] A. J. Syafiqoh, R. Mahardika, S. Amaria, E. Winaryati, and M. A. Haris, "Pemodelan Regresi Binomial Negatif untuk Mengevaluasi Faktor-faktor yang Mempengaruhi Kasus Tuberkulosis di Provinsi Jawa Barat," *J. MSA Mat. Dan Stat. Serta Apl.*, vol. 12, no. 1, pp. 15–23, Jun. 2024, doi: [10.24252/msa.v12i1.39450](https://doi.org/10.24252/msa.v12i1.39450).
- [8] J. Navelski and K. Odongo, "Making Use of PCA in the Presence of Multicollinearity: An Application to Predicting Body Fat Percentage," Department of Mathematics and Statistics & The School of Economic Sciences Washington State University, 2021.
- [9] N. A. Muhaa, L. M. Mulyono, M. R. Fadhilah, and Y. Umaidah, "Klasterisasi Tren Tuberkulosis Global dengan Principal Component Analysis (PCA) dan K-Means," *J. Pustaka Data Pus. Akses Kaji. Database Anal. Teknol. Dan Arsit. Komput.*, vol. 5, no. 1, pp. 132–143, 2025.
- [10] S. Ririhena and S. B. Loklomin, "Pendekatan Principal Component Analysis pada Data Demam Berdarah Dengue di Provinsi Maluku Utara," *BAREKENG J. Ilmu Mat. Dan Terap.*, vol. 14, no. 4, pp. 557–564, Dec. 2020, doi: [10.30598/barekengvol14iss4pp557-564](https://doi.org/10.30598/barekengvol14iss4pp557-564).
- [11] M. Z. Nasution, A. A. Nababan, K. U. Syaliman, S. Novelan, and M. Jannah, "Penerapan Principal Component Analysis (PCA) dalam Penentuan Faktor Dominan yang Mempengaruhi Pengidap Kanker Serviks," vol. 3, no. 1, 2019.

- [12] P. M. Ariani and E. Widodo, “Analisis Faktor Penyebab Penyakit DBD di Jawa Tengah Menggunakan Regresi Binomial Negatif,” *J. Kesehat. Vokasional*, vol. 3, no. 1, pp. 1–6, May 2018, doi: 10.22146/jkesvo.33870.
- [13] C. T. Fatmala, M. Hayati, R. Permatasari, M. Hudori, and D. Y. Dalimunthe, “Pemodelan Jumlah Kasus HIV/AIDS di Provinsi Lampung Menggunakan Regresi Binomial Negatif,” *J. Math. Theory Appl.*, vol. 6, no. 2, pp. 168–177, Oct. 2024, doi: 10.31605/jomta.v6i2.4069.
- [14] T. S. Prayoga and Suliadi, “Korelasi Rank-Spearman pada Hubungan Beberapa Variabel Produk Domestik Regional Bruto,” *J. Ris. Stat.*, pp. 137–144, Dec. 2024, doi: 10.29313/jrs.v4i2.5162.
- [15] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, 5th ed. New York: Chapman and Hall/CRC, 2010. doi: 10.1201/9781439896129.
- [16] A. Ifadah, “Analisis Metode Principal Component Analysis (Komponen Utama) dan Regresi Ridge dalam Mengatasi Dampak Multikolinieritas dalam Analisis Regresi Linear Berganda,” Skripsi, Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang, 2011.
- [17] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. United Kingdom: Cengage, 2019.
- [18] A. C. Rencher and W. F. Christensen, “Principal Component Analysis,” in *Methods of Multivariate Analysis*, John Wiley & Sons, Ltd, 2012, pp. 405–433. doi: 10.1002/9781118391686.ch12.
- [19] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.
- [20] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Regression Models*, 4th ed. New York: McGraw-Hill/Irwin, 2005.
- [21] R. Iknas, S. Satriani, and K. Nurfadilah, “Negative Binomial and Generalized Poisson Regression Model for Death Due to Dengue Hemorrhagic Fever Data,” *Eig. Math. J.*, pp. 39–48, Jun. 2023, doi: 10.29303/emj.v6i1.153.
- [22] F. Fatmasari, “Pendekatan Regresi Binomial Negatif untuk Data Berdistribusi Poisson yang Mengalami Over Dispersi (Studi Kasus DBD di Kota Malang),” Skripsi, Universitas Brawijaya, Malang, 2014.
- [23] Irwan and P. S. Devni, “Pemodelan Regresi Poisson, Binomial Negatif Pada Kasus Kecelakaan Kendaraan Bermotor di Lalu Lintas Sumatera Barat,” *Pros. Semin. Nas. Mat. Dan Pendidik. Mat. Penguatan Peran Mat. Dan Pendidik. Mat. Untuk Indones. Yang Lebih Baik*, 2013, Accessed: Nov. 28, 2025. [Online]. Available: <http://uny.ac.id>



© The Author(s) 2026. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License. Editorial of Journal of Mathematics: Theory and Applications, Department of Mathematics, Universitas Sulawesi Barat, Jalan Prof. Dr. Baharuddin Lopa, S.H., Talumung, Majene 91412, Sulawesi Barat.