



PEMODELAN PENYAKIT TUBERKULOSIS DI JAWA BARAT MENGUNAKAN *LINEAR MIXED MODEL* DENGAN PENALTI *SMOOTHLY CLIPPED ABSOLUTE DEVIATION*

VERA MAYA SANTI¹, KHAIRUNNISA HELMINA^{2*}, YUDI MAHATMA³

¹Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta

^{2,3}Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta

* khairunnisahelmina30@gmail.com

ABSTRAK

Tuberkulosis (TBC) adalah salah satu penyakit menular utama di Indonesia, dengan Jawa Barat sebagai provinsi dengan kasus terbanyak, mencapai lebih dari 13% dari total nasional. Penelitian ini menganalisis faktor-faktor yang memengaruhi penyebaran TBC di Jawa Barat menggunakan *Linear Mixed Model* (LMM) dengan penalti *Smoothly Clipped Absolute Deviation* (SCAD). Model ini dipilih untuk mengatasi korelasi antar amatan serta masalah multikolinearitas dan *overfitting*. Studi ini menggunakan data sekunder yang diperoleh dari Badan Pusat Statistik (BPS), mencakup 27 kabupaten/kota di Provinsi Jawa Barat. Variabel penelitian yaitu persentase individu yang terinfeksi tuberkulosis sebagai variabel terikat dengan melibatkan sejumlah 17 efek tetap, sedangkan kabupaten/kota menjadi efek acak dalam model. Hasil analisis menunjukkan bahwa LMM SCAD lebih efektif dalam menyederhanakan model dan meningkatkan generalisasi. Faktor signifikan yang memengaruhi penyebaran TBC meliputi kondisi rumah tidak layak huni, rumah sehat, laju pertumbuhan penduduk, dan pengeluaran per kapita. Evaluasi model menggunakan AIC, BIC, dan GCV menunjukkan bahwa LMM SCAD memiliki performa lebih baik dibandingkan model LMM.

Kata kunci: Tuberkulosis, LMM, Multikolinearitas, SCAD

ABSTRACT

Tuberculosis (TB) is a major infectious disease in Indonesia, with West Java contributing the highest number of cases, exceeding 13% of the national total. This study analyzes the factors influencing TB transmission in West Java using the Linear Mixed Model (LMM) with the Smoothly Clipped Absolute Deviation (SCAD) penalty. This model is chosen to address correlation among observations as well as issues of multicollinearity and overfitting. This study uses secondary data obtained from the Central Statistics Agency (BPS), covering 27 districts/cities in West Java Province. The research variable is the percentage of individuals infected with tuberculosis as the dependent variable involving 17 fixed effects, while the district/city is a random effect in the model. The results indicate that LMM SCAD effectively simplifies the model and improves generalization. Significant factors affecting TB transmission include inadequate housing conditions, healthy housing, population growth rate, and per capita expenditure. Model evaluation using AIC, BIC, and GCV confirms that LMM SCAD outperforms conventional LMM in handling multicollinearity.

Keywords: Tuberculosis, LMM, Multicollinearity, SCAD.

1 Pendahuluan

Tuberkulosis Paru (TBC) adalah penyakit menular akibat bakteri *Mycobacterium tuberculosis* yang menyerang paru-paru dan menyebar melalui udara. Indonesia menempati peringkat kedua dunia dengan 969 ribu kasus dan 93 ribu kematian pada 2022 [1]. Jawa Barat mencatat kasus tertinggi secara nasional, lebih dari 13%, akibat kepadatan penduduk dan tantangan kesehatan lainnya [2]. Tingginya kasus ini membutuhkan pendekatan ilmiah untuk mengidentifikasi faktor risiko dan menentukan langkah penanganan.

Penelitian sebelumnya telah mengidentifikasi sejumlah faktor risiko TBC, seperti status HIV, tingkat pengangguran, usia lanjut, dan lingkungan fisik [3], [4]. Dalam konteks Indonesia, variabel-variabel seperti fasilitas sanitasi, kepadatan penduduk, gizi balita, dan bahan bakar memasak juga berkontribusi signifikan terhadap penyebaran TBC. Namun, sebagian besar penelitian hanya menggunakan metode analisis linier sederhana yang kurang mampu menangani data kompleks dengan struktur kelompok atau dimensi tinggi.

Untuk mengatasi tantangan tersebut, pendekatan *Linear Mixed Model* (LMM) menjadi relevan. LMM memungkinkan analisis yang mempertimbangkan efek tetap dan acak, sehingga dapat menangani heteroskedastisitas serta variasi antar kelompok [9]. Metode ini telah digunakan dalam berbagai analisis data berkelompok, seperti dalam penelitian [5] yang mengevaluasi literasi membaca di Indonesia. Meski begitu, LMM dapat mengalami kesulitan dalam mengatasi multikolinearitas dan banyaknya variabel independen, terutama pada data berdimensi tinggi.

Untuk meningkatkan efisiensi model, metode penyusutan seperti *Smoothly Clipped Absolute Deviation* (SCAD) digunakan. SCAD menambahkan penalti pada variabel tetap sehingga dapat menyeleksi variabel yang paling relevan dan menghasilkan estimasi parameter yang efisien. Metode ini terbukti lebih unggul dibandingkan regresi ridge dan LASSO, terutama dalam kasus data berdimensi tinggi [7], [10], [11]. Dengan SCAD, model menjadi lebih sederhana dan mudah diinterpretasikan.

Penelitian ini bertujuan menganalisis faktor-faktor penyebab penyakit TBC di Jawa Barat menggunakan pendekatan *Linear Mixed Model* dengan penalti SCAD. Melalui metode ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam tentang variabel-variabel signifikan yang mempengaruhi penyebaran TBC, sehingga dapat mendukung pengambilan kebijakan pencegahan yang lebih efektif.

2 Tinjauan Pustaka

2.1 Data

Penelitian ini menggunakan data faktor penyebab tuberkulosis di 27 Kabupaten/Kota di Jawa Barat, meliputi satu variabel dependen dan 17 variabel independen. Data bersumber dari Badan Pusat Statistik tahun 2023, mencakup rumah tidak layak huni (X_1), rumah sehat (X_2), sanitasi layak (X_3), sumber air minum layak (X_4), jumlah penduduk (X_5), laju pertumbuhan penduduk (X_6), luas daerah (X_7), kepadatan penduduk (X_8), balita kurang gizi (X_9), imigrasi masuk (X_{10}), timbunan sampah (X_{11}), pengeluaran per kapita (X_{12}), penduduk miskin (X_{13}), usia harapan hidup (X_{14}), harapan lama sekolah (X_{15}), rata-rata lama sekolah (X_{16}), dan tingkat kelahiran (X_{17}).

2.2 *Linear Mixed Model*

Model Linier Campuran (MLC) atau *Linear Mixed Model* (LMM) menggabungkan analisis varians dan regresi linier dengan memasukkan efek acak untuk menjelaskan variasi antar dan dalam kelompok [9], [12]. Efek acak mengestimasi rata-rata variabel, sementara efek tetap mengestimasi varians [13].

Secara matematis *Linear Mixed Model* menurut [14] adalah sebagai berikut:

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\beta + \mathbf{Z}_{ij}\mathbf{u}_i + \varepsilon_{ij} \quad (1)$$

di mana, $\mathbf{Y}_{mn}$ adalah variabel respons untuk observasi ke- i pada kelompok ke- j , $\mathbf{X}_{mn}$ adalah matriks untuk efek tetap untuk observasi ke- i pada kelompok ke- j , β adalah vektor parameter untuk efek tetap, $\mathbf{Z}_{mn}$ adalah matriks untuk efek acak untuk observasi ke- i pada kelompok ke- j , $\mathbf{u}_m$ adalah vektor parameter untuk efek acak dari kelompok ke- i , ε_{ij} adalah nilai *error* / residual, i adalah jumlah amatan, j adalah jumlah kelompok, dengan $m = 1, 2, \dots, i$ dan $n = 1, 2, \dots, j$. Bentuk umum LMM secara sederhana dapat dituliskan sebagai [17] :

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon \quad (2)$$

2.3 Metode *Maximum Likelihood*

Metode *Maximum Likelihood* (ML) adalah pendekatan statistik untuk mengestimasi parameter model dengan memaksimalkan fungsi *likelihood* data yang diamati [15]. Misalkan $X_1, X_2, \dots, X_n$ adalah sampel acak dari fungsi kepadatan $f(x; \theta)$. Fungsi *likelihood* gabungan dinyatakan sebagai $L(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta)f(x_2; \theta)\dots f(x_n; \theta)$ [16]. Untuk peubah acak normal dan kontinu, *likelihood* adalah hasil perkalian fungsi kepadatan peluang setiap n [18].

2.4 *Linear Mixed Model SCAD*

Metode SCAD pada model LMM mengurangi fungsi *log-likelihood* dengan penalti SCAD untuk memaksimalkannya [20]. Penalti SCAD mendorong koefisien variabel yang tidak diperlukan menjadi nol [19]. Parameter λ mengontrol kekuatan penalti, p adalah jumlah parameter, dan β_j adalah koefisien untuk variabel ke- j dengan $j = 1, 2, \dots, k$, di mana k adalah jumlah variabel independen. Fungsi log-likelihood dengan penalti SCAD pada LMM dirumuskan sebagai:

$$\log L_{p_\lambda, LMM} = \log L_{LMM} - n\lambda \sum_{j=1}^d p_\lambda |\beta_j| \quad (3)$$

2.5 Uji Signifikansi Parameter

Pengujian dilakukan untuk menentukan apakah parameter model memiliki pengaruh yang signifikan terhadap variabel dependen. Uji signifikansi dilakukan baik secara parsial dan serentak.

2.5.1 Uji Wald

Uji Wald menilai signifikansi parameter, baik efek tetap maupun acak, secara parsial. Uji ini mengestimasi parameter $\hat{\beta}$ untuk menghasilkan statistik Wald (Z atau t) yang berdistribusi *Chi-square* dan nilai *p-value* dengan hipotesis sebagai berikut:

$H_0 : \beta = 0$ (parameter tidak memiliki efek signifikan)

$H_1 : \beta \neq 0$ (parameter memiliki efek signifikan)

Dan statistik uji:

$$W^2 = \left[\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right]^2 \quad (4)$$

di mana, $\hat{\beta}_i$ adalah nilai estimasi parameter dan $SE(\hat{\beta}_i)$ adalah standar *error*. Jika nilai *p-value* lebih kecil atau sama dengan taraf signifikansi, maka hipotesis H_0 ditolak dan H_1 diterima [21].

2.5.2 Uji Ratio Likelihood

Uji rasio *likelihood* membandingkan dua model LMM untuk menilai signifikansi parameter, baik efek tetap maupun acak. Uji ini menghitung rasio *likelihood* untuk menghasilkan statistik *Chi-square* dan *p-value*. Jika *p-value* lebih kecil dari α , parameter dianggap signifikan secara statistik.

Hipotesis uji:

$H_0 : \beta = 0$ (parameter tidak memiliki efek signifikan)

$H_1 : \beta \neq 0$ (minimal ada 1 parameter memiliki efek signifikan)

Dan statistik uji:

$$G = -2 \log \frac{L_0}{L_1} \quad (5)$$

di mana, L_0 adalah fungsi *likelihood* yang tidak memuat variabel independen dan L_1 adalah fungsi *likelihood* yang memuat variabel independen. Jika nilai *p-value* lebih kecil dari taraf signifikansi (α), maka hipotesis H_0 ditolak sehingga parameter signifikan secara statistik [22].

2.6 Kriteria Kebaikan Model

Kriteria kebaikan model digunakan untuk menilai kualitas dan efektivitas model dalam merepresentasikan realitas dan memprediksi secara akurat. Menurut [23] dan [24], beberapa kriteria yang digunakan adalah GCV, AIC, dan BIC. *General Cross Validation* (GCV) dihitung berdasarkan rumus,

$$GCV = \frac{\|y - X\beta\|^2}{n \left(1 - \frac{d}{n}\right)^2} \quad (6)$$

dengan, $\|y - X\beta\|^2$ adalah jumlah kuadrat dari residual, d adalah jumlah parameter efektif dalam model, dan n jumlah amatan dalam data.

Akaike Information Criterion (AIC) dihitung berdasarkan rumus berikut,

$$AIC = -2 \ln L + 2P \quad (7)$$

dengan, L adalah nilai dari fungsi *log-likelihood* dan P adalah jumlah variabel independen.

Bayesian Information Criterion (BIC) dihitung berdasarkan rumus berikut,

$$BIC = -2L + P \log(n) \quad (8)$$

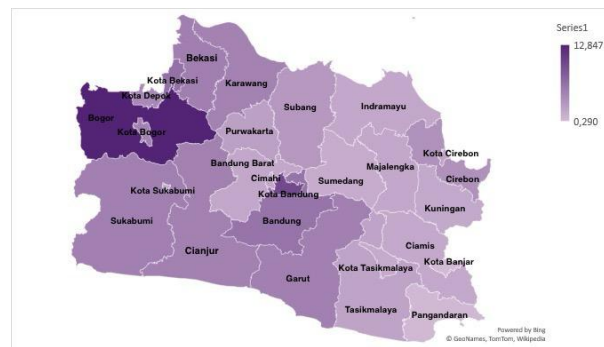
dengan, L adalah nilai dari fungsi *log-likelihood*, P adalah jumlah variabel independen, dan n adalah ukuran sampel.

Model terbaik adalah yang memiliki nilai AIC, BIC, dan GCV terendah, menunjukkan keseimbangan optimal antara kecocokan data dan kompleksitas, dengan preferensi pada model yang lebih sederhana.

3 Hasil dan Pembahasan

3.1 Analisis Deskriptif

Sebaran dari tuberkulosis di Jawa Barat pada tahun 2022 dapat dilihat pada gambar berikut.



Gambar 1: Nilai Lambda

Berdasarkan data yang ditampilkan pada peta, wilayah dengan jumlah kasus tertinggi ditandai dengan gradasi warna yang lebih gelap, sementara wilayah dengan jumlah kasus yang lebih rendah memiliki warna yang lebih terang. Kabupaten/Kota dengan jumlah kasus tuberkulosis tertinggi adalah Kabupaten Bogor dan terendah berada pada Kota Banjar.

3.2 Spesifikasi Model *Linear Mixed Model*

Pemodelan *Linear Mixed Model* (LMM) dilakukan dengan *fitting* menggunakan metode penduga Maximum Likelihood. Hasil dari pemodelan dengan LMM dapat dilihat pada tabel berikut.

Tabel 1: Model LMM

Variabel	Estimate	Std. Error	t-value	Variabel	Estimate	Std. Error	t-value
(Intercept)	3,704	0,093	39,761	X ₉	-0,028	0,152	-0,188
X ₁	1,333	0,371	3,592	X ₁₀	-0,311	0,180	-1,728
X ₂	-0,328	0,525	-0,624	X ₁₁	-0,008	0,210	-0,041
X ₃	0,012	0,575	0,021	X ₁₂	-0,410	0,170	-2,403
X ₄	-0,603	0,396	-1,522	X ₁₃	-0,340	0,246	-1,382
X ₅	0,076	0,272	0,281	X ₁₄	0,318	0,558	0,571
X ₆	1,433	0,475	3,017	X ₁₅	-0,407	0,218	-1,861
X ₇	-0,080	0,162	-0,494	X ₁₆	0,013	0,010	1,276
X ₈	0,121	0,374	0,323	X ₁₇	0,108	0,350	0,310
Variances: 0,201				Residual: 0,032			

Hasil analisis LMM menunjukkan model memiliki kecocokan yang baik dengan AIC (77, 4) dan BIC (103, 4) rendah serta variasi signifikan antar kabupaten/kota, menegaskan pentingnya efek kelompok. Koefisien variabel X₁ (rumah tidak layak huni) menunjukkan hubungan positif signifikan dengan infeksi tuberkulosis, yang berarti semakin banyak rumah tidak layak huni, semakin tinggi tingkat infeksi. Sebaliknya, koefisien X₂ (rumah sehat) menunjukkan hubungan negatif signifikan, mengindikasikan bahwa semakin banyak rumah sehat, semakin rendah tingkat infeksi tuberkulosis. Hal yang sama berlaku untuk variabel lainnya.

3.3 Uji Multikolinearitas

Uji multikolinearitas dilakukan untuk mengantisipasi adanya korelasi antara variabel dependen dengan variabel independen. Berikut adalah hasil uji multikolinearitas menggunakan VIF.

Tabel 2: Nilai VIF

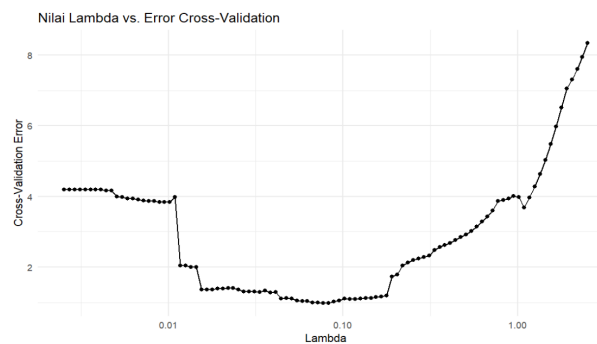
Variabel	VIF Value	Keterangan
X_1	15,296	Multikolinearitas
X_2	30,697	Multikolinearitas
X_3	36,728	Multikolinearitas
X_4	17,458	Multikolinearitas
X_5	8,242	-
X_6	25,041	Multikolinearitas
X_7	2,923	-
X_8	15,591	Multikolinearitas

Variabel	VIF Value	Keterangan
X_9	2,576	-
X_{10}	3,606	-
X_{11}	4,901	-
X_{12}	3,238	-
X_{13}	3,022	-
X_{14}	34,592	Multikolinearitas
X_{15}	5,314	-
X_{16}	9,564	-
X_{17}	13,594	Multikolinearitas

Terdapat 8 variabel independen yang nilai VIF > 10 yang mengindikasikan adanya multikolinearitas. Oleh karena itu, digunakan metode alternatif SCAD.

3.4 Spesifikasi Model *Linear Mixed Model* SCAD

Model LMM dengan penalti SCAD menggunakan tuning parameter λ yang dipilih berdasarkan *log-likelihood* optimal melalui *10-fold cross-validation*, memastikan kompleksitas model yang optimal. Grafik *cross-validation error* menunjukkan hubungan λ dengan kesalahan prediksi.

**Gambar 2:** Nilai Lambda

Berdasarkan gambar di atas dapat dilihat bahwa nilai λ optimum sebesar 0,178. Pendugaan parameter LMM dengan penalti SCAD dilakukan dengan menggunakan metode *Maximum Likelihood*. Berikut hasil pendugaan parameter LMM SCAD.

Tabel 3: Estimasi Parameter LMM dengan SCAD

Variabel	Koefisien	Variabel	Koefisien	Variabel	Koefisien
<i>Intercept</i>	3,704	X_6	1,337	X_{12}	-0,194
X_1	1,577	X_7	0,000	X_{13}	0,000
X_2	0,220	X_8	0,000	X_{14}	0,000
X_3	0,000	X_9	0,002	X_{15}	0,000
X_4	0,000	X_{10}	0,000	X_{16}	0,000
X_5	0,000	X_{11}	0,000	X_{17}	0,000

Variabel yang menjadi faktor utama penyebaran tuberkulosis meliputi X_1 , X_2 , X_6 , X_9 dan X_{12} . Koefisien positif menunjukkan hubungan positif, sedangkan koefisien negatif menunjukkan hubungan negatif dengan variabel dependen.

3.5 Pendugaan Ragam Acak (u)

Nilai ragam acak (u) menunjukkan perbedaan kontribusi setiap kelompok (kabupaten/kota) terhadap variabel dependen setelah mempertimbangkan efek tetap.

Tabel 4: Estimasi Parameter LMM dengan SCAD

Kota	Ragam acak	Kota	Ragam acak	Kota	Ragam acak
Purwakarta	0,587	Kota Cirebon	0,221	Kota Depok	-0,131
Kota Tasikmalaya	0,502	Kota Banjar	0,222	Bekasi	-0,215
Ciamis	0,489	Garut	0,157	Kota Bekasi	-0,238
Kuningan	0,489	Pangandaran	0,152	Kota Cimahi	-0,253
Cianjur	0,489	Kota Sukabumi	0,144	Majalengka	-0,386
Subang	0,471	Kota Bogor	0,123	Indramayu	-0,463
Bandung	0,436	Kota Bandung	0,014	Karawang	-0,540
Bogor	0,422	Sukabumi	0,016	Bandung Barat	-1,160

Berdasarkan analisis, Purwakarta memiliki efek acak paling tinggi (0,586), artinya kota ini sangat mendukung kenaikan nilai Y . Sebaliknya, Tasikmalaya memiliki efek acak paling rendah (-0,747), yang berarti kota ini justru menurunkan nilai Y . Untuk mengukur proporsi varians antar kelompok (kab/kota) dengan varians dalam kelompok (residual), dihitung *Interclass Correlation Coefficient* (ICC) dengan perhitungan berikut [25].

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} = \frac{0,201}{0,201 + 0,032} = 0,862$$

Nilai ICC sebesar 0,862 mendekati 1, yang menunjukkan bahwa sebagian besar variabilitas dalam data berasal dari perbedaan antar kelompok dibandingkan dengan variabilitas dalam kelompok. ICC di atas 0,75 umumnya dianggap sangat baik, menunjukkan bahwa pengukuran memiliki keandalan tinggi.

3.6 Uji Signifikansi Parameter

3.6.1 Uji Wald

Hasil uji Wald memberikan informasi penting tentang variabel mana yang perlu dipertahankan dalam model dan variabel mana yang dapat dihilangkan dengan hasil uji berikut.

Tabel 5: Hasil Uji Wald

Variabel	Koefisien	Standar Error	Z-score	p-value
(Intercept)	3,704	0,129	28,551	0,000e+00
X_1	1,495	0,212	7,050	1,781e-12*
X_2	0,334	0,183	1,823	6,816e-02*
X_6	1,375	0,192	7,144	9,032e-13*
X_9	0,116	0,147	0,793	4,273e-01
X_{12}	-0,353	0,104	-3,380	7,230e-04*

Intersep sebesar 3,704 merepresentasikan rata-rata variabel dependen saat variabel independen bernilai nol. Variabel X_1 , X_2 , X_6 , dan X_{12} signifikan ($p < 0,05$), sementara X_9 tidak signifikan ($p > 0,05$).

3.6.2 Uji Ratio Likelihood

Uji *ratio likelihood* dilakukan dengan membandingkan dua model yang dapat dilihat pada tabel di bawah.

Tabel 6: Hasil Uji *Ratio Likelihood*

Model	Log-Likelihood	Chi-square	df	Pr(>Chisq)
Reduced	-65,864			
Unreduced	-18,722	94,285	17	1,007e-12

Hasil LRT menunjukkan bahwa model kompleks (memuat seluruh variabel independen) lebih baik dari pada model sederhana (tidak memuat variabel independen) dalam menjelaskan hubungan variabel. Nilai $\text{Pr}(>\text{Chisq})$ adalah $1,007e - 12$ dimana, nilai ini lebih kecil dari taraf signifikansi sebesar 0,05, mengindikasikan bahwa model kompleks lebih baik dan signifikan.

3.7 Pemilihan Model

Nilai AIC, BIC, dan GCV dapat dibandingkan untuk memilih model terbaik.

Tabel 7: Perbandingan nilai AIC, BIC, dan GCV

Metode	LMM	LMM SCAD
AIC	77,443	33,743
BIC	103,359	41,518
GCV	1,083	0,198

Analisis LMM dengan penalti SCAD menunjukkan kinerja lebih baik dibandingkan model awal, dengan penurunan signifikan pada AIC, BIC, dan GCV. Model ini mengatasi overfitting, multikolinearitas, serta meningkatkan generalisasi dan akurasi prediksi tanpa mengurangi akurasi estimasi.

Berdasarkan model LMM dengan penalti SCAD, koefisien pada variabel independen dapat diinterpretasikan terhadap variabel dependen yaitu persentase rumah tidak layak huni (X_1) di Jawa Barat pada tahun 2022 diperkirakan sebesar 1,495% hal ini menandakan setiap kenaikan 1% rumah tidak layak huni akan meningkatkan penyebaran TBC sebesar 1,495%. Penelitian ini sejalan dengan penelitian [26] dan [27] yang menyimpulkan bahwa persentase rumah tidak layak huni menjadi faktor yang berpengaruh terhadap rata-rata jumlah kasus TBC di Indonesia. Rumah Sehat (X_2) dengan estimasi parameter sebesar 0,334%, hal ini menunjukkan bahwa kondisi rumah yang tidak memenuhi standar kesehatan memiliki peran dalam meningkatkan risiko penularan tuberkulosis. Penelitian ini sejalan dengan penelitian [28] yang menyatakan bahwa wilayah dengan rumah sehat yang rendah memiliki angka kasus TBC yang tinggi. Laju Pertumbuhan Penduduk (X_6) dengan estimasi sebesar 1,375% berarti bahwa setiap kenaikan 1% dalam laju pertumbuhan penduduk di suatu wilayah di Jawa Barat diperkirakan akan menyebabkan peningkatan 1,375% dalam jumlah kasus TBC. Pengeluaran Per kapita (X_{12}) memiliki pengaruh terhadap penyebaran tuberkulosis di Jawa Barat dengan koefisien sebesar $-0,353\%$. Ini berarti bahwa peningkatan pengeluaran per kapita berhubungan dengan penurunan tingkat penyebaran tuberkulosis. Hal ini sejalan dengan penelitian [29] dan [30] yang menyimpulkan bahwa pengeluaran per kapita mempengaruhi kejadian TBC.

4 Kesimpulan

Model *Linear Mixed Model* (LMM) dengan penerapan penalti SCAD menunjukkan hasil yang lebih baik dalam mengatasi masalah *overfitting* dan meningkatkan kemampuan generalisasi model. Penurunan kompleksitas model juga terlihat dari berkurangnya jumlah variabel independen yang signifikan sehingga menjadikan model lebih baik dalam memodelkan faktor-faktor penyebab tuberkulosis yang mengalami masalah multikolinearitas. Hal ini juga ditunjukkan berdasarkan

perhitungan nilai AIC, BIC, dan GCV. Beberapa faktor yang menjadi penyebab penularan tuberkulosis di Jawa Barat yaitu, rumah tidak layak huni, rumah sehat, laju pertumbuhan penduduk, dan pengeluaran perkapita.

5 Ucapan Terima Kasih

Terima kasih kepada Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Jakarta serta semua pihak yang telah membantu dalam penelitian.

Pustaka

- [1] World Health Organization. *Tuberculosis*. 2024.
- [2] Badan Pusat Statistik. *Provinsi Jawa Barat dalam Angka 2023*, 2023.
- [3] O. B. Ojo, S. Lougue, and W. A. Woldegerima, "Bayesian generalized linear mixed modeling of tuberculosis using informative priors," *PLoS ONE*, vol. 12, no. 3, p. e0172580, 2017.
- [4] Y. Masnarivan *et al.*, "Pemodelan faktor risiko tuberkulosis paru di Sumatera Barat," *Jambi Medical Journal: Jurnal Kedokteran dan Kesehatan*, vol. 10, no. 1, pp. 6880, 2022.
- [5] V. M. Santi, S. Azzahra, and D. Siregar, "Analisis skor literasi membaca siswa Indonesia menggunakan linear mixed models," *Journal of Mathematics Education, Science and Technology*, vol. 7, no. 2, pp. 116129, 2022.
- [6] V. M. Santi, A. N. Mutia, and Q. Meidianingsih. "Geographically weighted regression dalam menganalisis faktor-faktor yang mempengaruhi kasus tuberkulosis di Sumatera Utara". *Jurnal Ilmiah Matematika dan Ilmu Pengetahuan Alam*. Vol. 19, No. 2, pp 107-116, 2022.
- [7] V. M. Santi, K. A. Notodiputro, and B. Sartono. "Variable selection methods applied to the mathematics scores of Indonesian students based on convex penalized likelihood". *Journal of Physics: Conference Series*, Volume 1402, Issue 7, 2019.
- [8] V. M. Santi *et al.*, "Generalized linear mixed models by penalized Lasso in modelling the scores of Indonesian students," *Journal of Physics: Conference Series*, vol. 1869, no. 1, p. 012140, 2021. [Online]. Available: <https://doi.org/10.1088/1742-6596/1869/1/012140>
- [9] V. M. Santi, *et al.*, "Restricted Maximum Likelihood Estimation for Multivariate Linear Mixed Model in Analyzing PISA Data for Indonesian Students". *BAREKENG: Jurnal Ilmu Matematika dan Terapan*. Vol. 16, Issue 2, 607-614, 2021.
- [10] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 13481360, 2001.
- [11] I. Wulandari *et al*, *ICSA 2019: Proceedings of the 1st International Conference on Statistics and Analytics*, 2020. [Online]. Available: <https://doi.org/10.4108/eai.2-8-2019.2290485>
- [12] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge University Press, 2006.
- [13] V. A. Brown, "An introduction to linear mixed effect modeling in R," *Advances in Methods and Practices in Psychological Science*, vol. 4, no. 1, 2021.

- [14] Little *et al.*, *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc., 1996.
- [15] L. J. Bain and M. Engelhardt, *Introduction to Probability and Mathematical Statistics*, 2nd ed. Pacific Grove, CA: Brooks/Cole, 1992.
- [16] M. Mood *et al.*, *Introduction to the Theory of Statistics*. New York, NY: McGraw-Hill Book Company, 1986.
- [17] F. Scott, "The LASSO linear mixed model for mapping quantitative trait loci," M.S. thesis, The University of Adelaide, Australia, 2006.
- [18] M. O'Neill, *Matematika REML*. Statistical Advisory & Training Service Pty Ltd., Universitas Brawijaya, 2013.
- [19] Y. Kim, H. Choi, and H. Oh, "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association*, vol. 103, no. 484, 2008.
- [20] F. G. Yavuz and O. Arslan, "Variable selection in elliptical linear mixed model," *Journal of Applied Statistics*, vol. 47, no. 11, 2020.
- [21] D. Nachrowi and H. Usman, *Penggunaan Teknik Ekonometri*. Jakarta, Indonesia: PT Raja Grafindo Persada, 2002.
- [22] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York, NY: John Wiley & Sons, Inc., 1989.
- [23] H. Wang, R. Li, and C.-L. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553-568, 2007.
- [24] J. Zahro *et al.*, *Aplikasi Generalized Linear Model pada R*. Yogyakarta, Indonesia, 2018.
- [25] R. A. Fisher *et al.*, *Statistical Methods for Research Workers*. Edinburgh, U.K.: Oliver and Boyd, 1954.
- [26] N. A. Karima *et al.*, "Analisis faktor-faktor yang mempengaruhi jumlah kasus tuberkulosis di Indonesia menggunakan model geographically weighted Poisson regression," *EKSPONENSIAL*, vol. 12, no. 1, pp. 716, 2021.
- [27] S. Rahayu, "Pengaruh lingkungan fisik terhadap kejadian TB paru," 2018. [Online]. Available: <https://doi.org/10.31219/osf.io/8kmqw>.
- [28] V. Olivionita *et al.*, "Analisis spasial faktor-faktor yang mempengaruhi kejadian tuberkulosis di Malang Raya tahun 2020-2021," *Jurnal Epidemiologi Kesehatan Komunitas*, vol. 9, no. 1, pp. 6371, 2024.
- [29] A. Sejati and L. Sofiana, "Faktor-faktor terjadinya tuberkulosis," *Jurnal Kesehatan Masyarakat*, vol. 10, no. 2, pp. 122128, 2015.
- [30] A. H. Mahpudin and R. Mahkota, "Faktor lingkungan fisik rumah, respon biologis dan kejadian TBC paru di Indonesia," *Kesmas: National Public Health Journal*, vol. 1, no. 4, pp. 144153, 2007.