

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tuberkulosis merupakan salah satu penyakit menular disebabkan oleh bakteri *Mycobacterium tuberculosis*, hingga saat ini masih menjadi masalah kesehatan masyarakat global, termasuk Indonesia. Menurut laporan *World Health Organization* (WHO) tahun 2021, Indonesia menempati posisi kedua dengan jumlah kasus TB terbanyak di dunia setelah India, yaitu diperkirakan sebanyak 969.000 kasus. Angka ini meningkat sebesar 17% dibandingkan tahun 2020 yang tercatat sebanyak 824.000 kasus. TB juga merupakan salah satu penyebab kematian tertinggi akibat penyakit menular di Indonesia, dengan jumlah kasus yang terus meningkat dari tahun ke tahun. Oleh karena itu, upaya pencegahan dan penanggulangan TB menjadi prioritas, salah satunya dengan mengidentifikasi dan mengendalikan faktor-faktor yang mempengaruhi penyebaran penyakit ini (Khotimah & Suwanda, 2022).

Beberapa faktor yang mempengaruhi jumlah kasus TB antara lain seperti kebiasaan merokok dan status gizi pada penelitian (Avy, et.al, 2024). Jenis kelamin, usia di atas 65 tahun, serta persentase penduduk juga mempengaruhi jumlah kasus TB pada penelitian (Rolo, et.al, 2023). Penelitian lain juga menunjukkan bahwa ketersediaan tenaga medis memiliki pengaruh signifikan terhadap kasus TB (Prihatiningsih, et. al, 2020). Selanjutnya, penelitian yang dilakukan oleh Chen (2022) menunjukkan bahwa, variabel yang dapat menjelaskan jumlah kasus tuberkulosis di Jawa Barat adalah dipengaruhi oleh kelompok kesehatan dan lingkungan, seperti jumlah rumah sakit, jumlah puskesmas, jumlah tenaga medis, jumlah penderita gizi buruk, jumlah kasus penyakit HIV/AIDS, jumlah kecamatan, dan kepadatan penduduk. Untuk itu perlu dilakukan suatu pendekatan untuk memodelkan faktor-faktor yang diduga mempengaruhi jumlah kasus TB di Indonesia.

Analisis regresi merupakan metode statistik yang umum digunakan untuk memodelkan hubungan antara variabel respon (Y) dengan satu atau lebih variabel prediktor (X) dalam suatu model linear. Metode *Ordinary Least Square* (OLS) sering digunakan untuk estimasi parameter regresi, dengan ketentuan bahwa asumsi klasik harus terpenuhi (Desviona & Yanuar, 2020). Apabila asumsi klasik terpenuhi, estimator OLS bersifat *Best Linier Unbiased Estimator* (BLUE), Sehingga menghasilkan estimasi parameter yang akurat. Namun, jika terdapat pencilon atau pelanggaran asumsi, sifat baik estimator OLS dapat terganggu. Oleh karena itu, metode regresi *robust* seperti regresi kuantil dapat mengatasi masalah tersebut.

Regresi kuantil memiliki keunggulan dalam menangani adanya titik pencilon dalam variabel respon pada suatu model. Selain itu, model ini juga dapat menghasilkan hasil yang lebih efisien ketika terdapat distribusi yang tidak normal (Beyaztas & Shang, 2022). Estimasi parameter dari regresi kuantil juga lebih stabil dan konsisten (Nurfaizah, 2020). Beberapa penelitian telah menerapkan metode ini, seperti Oktavia, et. al, (2020) yang mengkaji tentang analisis estimasi parameter regresi kuantil dengan metode bootstrap. Selanjutnya penelitian yang dilakukan oleh Aprilia, et. al, (2020) tentang model regresi kuantil orde dua dalam menganalisis perubahan trombosit pasien demam berdarah. Selain itu, penelitian lainnya dilakukan oleh Halamuty, et. al (2022) yang menggunakan analisis regresi kuantil dalam mengatasi data pencilon, hasil penelitian yang didapatkan menyatakan estimasi parameter pada data yang mengandung pencilon dapat diatasi dengan analisis regresi kuantil.

Selain pencilon, masalah lain dalam analisis regresi adalah multikolinearitas, yang terjadi ketika terdapat korelasi yang kuat antara dua atau lebih variabel prediktor. Multikolinearitas dapat menyebabkan a estimasi parameter dan kesulitan dalam interpretasi hasil analisis. Untuk mengatasi aritas, dapat diatasi dengan menggunakan PCA (*Principal Component Analysis*) dan arisasi adalah teknik yang digunakan untuk mengurangi kompleksitas model dengan alti pada koefisien regresi, ada dua jenis regularisasi yang umum digunakan yaitu ridge, ti terdahulu yang telah mengaplikasikan metode tersebut, diantaranya Islamiyati et. al, an PCA, Sulistya et. al, (2023) menggunakan metode ridge, Liu et. al, (2024)



menggunakan LASSO. Dalam penelitian Rahmawati & Suratman (2022) yang membandingkan ridge dan LASSO untuk data yang mengandung multikolinearitas, hasil penelitian yang didapatkan menyatakan bahwa dalam hal interpretasi model regresi LASSO lebih unggul dibandingkan dengan regresi ridge.

Regresi kuantil LASSO merupakan metode gabungan antara regresi kuantil dan LASSO, digunakan untuk memodelkan hubungan antara variabel respon dengan variabel prediktor, regresi kuantil tidak mengasumsikan distribusi normal pada error dan mampu memberikan informasi yang lengkap mengenai pengaruh variabel prediktor di berbagai titik kuantil pada distribusi respon. Sementara itu, penalti LASSO digunakan untuk mencegah overfitting serta melakukan seleksi variabel, sehingga menghasilkan model yang lebih sederhana dan mudah diinterpretasikan. Penelitian Santri & Hanike (2020) menerapkan pemodelan statistical downscaling regresi kuantil LASSO dan analisis komponen utama untuk pendugaan curah hujan ekstrim, sedangkan Fauzan et. al, (2024) membandingkan regresi linear berganda, LASSO, ridge, elastic net, dan kuantil LASSO dalam memodelkan prediksi angka harapan hidup. Namun, penelitian-penelitian tersebut hanya menghasilkan estimasi titik dan belum menyertakan analisis inferensial seperti interval kepercayaan. Padahal, estimasi titik saja tidak cukup untuk menghasilkan kesimpulan statistik yang kuat. Diperlukan interval kepercayaan untuk memberikan ukuran ketidakpastian terhadap estimasi parameter.

Interval kepercayaan menunjukkan rentang nilai yang secara statistik diyakini mengandung nilai parameter populasi dengan tingkat keyakinan tertentu (Koenker & Bassett, 1987). Dalam regresi kuantil, interval kepercayaan dapat dilakukan dengan beberapa pendekatan yaitu pendekatan langsung (*direct*), *rank-score*, dan metode *resampling* seperti bootstrap (Hapsery, 2017). Pendekatan langsung sering digunakan karena efisien dan tidak memerlukan sampel ulang, taksiran interval kepercayaan yang digunakan pendekatan fungsi sparsity. Fungsi *sparsity* menggambarkan kerapatan lokal dari distribusi error pada kuantil tertentu, dan digunakan dalam menghitung varians dan standar error parameter regresi. Dalam konteks regresi kuantil penalized seperti LASSO, metode *sparsity* sangat relevan karena memungkinkan estimasi parameter yang lebih stabil dan efisien (Yuan & Lin, 2005).

Penelitian sebelumnya yang dilakukan oleh Hadijah (2022) berhasil menerapkan sparsity untuk membentuk interval kepercayaan pada regresi kuantil penalized spline, dan metode ini efektif dalam analisis data medis. Penelitian selanjutnya Khotimah (2022), juga menerapkan metode sparsity untuk membentuk interval kepercayaan yang akurat dalam memodelkan incident rate tuberkulosis paru pada model regresi kuantil. Namun demikian, penelitian mengenai pembentukan interval kepercayaan pada regresi kuantil LASSO, terutama dalam masalah kasus TB di Indonesia, masih sangat terbatas. Maka berdasarkan uraian tersebut, penulis tertarik untuk meneliti tentang estimasi interval kepercayaan regresi kuantil LASSO pada data tuberkulosis di Indonesia tahun 2021.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah yang akan dibahas pada penelitian ini sebagai berikut:

1. Bagaimana bentuk estimasi interval kepercayaan parameter regresi kuantil LASSO?
2. Bagaimana model regresi kuantil LASSO pada jumlah kasus tuberkulosis di Indonesia?
3. Bagaimana estimasi interval kepercayaan regresi kuantil LASSO pada data tuberkulosis di Indonesia?

1.3 Batasan Masalah

Berdasarkan rumusan masalah, ruang lingkup dalam penelitian ini akan dibatasi agar tidak menimbulkan penafsiran yang lebih luas, adapun ruang lingkup dalam penelitian ini dibatasi beberapa hal yaitu:



gunakan yaitu $\theta = 0,25; 0,50; 0,75$.

anakan adalah data tuberkulosis di Indonesia tahun 2021

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Mengestimasi interval kepercayaan dari parameter model regresi kuantil LASSO
2. Mendapatkan hasil model regresi kuantil LASSO pada jumlah kasus tuberkulosis di Indonesia.
3. Mendapatkan hasil estimasi interval kepercayaan dari parameter regresi kuantil LASSO pada data tuberkulosis di Indonesia

1.5 Manfaat Penelitian

Penelitian ini bermanfaat sebagai sumber informasi bagi pembaca tentang cara mengestimasi interval kepercayaan regresi kuantil LASSO, serta menambah wawasan pengetahuan tentang model regresi kuantil LASSO pada data yang mengandung masalah pencilan dan multikolinearitas. Selain itu, penelitian ini juga dapat bermanfaat sebagai sumber informasi bagi pemerintah tentang faktor-faktor apa saja yang mempengaruhi jumlah kasus tuberkulosis di Indonesia, sehingga dapat memberikan kontribusi dalam pengembangan kebijakan terkait pencegahan dan pengendalian tuberkulosis di Indonesia.

1.6 Kajian Teori

1.6.1 Regresi Kuantil

Regresi kuantil merupakan suatu pendekatan dalam analisis regresi yang dikenalkan oleh Koenker dan Bassett (1978). Pendekatan ini digunakan untuk menduga hubungan antara variabel respon dan variabel prediktor pada fungsi kuantil bersyarat tertentu, fungsi kuantil dapat diduga dari sebaran bersyarat variabel respon pada setiap nilai kuantil sesuai dengan kuantil yang diinginkan. Analisis regresi kuantil memiliki sifat yang *robust* terhadap data yang mengandung pencilan. Pencilan merupakan kondisi amatan yang menyimpang jauh dari pengamatan lainnya. Pencilan dapat menyebabkan data tidak berdistribusi normal sehingga dapat berpengaruh terhadap pengambilan kesimpulan pada sebuah penelitian (Sihombing et. al, 2022). Terdapat beberapa metode yang dapat digunakan untuk mendefinisikan adanya pencilan yang berpengaruh dalam koefisien regresi yaitu dengan scatter plot dan boxplot. Kelebihan dari regresi kuantil memberikan informasi yang lebih banyak mengenai distribusi dari sebuah data dibandingkan regresi mean dan regresi median (Rizki & Taqiuddin, 2022).

Misalkan terdapat data $\{X_{ij}, y_i\}$, $i = 1, 2, \dots, n$, himpunan pasangan dari variabel random yang berdistribusi secara independen dan identik dengan kuantil $\theta \in (0, 1)$. Data tersebut memiliki fungsi distribusi peluang bersyarat $F(y|X_i) = P(Y \leq y|X_i)$. Persamaan umum regresi kuantil linier khusus untuk kuantil bersyarat dari variabel respon Y , yaitu:

$$y_i(\theta) = \beta_0(\theta) + \beta_1(\theta)X_{i1} + \dots + \beta_k(\theta)X_{ik} + \varepsilon_i(\theta) \quad (1)$$

dengan

$y_i(\theta)$: variabel respon ke- i pada kuantil ke- θ

β_0 : intersep pada sumbu y , ke titik potong sumbu y

β_j : parameter dari variabel ke- j untuk $j = 1, 2, \dots, k$

θ : nilai kuantil (0,1)

X_{ij} : variabel prediktor ke- j dan pada pengamatan ke- i untuk $i = 1, 2, \dots, n$ dan $j = 1, 2, \dots, k$

$\varepsilon_i(\theta)$: error ke- i pada kuantil ke- θ

y_i : variabel respon

X_{ij} : variabel prediktor



kuantil disajikan dalam bentuk matriks, Persamaan (1) dapat ditulis sebagai berikut:

$$\begin{bmatrix} y_1(\theta) \\ y_2(\theta) \\ \vdots \\ y_n(\theta) \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0(\theta) \\ \beta_1(\theta) \\ \vdots \\ \beta_k(\theta) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(\theta) \\ \varepsilon_2(\theta) \\ \vdots \\ \varepsilon_k(\theta) \end{bmatrix} \quad (2)$$

Selanjutnya, bentuk matriks yang disajikan dapat ditulis dalam bentuk persamaan model linier sebagai berikut:

$$\mathbf{y}(\theta) = \mathbf{X}\boldsymbol{\beta}(\theta) + \boldsymbol{\varepsilon}(\theta) \quad (3)$$

dengan

$\mathbf{y}(\theta)$: vektor kolom berukuran $n \times 1$ dari variabel respon y

$\mathbf{X}$: matriks berukuran $n \times (k + 1)$ dengan baris n (pengamatan ke- i) pada kolom k variabel predictor ke- j dengan $i = 1, 2, \dots, n, j = 1, 2, \dots, k$.)

$\boldsymbol{\beta}(\theta)$: vektor kolom berukuran $(k + 1) \times 1$ dari parameter β_j dengan $j = 1, 2, \dots, k$

$\boldsymbol{\varepsilon}(\theta)$: vektor kolom berukuran $n \times 1$ dari error ε_i

Jika fungsi bersyarat dari kuantil ke- θ dengan variabel prediktor X tertentu, maka fungsi bersyarat tersebut didefinisikan sebagai berikut:

$$Q_{y_i}(\theta|X_{i1}, X_{i2}, \dots, X_{ik}) = Q_Y(\theta|X) = \mathbf{X}_i^T \boldsymbol{\beta}(\theta) \quad (4)$$

Koefisien $\boldsymbol{\beta}$ pada Persamaan (4) dapat diestimasi dengan meminimumkan fungsi objektif kuantil sebagai berikut:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(y_i - \mathbf{X}_i^T \boldsymbol{\beta}(\theta)) \quad (5)$$

dengan $y_i = \{y_1, y_2, \dots, y_n\}$ merupakan sampel random dengan variabel respon Y dan $X_i \in \mathbb{R}^k$ merupakan vektor kovariat, sedangkan ρ_{θ} merupakan *loss function* (Balami, 2017).

1.6.2 Estimasi Parameter Regresi Kuantil

Regresi kuantil diestimasi dengan meminimumkan jumlah absolut error yang lebih dikenal dengan *Least Absolute Deviation* (LAD), error pada regresi kuantil diberi bobot yang berbeda, bobot yang digunakan yaitu θ untuk nilai error yang lebih besar atau sama dengan nol dan $(1 - \theta)$ untuk error yang kurang dari nol. Perkalian antara error dengan bobot yang diberikan disebut sebagai *loss function* $P_{\theta}(\varepsilon)$ sebagai berikut:

$$P_{\theta}(\varepsilon) = \sum_{i=1, \varepsilon_i \geq 0}^n \theta |\varepsilon_i| + \sum_{i=1, \varepsilon_i < 0}^n (1 - \theta) |\varepsilon_i| \quad (6)$$

Hal tersebut, dalam regresi kuantil membuat fungsi kuantil ke- θ dari variabel Y dengan syarat X yang mempertimbangkan penduga $\beta(\theta)$, sehingga permasalahan tersebut dapat diperoleh solusinya yang dinyatakan pada Persamaan (7) berikut:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(\varepsilon) = \min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(y_i - Q_Y(\theta|X)) \quad (7)$$

dengan

y_i : variabel respon ke- i

$\rho_{\theta}(\varepsilon)$: *loss function*



kuantil dengan $\theta \in (0, 1)$

kuantil dari variabel Y dengan syarat X

di untuk kuantil bersyarat $Q_Y(\theta|X)$ didefinisikan sebagai berikut:

$$Q_Y(\theta|X) = \mathbf{X}_i^T \boldsymbol{\beta}(\theta) \quad (8)$$

Nilai estimasi terhadap parameter β dari regresi kuantil ke- θ diperoleh dengan meminimumkan *loss function* dari Persamaan (7) sehingga diperoleh persamaan berikut

$$\hat{\beta}(\theta) = \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(\varepsilon) = \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(y_i - \mathbf{X}_i^T \beta(\theta)) \quad (9)$$

dengan $\rho_{\theta}(\varepsilon)$ pada Persamaan (9) didefinisikan

$$\rho_{\theta}(\varepsilon) = \begin{cases} \theta \varepsilon, & \text{jika } \varepsilon \geq 0 \\ (1 - \theta) \varepsilon, & \text{jika } \varepsilon < 0 \end{cases}$$

Selanjutnya, akan dibuktikan bahwa *loss function* berbentuk asimetris dengan penjelasan berikut:

$$\rho_{\theta}(\varepsilon) = [\theta I(\varepsilon \geq 0) + (1 - \theta) I(\varepsilon < 0)] |\varepsilon|$$

ε : error dari estimasi parameter
 θ : kuantil (0,25;0,50;0,75)
 $I(\cdot)$: fungsi indikator yang terdefiniskan

Oleh karena itu, dapat ditunjukkan bahwa:

$$\rho_{\theta}(\varepsilon) = \begin{cases} \theta \varepsilon, & \varepsilon \geq 0 \\ (1 - \theta) \varepsilon, & \varepsilon < 0 \end{cases}$$

a. Untuk $\varepsilon \geq 0$

karena $\varepsilon \geq 0$, maka:

$$I(\varepsilon \geq 0) = 1$$

$$I(\varepsilon < 0) = 0$$

$$|\varepsilon| = \varepsilon$$

sehingga:

$$\begin{aligned} \rho_{\theta}(\varepsilon) &= [\theta I(\varepsilon \geq 0) + (1 - \theta) I(\varepsilon < 0)] |\varepsilon| \\ &= [\theta \cdot 1 + (1 - \theta) \cdot 0] \cdot \varepsilon \\ &= \theta \varepsilon \end{aligned}$$

b. Untuk $\varepsilon < 0$

karena $\varepsilon < 0$, maka:

$$I(\varepsilon \geq 0) = 0$$

$$I(\varepsilon < 0) = 1$$

$$|\varepsilon| = -\varepsilon$$

sehingga:

$$\begin{aligned} \rho_{\theta}(\varepsilon) &= [\theta I(\varepsilon \geq 0) + (1 - \theta) I(\varepsilon < 0)] |\varepsilon| \\ &= [\theta \cdot 0 + (1 - \theta) \cdot 1] \cdot (-\varepsilon) \\ &= (1 - \theta)(-\varepsilon) \\ &= (\theta - 1)\varepsilon \end{aligned}$$

Jika v merupakan fungsi X yang diketahui dan memiliki fungsi probabilitas $F_{y|x}(y)$ maka kuantil ke- θ dari x dapat dituliskan seperti berikut:

$$\min_{\beta} \theta \int_{i=1; \varepsilon_i \geq 0}^n |\varepsilon_i| dF_y(y) + (1 - \theta) \int_{i=1; \varepsilon_i < 0}^n |\varepsilon_i| dF_y(y)$$

bandingkan estimasi $\hat{\beta}(\theta)$, sehingga diperoleh solusi untuk permasalahan tersebut dapat berikut:



$$\hat{\beta}(\theta) = \min_{\beta \in \mathbb{R}^k} \{ \theta \sum_{i=1; \varepsilon_i \geq 0}^n |y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\theta)| + (1 - \theta) \sum_{i=1; \varepsilon_i < 0}^n |y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\theta)| \} \quad (10)$$

(Koenker, 2005).

Solusi dari Persamaan (10) tidak dapat diperoleh secara analitik, tetapi secara numerik. salah satu metode numerik yang digunakan adalah algoritma simpleks. Algoritma simpleks adalah salah satu metode pendugaan parameter secara numerik untuk regresi kuantil yang telah dikembangkan oleh Barrodan dan Robert pada tahun 1974. Algoritma simpleks memberikan solusi permasalahan program linear yang melibatkan beberapa variabel keputusan dengan bantuan komputasi (Davino, et. al, 2014).

1.6.3 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO merupakan salah satu metode regularisasi dengan penambahan penalti $L1$ (λ) pada regresi dengan metode kuadrat terkecil (MKT). Tujuan dari metode ini adalah untuk melakukan penyusutan terhadap parameter regresi, khususnya pada koefisien variabel prediktor yang memiliki korelasi tinggi terhadap error, sehingga nilai koefisien tersebut dapat menyusut mendekati nol atau tepat menjadi nol. Dengan demikian, LASSO berfungsi tidak hanya sebagai teknik regularisasi, tetapi juga sebagai metode seleksi variabel yang efektif, sekaligus mampu mengatasi masalah multikolinearitas dalam model regresi (Tibshirani, 1996).

Penduga koefisien LASSO diperoleh dengan meminimumkan fungsi tujuan berupa jumlah kuadrat sisaan (JKS) yang ditambahkan penalti λ , dengan persamaan sebagai berikut:

$$JKS_L = JKS + \lambda \sum_{j=1}^k |\beta_j| \quad (11)$$

dengan $JKS = \sum_{i=1}^n (y_i - f(X_i))^2$

$$f(X_i) = (\beta_0 + \sum_{j=1}^k \beta_j X_{ij})$$

Penduga koefisien LASSO dapat diformulasikan pada persamaan (12) berikut:

$$\hat{\beta}^L = \sum_{i=1}^n (y_i - \beta_0 + \sum_{j=1}^k \beta_j X_{ij})^2 + \lambda \sum_{j=1}^k |\beta_j| \quad (12)$$

dengan,

- n : jumlah pengamatan
- k : jumlah variabel prediktor
- y_i : nilai pengamatan variabel respon untuk pengamatan ke- i
- X_{ij} : nilai variabel prediktor ke- j pada pengamatan ke- i
- β_j : koefisien regresi untuk variabel prediktor ke- j
- β_0 : intersep model
- λ : parameter penalti yang mengontrol besarnya penyusutan

Nilai λ dapat diatur sesuai kebutuhan analisis. Semakin besar nilai λ , maka penalti terhadap koefisien regresi semakin besar, sehingga hanya sedikit variabel prediktor yang masuk ke dalam model. Sebaliknya, model kembali ke regresi linear biasa tanpa regularisasi. Secara umum, nilai λ yang menggunakan metode validasi silang, seperti yang dijelaskan dalam penelitian Atman (2022).



1.6.4 Regresi Kuantil LASSO

Penelitian yang dilakukan oleh Li & Zhu, (2008) mengembangkan model regresi kuantil dengan menambahkan penalti LASSO. Parameter λ merupakan parameter penalti yang bernilai positif. Semakin besar nilai λ , semakin banyak koefisien regresi yang disusutkan menuju atau menjadi nol, sehingga hanya variabel-variabel penting yang dipertahankan dalam model. Penduga koefisien regresi kuantil ke- θ untuk $\theta \in (0,1)$ setelah di tambahkan penalti LASSO dapat di tulis dalam bentuk *Lagrangian* sebagai berikut:

$$\hat{\beta}_L(\theta) = \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(y_i - \mathbf{X}_i^T \boldsymbol{\beta}(\theta)) + \lambda \sum_{j=1}^k |\beta_{\theta,j}| \quad (13)$$

Penduga koefisien LASSO tidak dapat diperoleh dalam bentuk tertutup seperti pada OLS atau regresi ridge, karena adanya penalti $|\beta_{\theta,j}|$ yang tidak terdeferensialkan dititik nol. Oleh karena itu, estimasi parameter diselesaikan secara numerik, salah satunya dengan menggunakan metode optimasi, seperti pemrograman kuadratik. Namun, sejak diperkenalkannya algoritma *Least Angle Regression* (LAR) oleh Efron et al. pada tahun 2004 (Hastie, et al., 2008), pendekatan LAR yang dimodifikasi dapat digunakan untuk mempercepat proses komputasi metode LASSO. Modifikasi ini memungkinkan efisiensi yang tinggi dibandingkan dengan pemrograman kuadratik. LAR merupakan algoritma regresi yang dapat dimodifikasi untuk memecahkan persoalan LASSO dengan mengikuti jalur solusi berbentuk *piecewise-linear*, dimana koefisien berubah secara bertahap.

Berikut tahapan-tahapan dalam algoritma LAR yang dimodifikasi untuk LASSO adalah sebagai berikut:

1. Seluruh variabel prediktor dibakukan (standarisasi) sehingga memiliki nilai tengah (mean) nol dan ragam (varians) satu. Tujuan dari pembakuan ini adalah agar seluruh prediktor berada pada skala yang sama, sehingga koefisien regresi dapat dibandingkan secara adil, terutama karena LASSO sensitif terhadap skala variabel.
2. Seluruh koefisien regresi diinisialisasi dengan nilai nol, yaitu $\beta_1, \beta_2, \dots, \beta_{10} = 0$. Sisaan awal dihitung sebagai $r = y - \hat{y}$, yaitu selisih antara nilai respons dan nilai rata-ratanya.
3. Variabel prediktor yang memiliki korelasi paling tinggi secara absolut dengan sisaan dipilih sebagai variabel aktif pertama. Korelasi ini menunjukkan variabel yang paling berkontribusi dalam menjelaskan variasi pada variabel respons.
4. Koefisien dari variabel terpilih diperbesar secara bertahap ke arah nilai koefisien kuadrat terkecil (least squares), hingga terdapat variabel lain yang memiliki korelasi absolut terhadap sisaan yang hampir sama besar.
5. Setelah ditemukan variabel kedua dengan korelasi yang hampir sama, arah peningkatan koefisien bergeser agar koefisien dari dua variabel aktif (X_j, X_m), bergerak bersama-sama menuju solusi kuadrat terkecil gabungan, sambil memperhatikan kontribusi masing-masing terhadap penurunan sisaan.
6. Jika selama proses peningkatan, salah satu koefisien non-nol menyusut kembali hingga mencapai nol, maka variabel tersebut dikeluarkan dari kelompok variabel aktif. Arah pergerakan koefisien kemudian dihitung ulang sesuai dengan gugus variabel aktif yang baru.
7. Langkah-langkah di atas diulang hingga semua variabel dimasukkan atau hingga tercapai jumlah langkah maksimal sebanyak $\min(n-1, k)$ di mana n adalah jumlah observasi dan k adalah jumlah variabel prediktor. Pada titik ini, diperoleh solusi kuadrat terkecil penuh.



ritas

menunjukkan terdapat hubungan di antara beberapa atau semua variabel yang l regresi. Terdapat dua jenis multikolinearitas yaitu multikolinearitas sempurna dan lak sempurna. Pada multikolinearitas sempurna terdapat hubungan linear antara satu variabel prediktor adalah fungsi linear dari variabel prediktor lain sedangkan ak sempurna terjadi apabila terdapat hubungan linear yang tidak sempurna antar

Adanya hubungan linier atau korelasi yang tinggi antar variabel prediktor dalam model regresi disebut dengan multikolinearitas. Salah satu cara untuk mengetahui keberadaan multikolinearitas dalam suatu model regresi adalah dengan menggunakan nilai *Tolerance* atau *Variance Inflation Factor* (VIF)

$$VIF_j = \frac{1}{1-R_j^2} \quad (14)$$

Dengan $j = 1, 2, \dots, p$ dan p adalah banyaknya variabel prediktor, R_j^2 adalah koefisien determinasi dari variabel prediktor yang diregresikan terhadap variabel prediktor lainnya. variabel dengan nilai $VIF > 10$ mengidentifikasi adanya multikolinearitas sehingga perlu dilakukan penanganan (Ihsan et. al, 2020)

1.6.6 Validasi Silang

Validasi silang adalah metode statistika untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua yaitu data *training* dan data *testing*. Pengukuran keberhasilan model menggunakan nilai *mean square error* (MSE). Pham (2019) menyatakan MSE dapat digunakan untuk menghitung deviasi dari hasil prediksi dengan nilai sebenarnya. Sehingga di peroleh persamaan berikut:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

Menurut James, et. al (2013), validasi silang k-lipat dilakukan dengan membagi data ke dalam k-lipat. Formula dari validasi silang k-lipat pada persamaan berikut:

$$CV_k = \frac{1}{k} \sum_{i=1}^k MSE_i$$

Menurut Izenman (2008), validasi silang yang sebaiknya digunakan adalah validasi silang 5-lipat atau 10-lipat karena menghasilkan nilai sisaan validasi silang dengan bias tinggi tetapi ragam rendah.

1.6.7 Interval Kepercayaan

Interval kepercayaan merupakan suatu rentang nilai yang memuat parameter populasi sesungguhnya, dengan batas bawah (lower) dan batas atas (upper). Kelebihan dari metode ini adalah dapat menjelaskan penyebaran data dengan menggunakan ukuran kuantil sesuai sebaran data, metode ini dapat menghasilkan hasil paling efisien untuk taksiran interval yang lebih sempit secara komputasi dibandingkan dengan metode lainnya (Zhou & Portnoy, 1996).

Dalam regresi kuantil, terdapat beberapa pendekatan yang dapat digunakan untuk menghitung interval kepercayaan dari parameter estimasi $\hat{\beta}(\theta)$. Salah satu pendekatan yang digunakan adalah pendekatan langsung, dengan setiap θ yang telah ditentukan, dengan taksiran interval formula sebagai berikut (Davino & Furno, 2014):

$$\Pr\left\{\hat{\beta}(\theta) - t_{\left(\frac{\alpha}{2}, df\right)} SE\left(\hat{\beta}(\theta)\right) \leq \beta(\theta) \leq \hat{\beta}(\theta) + t_{\left(\frac{\alpha}{2}, df\right)} SE\left(\hat{\beta}(\theta)\right)\right\} = 1 - \alpha \quad (15)$$

Menurut koenker (2005), Hao & Naiman (2007) salah satu karakteristik penting dari regresi kuantil adalah bahwa metode ini tidak mengasumsikan bahwa error berdistribusi normal. Oleh karena itu, dalam regresi kuantil, varians dari estimasi parameter tidak dapat dihitung secara langsung seperti dalam pendekatan OLS. Untuk mengatasi hal ini, digunakan asumsi asimtotik. Asimtotik adalah konsep dalam statistik yang



ketika jumlah sample semakin besar, maka distribusi dari estimator $\hat{\beta}(\theta)$ mengikuti asimtotik berikut:

$$\hat{\beta}(\theta) \xrightarrow{d} N(\hat{\beta}(\theta), Var \hat{\beta}(\theta))$$

asumsi asimtotik, dapat dihitung varians dari estimator koefisien regresi kuantil dengan asumsi sparsity, yaitu:

$$Var(\hat{\beta}(\theta)) \approx \frac{\theta(1-\theta)}{n[f(F^{-1}(\theta))]^2} (X^T X)^{-1} \quad (16)$$

dengan

$\theta(1 - \theta)$: varian dari *loss function* kuantil

n : jumlah pengamatan

$f(F^{-1}(\theta))$: fungsi densitas dari error pada kuantil ke- θ dihitung dengan kernel density estimator

$(\mathbf{X}^T \mathbf{X})^{-1}$: invers dari perkalian matriks transpose dikali matriks

$S(\theta)$: fungsi sparsity

Fungsi densitas $f(F^{-1}(\theta))$ merepresentasikan kepadatan dari distribusi error pada kuantil ke- θ , dengan $S(\theta) = \frac{1}{f(F^{-1}(\theta))}$ disebut sebagai fungsi sparsity. Fungsi inilah yang digunakan untuk menghitung varians dan selanjutnya untuk menghitung standar error.

Dari persamaan (16) tersebut, diperoleh standar error:

$$SE(\hat{\beta}(\theta)) = \sqrt{\text{Var}(\hat{\beta}(\theta))} \quad (17)$$

kemudian dibentuk interval kepercayaan berdasarkan distribusi normal.

1.6.8 Tuberkulosis

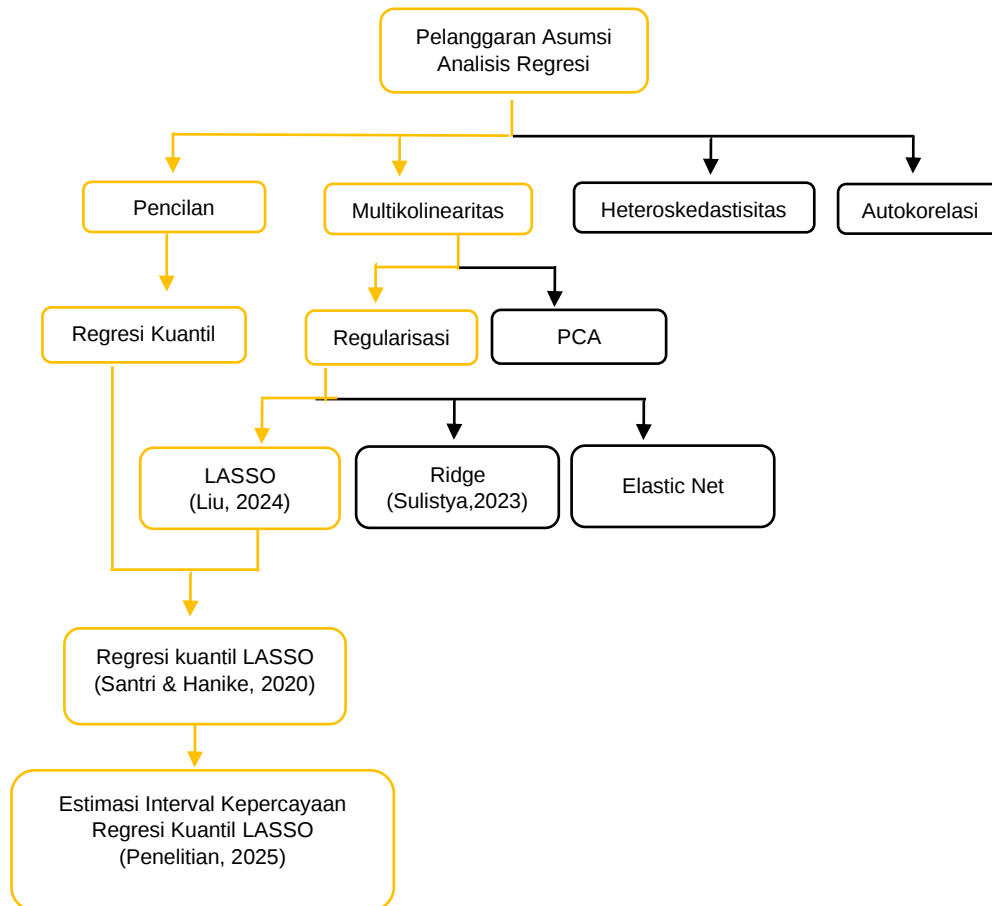
TB adalah suatu penyakit kronik menular yang disebabkan oleh bakteri *Mycobacterium tuberculosis*. Sebagian besar bakteri *tuberculosis* menyerang paru-paru tetapi juga dapat menyerang organ tubuh lainnya. Jumlah kasus yang terjadi setiap tahun juga dapat diturunkan dengan mengurangi prevalensi faktor risiko terkait kesehatan untuk tuberkulosis, misalnya merokok, diabetes, dan infeksi HIV, memberikan pengobatan pencegahan kepada orang dengan infeksi TB laten, tindakan pada determinan infeksi dan penyakit tuberkulosis yang lebih luas misalnya kemiskinan, kualitas perumahan dan kekurangan gizi. Orang yang terinfeksi bakteri TB berisiko 5-10% mengalami penyakit tuberkulosis sepanjang hidupnya. (WHO, 2021).

Penularan tuberkulosis biasanya terjadi di dalam ruangan yang gelap, dengan minim ventilasi di mana percik relik dapat bertahan di udara dalam waktu yang lebih lama. Cahaya matahari langsung dapat membunuh tuberkel basil dengan cepat, namun bakteri ini akan bertahan lebih lama di dalam keadaan yang gelap. Kontak dekat dalam waktu yang lama dengan orang terinfeksi meningkatkan risiko penularan. Apabila terinfeksi, proses sehingga paparan tersebut berkembang menjadi penyakit tuberkulosis aktif bergantung pada kondisi imun individu. Pada individu dengan sistem imun yang normal, 90% tidak akan berkembang menjadi penyakit tuberkulosis dan hanya 10% dari kasus akan menjadi penyakit tuberkulosis aktif (setengah kasus terjadi segera setelah terinfeksi dan setengahnya terjadi di kemudian hari). Risiko paling tinggi terdapat pada dua tahun pertama pasca-terinfeksi, dimana setengah dari kasus terjadi. Kelompok dengan risiko tertinggi terinfeksi adalah anak-anak dibawah usia 5 tahun dan lanjut usia. Orang yang mengalami kekurangan gizi 3 kali lebih berisiko dengan kondisi imun buruk lebih rentan mengalami penyakit TB aktif dibanding orang dengan kondisi sistem imun yang normal, secara global pada tahun 2020 terdapat 1,9 juta kasus TBD baru yang disebabkan karena kekurangan gizi. 50- 60% orang dengan HIV-positif yang terinfeksi TB akan mengalami penyakit TB yang aktif.



1.6.9 Kerangka Konsep

Kerangka konsep pada penelitian memiliki kaitan atau hubungan antara kerangka konsep satu dengan kerangka konsep lainnya dari masalah yang ingin diteliti. Kerangka konsep diperoleh dari konsep ilmu atau teori yang dipakai sebagai landasan penelitian. Adapun gambaran kerangka konsep dalam penelitian ini seperti pada Gambar 1 berikut.



Gambar 1. Kerangka Konsep



BAB II METODOLOGI PENELITIAN

2.1 Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh melalui profil kesehatan Indonesia tahun 2021 oleh kementerian kesehatan RI dengan alamat resmi <https://www.kemkes.go.id/>, Badan pusat statistika dengan alamat resmi <https://www.bps.go.id/> dan Kementerian dalam negeri dengan alamat resmi <https://gis.dukcapil.kemendagri.go.id/>.

2.2 Variabel Penelitian

Variabel yang digunakan pada penelitian ini terdiri dari variabel respon (Y) dan variabel prediktor (X). variabel pada penelitian disajikan pada Tabel 2 berikut:

Tabel 1. Variabel Penelitian

Variabel	Nama Variabel	Satuan	Definisi
Y	Jumlah kasus tuberkulosis tahun 2021 di Indonesia	Kasus	Kasus tuberkulosis adalah jumlah kasus terkonfirmasi tuberkulosis setiap provinsi di Indonesia.
X_1	Jumlah puskesmas	Unit	Jumlah puskesmas di setiap provinsi.
X_2	Jumlah rumah sakit	Unit	Jumlah rumah sakit umum di setiap provinsi
X_3	Jumlah tenaga medis	Orang	Jumlah tenaga medis di fasilitas kesehatan, meliputi dokter spesialis dan dokter umum di setiap provinsi.
X_4	Jumlah kasus HIV	Kasus	Jumlah kasus positif HIV di setiap provinsi
X_5	Presentase perokok penduduk umur 15 tahun ke atas	Persen	Hasil bagi antara jumlah perokok aktif penduduk aktif penduduk umur 15 tahun ke atas dengan jumlah penduduk di setiap provinsi dikalikan 100.
X_6	Presentase kabupaten/kota yang menerapkan Kawasan tanpa rokok	Persen	Hasil bagi antara jumlah kabupaten/kota yang menerapkan kawasan tanpa rokok dengan jumlah kabupaten/kota di setiap provinsi dikalikan 100.
X_7	Pengeluaran pemerintah di bidang kesehatan	Miliar Rp	Realisasi anggaran dekonsentrasi kementerian kesehatan.
X_8	Presentase penduduk miskin	Persen	Hasil bagi antara jumlah penduduk miskin dengan jumlah penduduk di setiap provinsi dikalikan 100.
X_9	Kepadatan penduduk	Jiwa/km ²	Hasil bagi antara jumlah penduduk dengan luas wilayah di setiap provinsi.
X_{10}	Jumlah lansia	Orang	Jumlah penduduk yang berusia 60 tahun ke atas.

2.3 Metode Analisis

Langkah-langkah yang akan dilakukan pada penelitian adalah sebagai berikut:



1. bentuk estimasi parameter pada model regresi kuantil LASSO sebagai berikut:

2. himpunan data berpasangan $(y_i, X_{ij}), i = 1, 2, \dots, n$ dan

3. $\dots, k$. Data mengandung pencilan sehingga dilakukan analisis regresi kuantil yang

4. pada persamaan umum regresi kuantil linier khusus untuk kuantil bersyarat $Q_{y|x}(\theta)$

5. variabel respon y_i , yaitu:

$$y_i = \beta_0(\theta) + \beta_1(\theta)X_{i1} + \dots + \beta_k(\theta)X_{ik} + \varepsilon_i(\theta)$$

- ii. Menyajikan model regresi kuantil disajikan dalam matriks, persamaan dapat ditulis sebagai berikut:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}(\boldsymbol{\theta})$$

- iii. Menyajikan solusi optimasi dengan meminimumkan error pada regresi kuantil dengan fungsi sebagai berikut:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta} \left(y_i - \mathbf{X}_i^T \boldsymbol{\beta}(\boldsymbol{\theta}) \right), \quad i = 1, 2, \dots, n.$$

- iv. Data yang digunakan juga terjadi multikolinearitas sehingga untuk mengatasi masalah tersebut dilakukan analisis regresi LASSO dengan penambahan penalti $L1(\lambda)$, Bentuk model regresi kuantil LASSO sebagai berikut:

$$\hat{\boldsymbol{\beta}}_L(\theta) = \min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^n \rho_{\theta}(y_i - \mathbf{X}_i^T \boldsymbol{\beta}(\boldsymbol{\theta})) + \lambda \sum_{j=1}^k |\beta_j|$$

Estimasi dilakukan dengan menggunakan algoritma komputasi.

- v. Mendapatkan interval kepercayaan dengan menggunakan pendekatan fungsi sparsity sebagai berikut :

1. Untuk mendapatkan interval kepercayaan membutuhkan distribusi dari estimator $\hat{\boldsymbol{\beta}}^L(\theta)$ dalam pendekatan asimtotik sebagai berikut:

$$\hat{\boldsymbol{\beta}}^L(\theta) \sim N(\boldsymbol{\beta}^L(\theta), \text{Var}(\hat{\boldsymbol{\beta}}^L(\theta)))$$

2. Dimisalkan T adalah statistik uji yang telah distandarisasi maka diperoleh

$$T = \frac{\hat{\boldsymbol{\beta}}^L(\theta) - \boldsymbol{\beta}^L(\theta)}{SE(\hat{\boldsymbol{\beta}}^L(\theta))} \sim N(0,1)$$

3. Adapun fungsi probabilitas didapatkan sebagai berikut:

$$P \left(-z_{1-\frac{\alpha}{2}} \cdot SE \hat{\boldsymbol{\beta}}^L(\theta) \leq \hat{\boldsymbol{\beta}}^L(\theta) - \boldsymbol{\beta}^L(\theta) \leq z_{1-\frac{\alpha}{2}} \cdot SE \hat{\boldsymbol{\beta}}^L(\theta) \right) = 1 - \alpha$$

4. Berdasarkan langkah 3 di dapatkan interval kepercayaan regresi kuantil LASSO.

2. Penerapan model interval kepercayaan regresi kuantil LASSO pada data jumlah kasus tuberkulosis di indonesia tahun 2021.

- Melakukan eksplorasi pada data yang digunakan.
- Melakukan standarisasi data
- Pemilihan kuantil dengan nilai kuantil yang digunakan adalah 0,25, 0,50, dan 0,75.
- Menentukan nilai lambda optimal
- Pemodelan regresi kuantil LASSO dengan menggunakan nilai lambda optimal.

dan estimasi interval kepercayaan dari regresi kuantil LASSO pada jumlah kasus tuberkulosis di indonesia tahun 2021
interpretasi hasil dan kesimpulan

