BAB I
PENDAHULUAN

1.1 Latar Belakang

Analisis regresi adalah metode statistik yang digunakan untuk memahami hubungan antara variabel dependen dan satu atau lebih variabel independen. Metode ini krusial dalam analisis data karena membantu mengidentifikasi faktor-faktor signifikan yang memengaruhi variabel dependen dan memungkinkan pembuatan model prediktif yang andal (Dani & Adrianingsih, 2021). Ketika data memiliki variabel dependen yang kontinu dapat menggunakan regresi linear umum. Namun, untuk variabel dependen yang bersifat kategorik, regresi logistik menjadi pilihan yang sangat efektif. Regresi logistik merupakan pemodelan variabel kualitatif dalam dua kategori (biner) atau lebih dari dua kategori (multinomial), serta menyediakan kerangka probabilitas untuk memprediksi peluang terjadinya suatu peristiwa berdasarkan variabel independen. Pendekatan ini memberikan manfaat dalam analisis data kategorik dan kompleks, terutama ketika pemahaman tentang probabilitas kejadian sangat dibutuhkan (Yumira dkk., 2017).

Regresi logistik biner adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen biner dan sejumlah variabel independen. Metode ini sering diterapkan untuk memprediksi peluang terjadinya peristiwa seperti diagnosis penyakit atau keputusan pembelian. Analisis pada metode ini melibatkan penggunaan fungsi logit untuk menghubungkan variabel dependen biner dengan variabel independen melalui transformasi logaritmik. Hal ini memungkinkan interpretasi hasil dalam bentuk probabilitas, yang memudahkan pemahaman dan penerapan di berbagai bidang (Wibowo dkk., 2021). Selain itu, metode ini juga menyediakan parameter *odds ratio*, yang berguna untuk mengukur kekuatan asosiasi antara variabel independen dan peluang terjadinya suatu peristiwa. Namun, metode ini memiliki kelemahan, terutama ketika dihadapkan pada masalah multikolinearitas.

Multikolinearitas adalah kondisi ketika dua atau lebih variabel independen dalam model regresi memiliki korelasi tinggi satu sama lain. Kondisi ini bisa menyebabkan masalah serius dalam analisis regresi, seperti mengurangi stabilitas estimasi parameter, memperbesar *standar error*, dan menurunkan keakuratan model (Azizah dkk., 2021) Untuk mengatasi multikolinearitas, biasanya dilakukan dengan menghapus atau menambah variabel independen dalam penelitian, atau menggunakan teknik regularisasi. Salah satu teknik yang sering digunakan adalah regresi *ridge*. Metode ini menambahkan penalti terhadap besarnya koefisien regresi, sehingga mengurangi variabilitas estimasi parameter dan meningkatkan stabilitas model (Putri & Suliadi, 2023). Selain itu, regresi *ridge* mampu menjaga stabilitas dan keakuratan model meskipun terdapat multikolinearitas di antara variabel independen, sehingga cocok untuk analisis data dengan hubungan variabel yang kompleks (Zahari dkk., 2014). Putra dan Ratnasari (2015) menggunakan regresi logistik *ridge* untuk memodelkan Indeks Pembangunan Manusia (IPM) di Jawa Timur yang

mengatasi masalah multikolinearitas antara variabel independen. Hasilnya menunjukkan bahwa model tersebut mampu menjelaskan IPM dengan akurasi mencapai 97.37%.

Optimasi model regresi logistik biner dapat dilakukan dengan berbagai pendekatan, termasuk metode yang lebih fleksibel untuk menangkap pola hubungan yang kompleks. Salah satu pendekatan tersebut adalah regresi nonparametrik, yang tidak memerlukan asumsi khusus tentang bentuk hubungan antara variabel dependen dan independen. Hal ini membuatnya sangat efektif dalam menganalisis pola yang tidak linier atau tidak teratur (Islamiyati, 2014). Salah satu metode populer dalam regresi nonparametrik adalah *spline truncated*. Metode ini merupakan metode yang membagi data menjadi beberapa segmen dan menyesuaikan bentuk kurva, sehingga menghasilkan model yang adaptif terhadap perubahan pola data. Penelitian oleh Dani dan Adrianingsih (2021) menunjukkan bahwa *spline truncated* efektif dalam memodelkan data nonlinier dengan nilai R^2 yang tinggi. Selain Itu, Metode ini juga mampu mengurangi *Mean Squared Error* (MSE) serta meningkatkan akurasi model secara signifikan dengan membagi domain data menjadi segmen-segmen dan mencocokkan fungsi polinomial di tiap segmen (Nendi & Wibowo, 2019).

Beberapa penelitian telah melakukan kombinasi regresi logistik dengan *spline truncated*. Islamiyati dkk. (2023) menggunakan model regresi logistik *spline* biner untuk menganalisis data status gizi anak di Kabupaten Barru, Sulawesi Selatan. Hasil yang diperoleh menunjukkan bahwa model tersebut mampu menjelaskan status gizi anak dengan akurasi klasifikasi sebesar 87.5%. Sementara itu, Arifin dkk. (2023) menerapkan model regresi logistik ordinal *spline* pada data gizi balita dengan tiga kategori gizi di Gowa. Hasil yang diperoleh menunjukkan bahwa model tersebut mampu menjelaskan gizi balita dengan capaian akurasi sebesar 92.25% dan menunjukkan bahwa anak usia 18-24 bulan memiliki pola nutrisi yang berbeda. Namun, penelitian-penelitian tersebut belum mempertimbangkan adanya multikolinearitas, terutama pada data yang memiliki hubungan kompleks antara variabel-variabel independen, seperti data status gizi balita yang dapat memengaruhi stabilitas dan akurasi estimasi parameter dalam model regresi.

Status gizi balita adalah indikator penting untuk menilai kualitas kesehatan anak-anak di suatu daerah. Anak-anak yang mengalami kekurangan gizi, baik gizi buruk maupun gizi kurang, rentan terhadap gangguan perkembangan fisik dan mental. Gizi buruk, termasuk *wasting* (kerdil) dan *underweight* (berat badan rendah), dapat menurunkan daya tahan tubuh anak terhadap infeksi, meningkatkan risiko kematian, serta memengaruhi kemampuan belajar dan tumbuh kembangnya (Yulianto dkk., 2022). Selain itu, gizi buruk juga mengganggu perkembangan otak, yang berisiko menyebabkan keterlambatan kognitif, kesulitan belajar, dan gangguan motorik. Anak-anak yang mengalami gizi buruk lebih rentan terhadap infeksi, dan kekurangan gizi yang berkepanjangan dapat memengaruhi kesehatan mereka hingga dewasa. Salah satu dampak paling serius dari gizi buruk adalah *stunting*, yaitu gangguan pertumbuhan yang menyebabkan tinggi badan anak lebih rendah dari standar usianya. Pada tahun 2023, prevalensi *stunting* di Kabupaten Gowa tercatat sebesar 21.5% (SGI, 2023), yang menunjukkan bahwa tantangan dalam

penanganan gizi buruk dan stunting masih perlu menjadi prioritas utama untuk memastikan kesehatan anak-anak di masa depan.

Beberapa faktor yang diduga memengaruhi status gizi balita adalah berat badan, tinggi badan, dan usia (Arifin dkk., 2023). Selain itu, penelitian lain juga menyebutkan bahwa status gizi balita dipengaruhi oleh berat badan lahir dan tinggi badan lahir (Rohmah & Nadhiroh, 2024). Berbagai faktor ini membuat hubungan antara variabel independen dan status gizi balita menjadi lebih kompleks, sehingga memungkinkan terjadi masalah multikolinearitas saat melakukan analisis data. Oleh karena itu, diperlukan pendekatan yang lebih adaptif. Metode regresi logistik *ridge* dengan estimator *spline truncated* dapat menjadi solusi karena kemampuannya menangkap pola data yang kompleks dan memastikan kestabilan model. Penelitian ini akan berfokus pada pembuatan dan penerapan model tersebut untuk memenuhi kebutuhan analisis yang lebih akurat serta memberikan dasar ilmiah yang kuat guna mendukung upaya intervensi gizi yang lebih terarah dan berbasis bukti. Dengan memanfaatkan temuan penelitian sebelumnya, penelitian ini akan mengidentifikasi faktor risiko dan pola pertumbuhan balita. Ini diharapkan dapat meningkatkan analisis data dan memungkinkan intervensi yang lebih efektif.

1.2 Batasan Masalah

Untuk membatasi ruang lingkup permasalahan pada penelitian ini, maka diberikan beberapa batasan bahwa:

1. Pemodelan *spline truncated* dalam regresi logistik biner *ridge* hanya dibatasi pada orde satu (linier) dan orde dua (kuadratik)
2. Metode penaksiran parameter yang digunakan adalah *Maximum Likelihood Estimation* (MLE).
3. Pemilihan titik knot optimal yang dilakukan hanya sampai pada dua titik knot yang dipilih berdasarkan metode *Generalized Cross-Validation* (GCV).

1.3 Tujuan Penelitian dan Manfaat Penelitian

Berdasarkan latar belakang, maka diperoleh tujuan penelitian sebagai berikut:

1. Mengestimasi parameter model regresi logistik biner *ridge* dengan estimator *spline truncated* pada data yang mengandung multikolinieritas.
2. Memodelkan hubungan antara status gizi balita dengan faktor-faktor yang memengaruhinya di Kabupaten Gowa pada tahun 2023, dengan pendekatan regresi logistik biner *ridge* dengan estimator *spline truncated*.

Adapun penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan wawasan mengenai faktor-faktor yang memengaruhi status gizi balita di Kabupaten Gowa.
2. Sebagai sumber pengetahuan dan informasi mengenai model status gizi balita yang dihasilkan melalui regresi logistik biner *ridge* dengan estimator *spline truncated*.

1.4 Teori

1.4.1 Multikolinieritas

Multikolinearitas terjadi ketika terdapat hubungan linear yang hampir sempurna antara kolom-kolom pada matriks X. Jika hubungan linear tersebut sempurna, hal ini akan menyebabkan nilai determinan $X^T X$ menjadi nol, yang dikenal sebagai multikolinearitas sempurna (Guan & Fu, 2022). Multikolinearitas memengaruhi varians koefisien regresi, yang pada akhirnya dapat mengurangi keandalan hasil regresi. Ketika variabel-variabel independen memiliki korelasi tinggi, interpretasi pengaruh setiap variabel terhadap variabel dependen menjadi tidak jelas (Shrestha, 2020). Salah satu metode paling sederhana dan umum untuk mendeteksi multikolinearitas adalah melalui analisis matriks korelasi. Metode ini menilai tingkat hubungan antara variabel independen dengan menggunakan koefisien korelasi *Pearson*, yang berkisar antara -1 dan 1 (Upendra dkk., 2023). Koefisien korelasi *Pearson* r dapat dihitung dengan persamaan (14) berikut

$$r_{ab} = \frac{n(\sum x_a x_b) - (\sum x_a)(\sum x_b)}{\sqrt{[n\sum x_a^2 - (\sum x_a)^2]\left[n\sum x_b^2 - (\sum x_b)^2\right]}}; a = 1,2,\dots,p \text{ dan } b = 1,2,\dots,p \qquad (1)$$

Koefisien korelasi *pearson* antara dua variabel independen, x_a dan x_b, dilambangkan dengan r_{ab}. Dalam hal ini, n merupakan jumlah observasi, x_a adalah variabel independen $ke-a$, dan x_b adalah variabel independen $ke-b$. Setelah semua koefisien korelasi ini dihitung, hasilnya dapat disusun dalam bentuk matriks korelasi $\boldsymbol{R}$, yang merepresentasikan hubungan antara variabel independen dalam model regresi. Matriks korelasi dirumusukan sebagai berikut:

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

Matriks $\boldsymbol{R}$ memiliki sifat bahwa elemen diagonal utama (r_{ab}) selalu bernilai 1, karena setiap variabel memiliki korelasi sempurna dengan dirinya sendiri. Dengan kata lain, jika $a = b$, maka $r_{ab} = 1$. Dalam praktiknya, nilai koefisien korelasi r_{ab} yang melebihi 0,8 atau kurang dari $-0{,}8$ dianggap menunjukkan korelasi yang sangat tinggi antara variabel independen x_a dan x_b. Korelasi yang tinggi ini dapat menjadi indikasi adanya masalah multikolinearitas, yang dapat mengurangi keakuratan model regresi (Kim, 2019).

Untuk menguji signifikansi koefisien korelasi r_{ab}, kita dapat merumuskan hipotesis sebagai berikut:

H_0: $r_{ab} = 0$, untuk $a \neq b$

H_1: $r_{ab} \neq 0$, untuk $a \neq b$

Uji signifikansi koefisien korelasi dilakukan dengan menggunakan statistik uji t yang dihitung dengan rumus:

$$t = \frac{r_{ab}\sqrt{n-2}}{\sqrt{1-r_{ab}^2}} \qquad (2)$$

dengan n adalah jumlah observasi. Statistik t ini kemudian dibandingkan dengan nilai kritis t dari tabel distribusi t dengan derajat kebebasan $df = n - 2$ pada tingkat signifikansi ($\alpha = 0.05$)). Kriteria pengujian adalah sebagai berikut:

- Jika $|t| > t_{kritis}$, maka H_0 ditolak, yang berarti hubungan linear antara x_a dan x_b signifikan secara statistik.
- Jika $|t| \leq t_{kritis}$, maka H_0 gagal ditolak, yang berarti hubungan linear antara x_a dan x_b tidak signifikan secara statistik.

Jika hasil uji t menunjukkan bahwa H_0 ditolak untuk r_{ab}, maka hubungan linear antara variabel independen signifikan secara statistik, yang memperkuat indikasi adanya multikolinearitas (Sanny & Dewi, 2020).

1.4.2 Regresi Logistik Biner

Regresi logistik biner merupakan model statistik yang bertujuan untuk menganalisis hubungan antara satu atau lebih variabel independen dengan variabel dependen yang bersifat kategori dan terdiri dari dua kategori, seperti 'sukses' dan 'gagal' atau 'ya' dan 'tidak'. Model ini mengestimasi probabilitas terjadinya suatu peristiwa berdasarkan fungsi logit dari kombinasi linear variabel independen, yang dapat berupa numerik maupun kategori (Sofiah & Hajarisman, 2023).

Model regresi logistik biner mengasumsikan variabel dependen y sebagai variabel biner dengan dua kategori, yaitu $y = 0$ atau $y = 1$. Misalkan untuk setiap pengamatan i, kita mendefinisikan $x_i = x_{i1}, x_{i2}, \dots, x_{ip}$ sebagai variabel independen yang terdiri dari p variabel. Model regresi logistik mengasumsikan bahwa variabel dependen Y_i pada pengamatan ke-i mengikuti distribusi bernoulli, dengan probabilitas sukses $\pi(x_i)$ untuk $y_i = 1$, yang ditunjukan pada Persamaan (3).

$$y_i|x_{i1}, x_{i2}, \dots, x_{ip} \sim Bernoulli(\pi(x_i)) \tag{3}$$

dengan $\pi(x_i)$ adalah probabilitas sukses $y_i = 1$ yang dihitung berdasarkan fungsi logistik dari kombinasi linier dari variabel independen, yang dinyatakan pada Persamaan (4) :

$$\pi(x_i) = P(y_i = 1|x_i) = \frac{e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}} \tag{4}$$

Persamaan (4) merupakan fungsi logit dengan β_0 adalah intersep (Koefisien untuk konstanta) dan β_j adalah koefisien variabel independen $x_{i1}, x_{i2}, \dots, x_{ip}$. Fungsi pada Persamaan (4) mengubah kombinasi linier variabel independen menjadi probabilitas yang memungkinkan model untuk memprediksi peluang terjadinya kejadian biner.

Untuk mempermudah estimasi parameter, digunakan transformasi logit dari probabilitas $\pi(x_i)$ yang mengubah probabilitas menjadi *log-odds* yang ditunjukan pada persamaan (5):

$$g(x_i) = logit[\pi(x_i)] = ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] \tag{5}$$

dengan transformasi logit, model regresi logistik menghubungkan *log-odds* dengan kombinasi linier variabel independen. Menggunakan Persamaan (5), bentuk linier terhadap parameter regresi dapat dituliskan seperti pada Persamaan (6):

$$g(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{6}$$

dengan $g(x_i)$ adalah *log-odds*, yang menghubungkan kombinasi linier variabel independen dengan hasil yang diinginkan.

1.4.3 *Maximum Likelihood Estimation* (MLE)

Tujuan utama dalam regresi logistik biner adalah mengestimasi parameter β yang memaksimalkan probabilitas bahwa model merefleksikan data yang diamati. Model ini diasumsikan mengikuti distribusi Bernoulli, yang menggambarkan probabilitas kejadian $\pi(x_i)$, yaitu probabilitas $y = 1$ berdasarkan kombinasi linier dari parameter β dan variabel independen x_{ij} seperti yang telah dijelaskan dalam persamaan (4). Setiap probabilitas individu pada pengamatan ke-i dapat dituliskan sebagai:

$$P(\, y_i = 1 \mid x_i \,) = \pi(x_i) \quad \text{dan} \quad P(\, y_i = 0 \mid x_i \,) = 1 - \pi(x_i) \tag{7}$$

dengan demikian, fungsi *likelihood* $L(\beta)$ untuk seluruh n pengamatan adalah hasil perkalian dari probabilitas untuk masing-masing pengamatan, yang dirumuskan sebagai berikut:

$$L(\beta) = \prod_{i=1}^{n} P(y_i|\boldsymbol{X}_i) = \prod_{i=1}^{n} \left[\pi(x_i)^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i}\right] \tag{8}$$

Probabilitas gabungan bahwa seluruh data yang diamati sesuai dengan model yang dibangun dijelaskan oleh fungsi *likelihood* pada Persamaan (7). Untuk menyederhanakan perhitungan, fungsi *likelihood* ini diubah menjadi fungsi log-*likelihood* $\ell(\beta)$ dengan mengambil logaritma natural dari $L(\beta)$, sehingga menghasilkan Persamaan (9):

$$\ell(\beta) = \sum_{i=1}^{n} \big[y_i \ln \pi\,(x_i) + (1 - y_i) \ln\big(1 - \pi(x_i)\big)\big] \tag{9}$$

Fungsi *log-likelihood* dipilih karena secara numerik lebih stabil dan lebih mudah untuk didiferensiasi dibandingkan dengan fungsi *likelihood* yang asli. Untuk menentukan nilai parameter β yang memaksimalkan fungsi *log-likelihood*, maka melakukan diferensiasi terhadap parameter β_j. Turunan pertama dari *log-likelihood* terhadap β_j datap dilihat pada persamaan (10):

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \big(y_i - \pi(x_i)\big) x_{ij} \tag{10}$$

Persamaan (10) di kenal sebagai *gradient* dan mengukur kontribusi setiap pengamat terhadap perubahan parameter β_j. Residual $y_i - \pi(x_i)$ mengambarkan deviasi atara nilai aktual dan nilai prediksi. Untuk menghitung turunan dari fungsi *sigmoid* $\pi(x_i)$ terhadap β_j digunakan rumus pada Persamaan (11):

$$\frac{\partial \pi(x_i)}{\partial \beta_j} = \pi(x_i)\big(1 - \pi(x_i)\big) x_{ij} \tag{11}$$

dengan demikian, persamaan untuk turunana pertama log-likelihood menjadi sistem persamaan nin-linier yang membutuhkan metode numerik untuk diselesaikan.

Metode *Newton-Raphson* merupakan metode iteratif yang sering digunakan untuk menyelesaikan sistem persamaan non-linier, dengan memperbarui nilai parameter β menggunakan informasi dari gradien dan matriks Hessian (matriks turunan kedua *log-likelihood*). Pembaruan parameter pada iterasi ke$-t$ dilakukan dengan rumus pada Persamaan (12) berikut:

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{H}^{-1}\boldsymbol{g} \tag{12}$$

dengan

$\boldsymbol{g}$: *gradient* (vektor turunan pertama *log-likelihood*)

$\boldsymbol{H}$: *Matriks Hessian(matriks turunan kedua log-likelihood)*,

Iterasi pada Persamaan (12) berlangsung hingga perbedaan antar parameter pada setiap langkah iterasi menjadi sangat kecil, menunjukkan bahwa konvergensi telah tercapai. Proses ini memungkinkan estimasi parameter yang paling sesuai dengan data yang diamati (Erviana & Karyana, 2022).

1.4.4 Regresi Logistik Biner *Ridge*

Regresi logistik biner *ridge* adalah pengembangan dari regresi logistik biner, yang bertujuan untuk mengatasi masalah multikolinearitas dan *overfitting.* Teknik ini melibatkan penambahan penalti *ridge* (regularisasi $L2$) ke dalam fungsi log-likelihood standar untuk mengecilkan nilai besar pada parameter β_j. Penalti ini berfungsi untuk menjaga stabilitas model, terutama ketika ada banyak variabel independen atau ketika terdapat korelasi tinggi antar variabel (Jadhav, 2020).

Fungsi log-likelihood dalam regresi logistik biner dijelaskan pertama kali melalui Persamaan (9). Pada regresi logistik *ridge*, fungsi ini dimodifikasi dengan menambahkan penalti *ridge* $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$, menghasilkan bentuk yang dijelaskan pada Persamaan (13)

$$\ell_{ridge}(\beta) = \ell(\beta) - \lambda\|\beta\|^2$$

$$\ell_{ridge}(\beta) = \sum_{i=1}^{n} \left[y_i \ln \pi\,(x_i) + (1 - y_i) \ln\big(1 - \pi(x_i)\big)\right] - \lambda\|\beta\|^2 \tag{13}$$

dengan λ adalah parameter penalti yang mengontrol kekuatan regulasi dan penalti $\|\beta\|^2$ digunakan untuk mengurangi nilai besar pada parameter β_j. Penambahan penalti ini membantu mengontrol ukuran koefisien model, sehingga mencegah *overfitting* dan memastikan estimasi parameter lebih stabil (Guan & Fu, 2022).

Turunan pertama dari *log-likelihood* terhadap β_j (gradien*)* yang digunakan untuk estimasi parameter β pada regresi logistik diberikan pada Persamaan (10). Ketika penalti *ridge* ditambahkan, gradien dimodifikasi menjadi:

$$g_{ridge}(\beta) = U(\beta) - 2\lambda\beta$$

$$g_{ridge}(\beta) = \sum_{i=1}^{n} \big(y_i - \pi(x_i)\big)x_{ij} - 2\lambda\beta \tag{14}$$

Penambahan penalti $-2\lambda\beta$ memastikan parameter β_j tetap kecil, sehingga menjaga stabilitas estimasi, terutama dalam kondisi multikolinearitas. Estimasi parameter β pada metode regresi logistik *ridge* dilakukan dengan metode iteratif *Newton-Raphson*, yang memperbarui nilai parameter berdasarkan gradien $\mathrm{U}_{\mathrm{ridge}}(\beta)$

dan matriks Hessian. Persamaan pembaruan parameter pada iterasi ke−t dinyatakan sebagai :

$$\beta^{(t+1)} = \beta^{(t)} - \boldsymbol{H_{ridge}}^{-1}\boldsymbol{g_{ridge}}\left(\beta^{(t)}\right) \quad (15)$$

dengan $\mathrm{H}_{\mathrm{ridge}}$ adalah matriks Hessian yang telah disesuaikan dengan penalti *ridge* $\mathrm{H}_{\mathrm{ridge}} = \mathrm{H} + 2\lambda\mathrm{I}$ (I adalah matriks identitas berukuran $(p\ x\ p)$ dan $g_{\mathrm{ridge}}(\beta)$ dihitung berdasarkan $g\left(\beta^{(t)}\right) + 2\lambda\beta^{(t)}$ dengan $2\lambda\beta^{(t)}$ adalah komponen penalti *ridge*. Penyesuaian ini memastikan bahwa penalti *ridge* memengaruhi sensitivitas pembaruan parameter, sehingga hasil estimasi lebih stabil.

Parameter regularisasi λ berada dalam rentang $0 \leq \lambda \leq 1$, dengan nilai $\lambda = 0$ akan menghasilkan model regresi biasa tanpa penalti, sementara $\lambda = 1$ akan memberikan penalti maksimal terhadap koefisien regresi. Pemilihan nilai λ yang tepat sangat penting, karena nilai yang terlalu besar akan menyebabkan model *underfitting*, sedangkan nilai yang terlalu kecil dapat menyebabkan model overfitting (Putri & Suliadi, 2023).Oleh karena itu, *Generalized Cross Validation* (GCV) sering digunakan untuk memilih nilai λ yang optimal, yang memberikan keseimbangan antara bias dan varians dalam model. Dengan adanya penalti ini, regresi *ridge* mampu mengurangi multikolinearitas dan menjaga stabilitas estimasi meskipun terdapat hubungan korelasi tinggi antara variabel independen. Sebagai hasilnya, model regresi menjadi lebih dapat diandalkan untuk melakukan prediksi dan memberikan interpretasi koefisien yang lebih stabil.

1.4.5 Pengujian Signifikansi Parameter

Pengujian signifikansi parameter adalah langkah yang sangat penting dalam regresi logistik biner, karena berfungsi untuk mengevaluasi sejauh mana variabel independen memengaruhi variabel dependen yang bersifat biner (Rohmah dkk., 2023). Ada dua jenis pengujian yang umumnya dilakukan dalam analisis ini, yaitu uji simultan yang menguji pengaruh seluruh variabel independen secara bersamaan, dan uji parsial yang menilai pengaruh masing-masing variabel independen terhadap variabel dependen secara individual.

Salah satu metode pengujian simultan yang umum digunakan dalam regresi logistik adalah uji *Likelihood Ratio* (LR). Metode ini membandingkan performa model penuh, yang mencakup semua variabel independen, dengan model terbatas yang tanpa variabel independen. Dengan membandingkan kedua model ini, uji LR memberikan wawasan tentang relevansi keseluruhan model serta kontribusi signifikan dari variabel independen terhadap prediksi variabel dependen (Simbolon dkk., 2019).

H_0: $\beta_1 = \beta_2 = \ ... = \beta_j = 0$ (tidak ada pengaruh signifikan)
H_1: Minimal terdapat satu $\beta_j \neq 0$, untuk $j = 1,2, ... p$

Statistik uji LR didasarkan pada perbandingan antara log-likelihood dari model penuh (lnL_1) dan model terbatas (lnL_0). Statistik uji LR dirumuskan pada persamaan (16):

$$D = -2 \times (lnL_0 - lnL_1) \quad (16)$$

Statistik D mengikuti distribusi *chi-square* (χ^2) dengan derajat kebebasan (df) sama dengan jumlah parameter β_j yang diuji. Penolakan hipotesis dilakukan jika nilai $D > \chi^2_{p;\alpha}$ atau $p - value < \alpha$, yang menunjukkan bahwa model dengan variabel independen lebih efektif dalam menjelaskan data.

Selain uji simultan, uji parsial dilakukan untuk mengevaluasi pengaruh masing-masing variabel independen secara individu terhadap variabel dependen. Untuk itu, digunakan Uji Wald dengan hipotesis sebagai berikut:

H_0: $\beta_j = 0$ (tidak berpengaruh signifikan terhadap variabel dependen)

H_1: $\beta_j \neq 0$ (berpengaruh signifikan terhadap variabel dependen)

Statistik uji Wald:

$$W_j = \frac{\widehat{\beta_j}}{se\left(\widehat{\beta_j}\right)} \sim N(0,1) \tag{17}$$

dengan $\widehat{\beta_j}$ adalah penduga β_j dan $se\left(\widehat{\beta_j}\right)$ adalah standar error penduga β_j. Statistik W_j mengikuti distribusi *chi-square* (χ^2) dengan derajat kebebasan 1. Kriteria pengujian adalah tolak H_0 jika $|W_j| > \chi^2(\alpha, 1)$ atau $p - value < \alpha$.

1.4.6 Regresi Nonparametrik *Spline Truncated*

Regresi nonparametrik merupakan metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen dan independen tanpa perlu menetapkan bentuk fungsi tertentu sebelumnya. Berbeda dengan regresi parametrik yang mengharuskan adanya asumsi mengenai bentuk fungsi hubungan, regresi nonparametrik menawarkan fleksibilitas lebih tinggi karena kurva regresinya langsung ditentukan oleh data yang ada (Suryono dkk., 2024). Fleksibilitas ini sangat berguna, terutama dalam kasus dengan pola hubungan yang kompleks atau tidak diketahui, seperti data demografi yang berfluktuasi atau data ekonomi dengan tren yang sulit diprediksi (Octavanny dkk., 2021). Secara umum, model regresi nonparametrik dapat dinyatakan dengan persamaan (18) sebagai berikut:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,2,\dots,n \tag{18}$$

dengan

y_i : variabel dependen pada amatan ke-i,

$f(x_i)$: fungsi regresi yang diestimasi secara flkesibel,,

x_i : variabel independen pada amatan ke-i,

ε_i : *residual* pada amatan ke-i

Spline truncated adalah teknik dalam regresi nonparametrik yang digunakan untuk memodelkan hubungan antara variabel dependen dan independen, khususnya ketika hubungan tersebut tidak diketahui atau mengalami perubahan signifikan di beberapa sub-interval. Teknik ini memungkinkan penangkapan perubahan hubungan pada interval yang berbeda dengan menggunakan titik knot (Firpha & Achmad, 2022). Pada dasarnya, regresi *spline truncated* terdiri dari dua komponen utama: fungsi polinomial yang memodelkan hubungan pada interval kontinu, dan komponen *truncated* yang memungkinkan penyesuaian fungsi pada interval spesifik. Dengan adanya titik knot, *spline* ini menyesuaikan bentuk kurvanya agar lebih fleksibel dan

sesuai dengan pola data yang kompleks (Handayani dkk., 2023). Secara matematis, fungsi *spline truncated* dapat ditulis sebagai berikut:

$$f(x_i) = \beta_0 + \sum_{k=1}^{q} \beta_k x_i^k + \sum_{h=1}^{r} \beta_{q+h}(x_i - K_h)_+^q \, \varepsilon_i, \qquad i = 1,2,\dots,n \qquad (19)$$

dengan β_k adalah koefisien polinomial untuk komponen pertama, β_{q+h} adalah koefisien pada komponen *truncated*, K_h menunjukkan posisi titik knot, r adalah jumlah titik *knot* sedangkan nilai q menunjukkan derajat polinomial serta $(x_i - K_h)_+^q$ disebut sebagai fungsi *truncated,* yang didefinisikan sebagai:

$$(x_i - K_h)_+^q = \begin{cases} (x_i - K_h)^q, & x \geq K_h \\ 0 \quad , & x < K_h \end{cases}$$

Komponen fungsi *truncated* $(x_i - K_h)_+^q$ bertujuan untuk mengubah bentuk fungsi regresi pada titik-titik tertentu, yang dikenal sebagai titik knot (K_h), yang mencerminkan perubahan signifikan dalam data (Nurcahayani dkk., 2021). Fungsi *truncated* ini memberikan fleksibilitas yang diperlukan untuk menangkap perubahan lokal dalam data, yang biasanya tidak dapat diakomodasi oleh model regresi polinomial biasa (Pratiwi, 2020).

1.4.7 Pemilihan Titik Knot Optimal

Pemilihan model terbaik merupakan langkah krusial dalam regresi nonparametrik, khususnya dalam regresi *spline truncated*. Salah satu kriteria yang umum digunakan adalah *Generalized Cross Validation* (GCV), yang memberikan nilai optimal asimtotik tanpa bergantung pada varians populasi yang tidak diketahui serta tetap konsisten meskipun data mengalami transformasi (Utami dkk., 2020). Dalam regresi *spline truncated*, metode ini membantu menentukan titik knot optimal dengan memilih nilai GCV terkecil, yang menunjukkan model dengan kesalahan prediksi minimum. Persamaan GCV untuk menentukan titik knot optimal pada regresi *spline truncated* seperti pada Persamaan (20)

$$GCV(K) = \frac{MSE}{\left[1 - \frac{trace\left(\boldsymbol{A}(\boldsymbol{k})\right)}{n}\right]^2} \qquad (20)$$

dengan

n : jumlah pengamatan,
K : jumlah titik knot,
$\boldsymbol{I}$: matriks identitas,
$\boldsymbol{A}(\boldsymbol{k})$: matriks estimasi yang dihitung berdasarkan titik knot yang digunakan.

Mean Squared Error (MSE) diperoleh melalui $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ dengan y_i adalah nilai aktual $\hat{y}_i$ adalah nilai prediksi.

1.4.8 Ketepatan Klasifikasi Model

Ketepatan klasifikasi model dalam regresi logistik dievaluasi melalui *confusion matrix*, yang merangkum hasil prediksi model dengan membandingkannya dengan nilai aktual variabel dependen. Alat ini memberikan gambaran lebih komprehensif

mengenai performa model dalam mengklasifikasikan data, dengan menunjukkan jumlah prediksi yang benar dan salah untuk setiap kategori. *Confusion matrix* membagi hasil prediksi model menjadi empat kategori utama: *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Berdasarkan informasi ini, berbagai metrik performa model dapat dihitung, salah satunya adalah akurasi, yang menunjukkan seberapa baik model dalam melakukan klasifikasi secara keseluruhan (Ohsaki dkk., 2017).

Tabel 1. Confusion Matrix

Aktual	Prediksi	
	Positive	*Negative*
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

Akurasi dihitung berdasarkan *confusion matrix* pada Tabel 1, yang mengukur proporsi prediksi benar dibandingkan dengan total prediksi. Rumus perhitungannya dapat dilihat pada Persamaan (21).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

Persamaan (21) menunjukkan seberapa sering model membuat prediksi yang benar dengan membandingkan jumlah prediksi yang benar (TP dan TN) terhadap total prediksi (TP, TN, FP, dan FN). *Confusion matrix* memberikan gambaran yang jelas tentang kinerja model regresi logistik dalam klasifikasi dan membantu mengidentifikasi area model yang mungkin melakukan kesalahan.

1.4.9 Status Gizi Balita

Masa balita, yang sering disebut sebagai periode emas, adalah fase kritis dalam perkembangan seorang anak yang menentukan dasar bagi pertumbuhan fisik, kognitif, emosional, dan sosial yang optimal. Periode ini ditandai oleh percepatan perkembangan yang signifikan dan sensitivitas tinggi terhadap berbagai pengaruh lingkungan seperti nutrisi, stimulasi, dan pengasuhan. Nutrisi adalah salah satu faktor utama yang mendukung pertumbuhan dan perkembangan balita. Asupan gizi yang cukup dan seimbang mendukung pertumbuhan fisik serta perkembangan kognitif, motorik, dan daya tahan tubuh. Kekurangan nutrisi pada periode ini dapat menyebabkan berbagai dampak negatif, termasuk hambatan dalam pertumbuhan, defisit kognitif, dan peningkatan risiko penyakit (Yulianto dkk., 2022).

Balita membutuhkan zat gizi makro seperti karbohidrat, protein, dan lemak, serta zat gizi mikro seperti vitamin dan mineral, yang penting untuk perkembangan fisik dan kognitif. Kekurangan gizi yang adekuat dapat menyebabkan malnutrisi, stunting, dan kerentanan terhadap penyakit infeksi. Stunting, yang merupakan kondisi tinggi badan rendah akibat kekurangan gizi kronis, menjadi masalah besar di Indonesia karena dampaknya terhadap kemampuan belajar dan produktivitas anak di masa depan. Penelitian menunjukkan bahwa intervensi nutrisi yang tepat selama periode emas sangat penting untuk mencegah gangguan pertumbuhan dan meningkatkan kualitas hidup anak secara keseluruhan (Pratama dkk., 2023).

Ketika menilai status gizi balita, salah satu metode yang digunakan adalah antropometri. Ini meliputi pengukuran berat badan, tinggi badan, serta parameter lain yang membantu menilai proporsi dan komposisi tubuh anak (Isni & Dinni, 2020). Berdasarkan SK Menteri Kesehatan No. 2 Tahun 2022, terdapat dua metode utama dalam penilaian status gizi balita, yaitu berdasarkan Berat Badan menurut Panjang Badan atau Tinggi Badan (BB/PB atau BB/TB) dan Berat Badan menurut Umur (BB/U). Setiap metode ini memiliki kategori status gizi yang ditentukan berdasarkan ambang batas Z-Score, yaitu standar deviasi yang menunjukkan jarak antara ukuran yang diukur dengan nilai rata-rata yang diharapkan untuk usia tertentu.

Tabel 2 menunjukkan kategori status gizi berdasarkan BB/TB, yang mencakup kategori gizi buruk hingga obesitas, sesuai dengan Z-Score yang telah ditentukan. Ini adalah panduan penting untuk mengidentifikasi kondisi gizi balita, yang selanjutnya dapat digunakan sebagai dasar perencanaan intervensi nutrisi.

Tabel 2. Kategori dan ambang batas status gizi balita berdasarkan BB/TB

Indeks	Kategorik Status Gizi	Ambang Batas (Z-Score)
Berat Badan menurut Tinggi Badan (BB/TB) anak usia 0 – 60 bulan	Gizi Buruk	< –3 SD
	Gizi Kurang	–3 SD sd –2 SD
	Gizi Baik	–2 SD sd +1 SD
	Berisiko Gizi Lebih	> +1 SD sd +2 SD
	Gizi Lebih	> +2 SD sd +3 SD
	Obesitas	> +3 SD

Sumber: Kementerian Kesehatan (2020)

Selain itu, terdapat juga TB/U yang memfokuskan pada perbandingan tinggi badan anak dengan umur. Tabel 3 menggambarkan kategori status gizi balita berdasarkan ambang batas Z-Score untuk TB/U, mulai dari tinggi badan sangat pendek hingga tinggi.

Tabel 3. Kategori dan ambang batas status gizi balita berdasarkan TB/U

Indeks	Kategorik Status Gizi	Ambang Batas (Z-Score)
Tinggi Badan menurut Umur (TB/U) anak usia 0 – 60 bulan	Sangat Pendek	< –3 SD
	Pendek	–3 SD sd –2 SD
	Normal	–2 SD sd +3 SD
	Tinggi	> +3 SD

Sumber: Kementerian Kesehatan (2020)

Penilaian status gizi menggunakan kategori-kategori di atas memberikan gambaran yang jelas mengenai kondisi kesehatan anak, yang sangat penting untuk merencanakan intervensi yang tepat guna mendukung tumbuh kembang balita.

BAB II
METODE PENELITIAN

2.1 Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder, yaitu data Status Gizi Balita di Kabupaten Gowa pada tahun 2023. Data ini diperoleh dari Dinas Kesehatan Kabupaten Gowa. Adapun data lengkapnya terdiri dari 2092 amatan yang dapat dilihat Pada Lampiran 1.

2.2 Variabel Penelitian

Terdapat satu variabel dependen (y) dan lima variabel independen (x) dalam penelitian ini. Variabel respon yang dianalisis adalah status gizi balita, yang dikategorikan menjadi dua kelompok: gizi kurang (gizi buruk, gizi kurang) dan gizi baik Adapun variabel independe meliputi faktor-faktor yang memengaruhi status gizi balita, seperti yang tercantum pada Tabel 2.

Tabel 4. Variabel Penelitian

Variabel	Keterangan	Kategori
y	Status Gizi Balita	0 = Gizi Baik & 1 = Gizi Buruk
x_1	Berat Badan Lahir	Rasio
x_2	Tinggi Badan Lahir	Rasio
x_3	Usia	Rasio
x_4	Berat Badan	Rasio
x_5	Tinggi Badan	Rasio

Sumber: Dinas Kesehatan Kabupaten Gowa (2023)

2.3 Metode Analisis Data

Langkah-langkah analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Mengestimasi parameter model regresi logistik biner *ridge* dengan estimator *spline truncated* menggunakan MLE.
2. Memodelkan keluarga Status Gizi Balita di Kabupaten Gowa pada tahun 2023 yang mengandung multikolinieritas dengan menggunakan regresi logistik biner *ridge* dengan estimator *spline truncated.* Pemodelan ini dilakukan dengan menggunakan bantuan *software* R-Studio. Adapun analisisnya sebagai berikut:
   a. Melakukan analisis statistik deskriptif untuk variabel dependen dan setiap variabel independen yang berhubungan dengan status gizi balita.
   b. Melakukan uji multikolinieritas dengan membangun matriks korelasi antar variabel independen untuk mengetahui apakah ada hubungan linier yang tinggi antar variabel menggunakan Persamaan (6).
   c. Melakukan pemodelan status gizi balita menggunakan regresi logistik *ridge* dengan estimator *spline truncated* untuk orde linier (orde 1) dan kuadratik (orde 2) pada satu dan dua titik knot.

- Menentukan titik-titik knot untuk masing-masing model, yaitu model linier dan kuadratik pada satu dan dua titik knot berdasarkan nilai GCV terkecil.
- Mencari nilai lambda optimal pada setiap model dengan menguji nilai antara 0 hingga 1 dengan interval 0.1 dan memilih yang meminimalkan nilai GCV.
- Melakukan estimasi parameter pada setiap model menggunakan MLE dengan penambahan penalti *ridge*.
- Melakukan proses iterasi *Newton-Raphson* hingga konvergensi tercapai, yaitu ketika $\left|\beta^{(t+1)} - \beta^{(t)}\right| \leq 10^{-6}$

d. Pemilihan model terbaik yang dipilih berdasarkan nilai GCV terendah, yang dihitung menggunakan Persamaan (20).
e. Melakukan uji signifikansi parameter secara simultan pada model terbaik menggunakan Persamaan (16) dan uji parameter secara parsial menggunakan Persamaan (17).
f. Menginterpretasi hasil estimasi koefisien model terbaik dan mengukur hasil klasifikasi model terbaik menggunakan Persamaan (21).
g. Menarik kesimpulan berdasarkan hasil estimasi dan analisis serta evektivitas model yang dibangun.