

Analisis Regresi dengan Metode LASSO Menggunakan Algoritma LARS untuk Mengatasi Multikolinearitas

Indah Alvionita *, Reny Rian Marlina

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

indahalvionita7@gmail.com, renyrianmarlina@unisba.ac.id

Abstract. Multicollinearity is a common problem in multiple linear regression, leading to unstable coefficient estimates, high variance, and reduced prediction accuracy. This study aims to develop a stable and efficient regression model using the Least Absolute Shrinkage and Selection Operator (LASSO) method with the Least Angle Regression (LARS) algorithm. The study focuses on socioeconomic factors affecting household expenditure in Indramayu Regency. The data were obtained from the 2024 National Socioeconomic Survey (SUSENAS) via the SILASTIK database. The dependent variable is Household Expenditure, with independent variables including Household Size (ART), daily per capita consumption of Calories, Carbohydrates, Fat, Protein, and Weekly Expenditure from Purchases. Analysis shows high multicollinearity among consumption variables in the OLS model. Using LASSO, only three significant variables remain: Household Size, Fat Consumption, and Protein Consumption. The model yields a pseudo R^2 of 0.3382682, nearly equal to the OLS R^2 of 0.3384315. The resulting regression equation is:

$$\hat{Y} = -2015769 + 937669.5X_1 + 46741.4X_4 + 13706X_5$$

LASSO successfully simplifies the model structure without compromising predictive power, making it an effective method for socioeconomic analysis involving highly correlated variables.

Keywords: *LASSO, LARS, multicollinearity.*

Abstrak. Multikolinearitas merupakan masalah umum dalam regresi linier berganda yang menyebabkan estimasi koefisien tidak stabil, varians tinggi, dan akurasi prediksi rendah. Penelitian ini bertujuan membangun model regresi yang stabil dan efisien menggunakan metode *Least Absolute Shrinkage and Selection Operator* (LASSO) dengan algoritma *Least Angle Regression* (LARS). Fokus penelitian adalah faktor sosial ekonomi yang memengaruhi pengeluaran rumah tangga di Kabupaten Indramayu. Data diperoleh dari Survei Sosial Ekonomi Nasional (SUSENAS) 2024 melalui basis data SILASTIK. Variabel dependen adalah Pengeluaran Rumah Tangga, sedangkan variabel independen meliputi: Jumlah ART, Konsumsi Kalori, Karbohidrat, Lemak, Protein per kapita per hari, serta Pengeluaran Seminggu dari Pembelian. Hasil analisis menunjukkan adanya multikolinearitas tinggi antar variabel konsumsi pada model MKT. Dengan LASSO, model menyisakan tiga variabel signifikan: Jumlah ART, Konsumsi Lemak, dan Konsumsi Protein, dengan nilai *pseudo* R^2 sebesar 0.3382682, hampir setara dengan R^2 model MKT sebesar 0.3384315. Model regresi LASSO yang diperoleh adalah:

$$\hat{Y} = -2015769 + 937669.5X_1 + 46741.4X_4 + 13706X_5$$

LASSO terbukti menyederhanakan model tanpa kehilangan kekuatan prediksi, sehingga efektif untuk analisis sosial ekonomi dengan banyak variabel berkorelasi.

Kata Kunci: *LASSO, LARS, Multikolinearitas.*

A. Pendahuluan

Multikolinearitas merupakan salah satu masalah umum dalam regresi linier berganda yang berdampak signifikan terhadap kestabilan estimasi koefisien dan keakuratan prediksi model. Ketika dua atau lebih variabel independen saling berkorelasi tinggi, interpretasi pengaruh masing-masing variabel terhadap variabel dependen menjadi tidak jelas, dan model regresi klasik cenderung memberikan estimasi yang tidak stabil (Gujarati & Porter, 2009). Permasalahan ini sering dijumpai pada data sosial ekonomi, terutama ketika melibatkan banyak variabel yang saling berkaitan erat.

Salah satu pendekatan modern yang efektif untuk mengatasi multikolinearitas adalah metode Least Absolute Shrinkage and Selection Operator (LASSO), yang diperkenalkan oleh Tibshirani (1996). Metode ini memungkinkan seleksi variabel secara otomatis, menghasilkan model yang lebih sederhana tanpa kehilangan kekuatan prediksi. Untuk memaksimalkan efisiensi perhitungannya, LASSO sering dikombinasikan dengan algoritma Least Angle Regression (LARS) yang dikembangkan oleh Efron dkk. (2004). Kombinasi keduanya memberikan solusi optimal dalam pemodelan regresi yang melibatkan banyak variabel berkorelasi tinggi.

Berdasarkan data Survei Sosial Ekonomi Nasional (SUSENAS) tahun 2024, Kabupaten Indramayu memiliki tingkat kemiskinan tertinggi di Jawa Barat, dengan rata-rata pengeluaran rumah tangga yang tergolong rendah. Informasi ini mencerminkan kondisi sosial ekonomi masyarakat dan penting untuk dianalisis lebih lanjut. Namun, kompleksitas antar variabel seperti jumlah anggota rumah tangga, konsumsi kalori, konsumsi gizi, dan pengeluaran mingguan sering menimbulkan gejala multikolinearitas yang mengganggu analisis regresi konvensional (BPS, 2024; Yuliana & Haryono, 2020).

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menerapkan metode regresi LASSO menggunakan algoritma LARS dalam menganalisis pengeluaran rumah tangga di Kabupaten Indramayu. Tujuan utamanya adalah menghasilkan model yang stabil, efisien, dan akurat dengan mengatasi permasalahan multikolinearitas pada variabel-variabel sosial ekonomi yang memengaruhi pengeluaran rumah tangga. Oleh karena itu, penelitian ini diberi judul “Analisis Regresi dengan Metode Least Absolute Shrinkage and Selection Operator (LASSO) Menggunakan Algoritma LARS untuk Mengatasi Multikolinearitas.”

B. Metode

Bahan

Penelitian ini menggunakan metode *Least Absolute Shrinkage and Selection Operator* (LASSO) dengan algoritma *Least Angle Regression* (LARS). Penelitian ini menggunakan data sekunder tahun 2024 yang diperoleh dari situs resmi Sistem Informasi Layanan Statistik (SILASTIK) yang dikelola oleh Badan Pusat Statistik (BPS). Data yang digunakan merupakan data wilayah Kabupaten Indramayu, Provinsi Jawa Barat. Dalam penelitian ini, terdapat 1 variabel dependen (Y) dan 6 variabel independen (X) yang digunakan untuk menganalisis faktor-faktor yang memengaruhi variabel utama.

Tabel 1. Variabel Penelitian

Variabel	Nama Variabel	Keterangan
Y	Pengeluaran Sebulan	Rata rata Pengeluaran Rumah Tangga Sebulan
X ₁	Jumlah ART	Jumlah Anggota Rumah Tangga
X ₂	Kalori	Banyaknya Konsumsi Kalori Perkapita Sehari
X ₃	Karbohidrat	Banyaknya Konsumsi Karbohidrat Perkapita Sehari
X ₄	Lemak	Banyaknya Konsumsi Lemak Perkapita Sehari
X ₅	Protein	Banyaknya Konsumsi Protein Perkapita Sehari
X ₆	Pengeluaran Seminggu	Nilai Pengeluaran Seminggu Satu Rumah Tangga yang Berasal dari Pembelian

Standarisasi Data

Standarisasi *Z-score* menyamakan skala variabel berbeda satuan dengan mengukur jarak data dari rata-rata dalam satuan deviasi standar. Nilai positif/negatif menunjukkan posisi relatif terhadap rata-rata (Hastie et al., 2017).

$$Z_j = \frac{X_j - \bar{X}}{S} \quad \dots(1)$$

Analisis Regresi Linear Berganda

Analisis regresi adalah metode statistik untuk mengukur hubungan antara variabel dependen dan independen, baik satu (regresi linier sederhana) maupun lebih (regresi linier berganda atau regresi klasik), dengan estimasi parameter umumnya menggunakan metode kuadrat terkecil atau kemungkinan maksimum (Kutner, Nachtsheim, & Neter, 2004; Gujarati, 2003).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad \dots(2)$$

Uji Multikolinearitas

Multikolinearitas dalam regresi linier berganda perlu diuji karena dapat membuat estimasi parameter tidak signifikan dan model menjadi tidak stabil (Sujarweni, 2015). Kondisi ini terjadi saat variabel independen saling berkorelasi tinggi, yang dapat dideteksi dengan *Variance Inflation Factor* (VIF); nilai VIF > 10 menunjukkan multikolinearitas tinggi dan menurunkan presisi model (Gujarati & Porter, 2009). VIF dihitung dengan rumus:

$$VIF_j = \frac{1}{1 - R_j^2} \quad \dots(3)$$

Pengujian Hipotesis

Menurut Gujarati (2004), pengujian hipotesis adalah prosedur statistik untuk menilai kebenaran pernyataan tentang parameter populasi berdasarkan data sampel. Dalam regresi, uji hipotesis digunakan untuk mengevaluasi signifikansi pengaruh variabel independen terhadap dependen, baik secara simultan (uji F) maupun parsial.

Pengujian simultan dilakukan menggunakan uji F, yang menguji apakah semua variabel independen secara bersama-sama berpengaruh signifikan terhadap variabel dependen. Rumus uji F adalah:

$$F = \frac{(R^2/k)}{((1 - R^2)/(n - k - 1))} \quad \dots(4)$$

Jika nilai F-hitung > F-tabel atau *p-value* < α , maka H_0 ditolak (Gujarati, 2004).

Sementara itu, uji t digunakan untuk menguji pengaruh setiap variabel independen secara individual. Rumus uji t sebagai berikut:

$$t = \frac{b_i}{SE(b_i)} \quad \dots(5)$$

Keterangan:

b_i = Koefisien regresi variabel independen ke-i

$SE(b_i)$ = Standard error dari koefisien regresi

H_0 ditolak jika t-hitung > t-tabel atau *p-value* < α , artinya variabel tersebut berpengaruh signifikan terhadap variabel dependen (Gujarati, 2004). Kedua uji ini penting dalam menentukan keandalan model regresi yang dibangun.

Koefisien Determinasi (R^2)

Koefisien determinasi (R^2) mengukur proporsi variasi variabel dependen yang dijelaskan oleh variabel independen, dengan nilai antara 0 dan 1; semakin mendekati 1, model dianggap semakin baik (Ghazali, 2011). Rumus koefisien determinasi adalah:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \text{ dimana } 0 \leq R^2 \leq 1 \quad \dots(6)$$

R^2 dihitung dari rasio jumlah kuadrat regresi (JKR) terhadap jumlah kuadrat total (JKT) (Sembiring, 1995), namun nilainya cenderung meningkat seiring penambahan variabel, meskipun tidak signifikan. Untuk mengatasinya, digunakan *adjusted* R^2 yang mempertimbangkan jumlah observasi dan variabel, sehingga lebih akurat dalam menilai kualitas model (Gujarati, 2003). Rumusnya adalah:

$$R_a^2 = 1 - \frac{k - 1}{n - 1} \left(\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right) \quad \dots(7)$$

dengan n adalah jumlah observasi dan k jumlah variabel. Oleh karena itu, penelitian ini menggunakan *adjusted* R^2 untuk mengevaluasi kebaikan model regresi.

Least Absolute Shrinkage and Selection Operator (LASSO)

Metode Least Absolute Shrinkage and Selection Operator (LASSO) oleh Tibshirani (1996) merupakan teknik regresi untuk mengatasi multikolinearitas dengan menyusutkan sebagian koefisien ke nol, sehingga mengurangi variansi dan melakukan seleksi variabel signifikan. Secara umum, model regresi LASSO dituliskan sebagai:

$$Y = X\hat{\beta}^{lasso} + \varepsilon \quad \dots(8)$$

Keterangan:

Y = vektor yang berisi nilai variabel dependen dengan ukuran ($n \times 1$)

X = matriks yang berisi variabel independen dengan ukuran ($n \times p$)

$\hat{\beta}^{lasso}$ = vektor yang berisi koefisien LASSO dengan ukuran $(k+1) \times 1$

ε = vektor yang berisi nilai galat dengan ukuran ($n \times 1$)

Estimasi koefisien LASSO diperoleh dengan menyelesaikan fungsi kuadrat yang dibatasi oleh kendala L1 (jumlah absolut koefisien), yaitu:

$$\hat{\beta}_{lasso} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad \dots(9)$$

(Efron et al., 2004). Di sini, λ adalah tuning parameter yang mengatur tingkat penyusutan. Semakin besar nilai λ , semakin banyak koefisien yang disusutkan ke nol. Tuning parameter ini dikaitkan dengan skala s menggunakan rumus:

$$s = \frac{t}{\sum_{j=1}^p |\hat{\beta}_j^0|} \quad \dots(10)$$

dengan, $t = \sum_{j=1}^p |\hat{\beta}_j^0|$, dimana $\hat{\beta}_j^0$ adalah koefisien MKT dari model awal tanpa regularisasi, dan nilai s yang optimal diperoleh melalui metode validasi silang (Dewi, 2010).

Least Angle Regression Shrinkage (LARS)

Least Angle Regression (LARS) adalah algoritma efisien untuk estimasi parameter regresi LASSO yang dikembangkan oleh Efron et al. (2004), dengan memodifikasi metode *forward selection* secara iteratif untuk memilih variabel berdasarkan korelasi tertinggi dengan residual (Hastie et al., 2008; Pusporini, 2012). Berikut langkah- langkah estimasi koefisien LASSO dengan LARS (Andana et al., 2017):

1. Mengidentifikasi komponen gradien antara variabel independen dan residual dari masing-masing variabel independen, dengan rumus sebagai berikut:

$$\hat{c}_j = X^t (y - \hat{\mu}) \quad \dots(11)$$

2. Mengidentifikasi komponen gradien terbesar dalam nilai absolut, menggunakan langkah-langkah berikut:

$$\hat{C} = \max\{|\hat{c}_j|\} \quad \dots(12)$$

maka didapatkan $s_j = \text{sign}\{\hat{c}_j\}$ untuk $j \in A$

3. Menentukan X_A , di mana himpunan A merupakan himpunan indeks aktif dari variabel independen $\{1, 2, 3, \dots, m\}$. Himpunan indeks aktif A ditentukan berdasarkan nilai korelasi mutlak terbesar. Matriks didefinisikan sebagai:

$$X_A = (\dots, s_j X_j, \dots)_{j \in A} \quad \dots(13)$$

dengan s_j bernilai ± 1 , sehingga

$$G_A = X_A^t X_A \text{ dan } A_A = (1_A^t G_A^{-1} 1_A)^{-\frac{1}{2}} \quad \dots(14)$$

4. Menghitung nilai vektor *equiangular*, yaitu vektor yang membagi sudut antara kolom-kolom menjadi sama besar, dengan ketentuan besar sudutnya kurang dari 90° . Nilai vektor *equiangular* dihitung dengan menggunakan rumus berikut:

$$u_A = X_A w_A \text{ dengan } w_A = A_A^{-1} 1_A \quad \dots(15)$$

5. Menghitung vektor *inner product*:

$$a = X^t u_A \quad \dots(16)$$

6. Menghitung $\hat{\mu}_A$ dengan $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{\gamma} u_A$

$$\hat{\gamma} = \min_{j \in A^c}^+ \left\{ \frac{\hat{C} - \hat{c}_j}{A_A - a_j}, \frac{\hat{C} + \hat{c}_j}{A_A + a_j} \right\} \quad \dots(17)$$

$\min_{j \in A^c}^+$ menunjukkan bahwa yang dipilih adalah nilai positif terkecil dari j yang tidak termasuk dalam himpunan A . Pada tahap akhir, nilai tersebut dihitung dengan menggunakan rumus:

$$\hat{\gamma}_m = \frac{\hat{C}_m}{A_m} \quad \dots(18)$$

7. Memperbaharui nilai $\hat{\mu}_A$, dengan menggunakan:

$$\hat{\gamma}_m = \frac{\hat{C}_m}{A_m} \quad \dots(19)$$

8. Algoritma LARS untuk LASSO memerlukan rumus:

$$\text{sign}(\hat{\beta}_j) = \text{sign}(\hat{c}_j) = s_j \quad \dots(20)$$

9. Langkah ke-8 dapat dicapai jika:

$$\tilde{\gamma} = \min_{\gamma_j > 0} \{\gamma_j\} \quad \dots(21)$$

$$\text{dengan } \gamma_j = \frac{-\hat{\beta}_j}{s_j w_{Aj}}$$

Jika $\tilde{\gamma}$ lebih kecil dari $\hat{\gamma}$ maka $\hat{\beta}_j$ (γ) tidak dapat dianggap sebagai solusi untuk LASSO karena batasan tanda tidak terpenuhi. Oleh karena itu, algoritma LARS untuk LASSO harus mengikuti ketentuan berikut: jika $\tilde{\gamma} < \hat{\gamma}$ hentikan proses LARS pada tahap $\tilde{\gamma} = \hat{\gamma}$ dan keluarkan variabel j dari perhitungan pada langkah selanjutnya. Variabel j akan dimasukkan kembali dalam perhitungan LARS pada tahap berikutnya.

10. Periksa apakah $\text{sign}(\hat{\beta}_j) = \text{sign}(\hat{c}_j) = s_j$, pada seleksi variabel pertama, jika sudah diperoleh $\text{sign}(\hat{\beta}_j) = \text{sign}(\hat{c}_j) = s_j = 1$, apabila tanda-tandanya sudah sesuai, maka proses seleksi variabel dapat dilanjutkan ke tahap berikutnya.

11. Melakukan langkah-langkah yang sama pada setiap proses seleksi variabel hingga semua variabel independen telah terseleksi.

Validasi Silang

Validasi silang (*cross-validation*) adalah teknik untuk mengevaluasi kinerja model dan mencegah overfitting dengan membagi data menjadi pelatihan dan uji. Dalam *k-fold cross-validation*, data dibagi menjadi k subset dan proses diulang hingga semua subset menjadi data uji (James et al., 2013).

$$CV = \frac{1}{k} \sum_{i=1}^n MSE_i \quad \dots(22)$$

Validasi silang efektif karena seluruh data digunakan bergiliran untuk pelatihan dan pengujian. Dalam regresi LASSO, nilai s terbaik dipilih dengan meminimalkan nilai CV menggunakan mode *fraction* (Nagarajan et al., 2014).

C. Hasil Penelitian dan Pembahasan

Standarisasi Data

Seluruh variabel distandarisasi dengan *Z-score* agar skala antarvariabel setara. Ini membuat model regresi lebih stabil dan koefisien lebih mudah diinterpretasikan.

Metode Kuadrat Terkecil (MKT)

Model regresi linier digunakan untuk menganalisis hubungan variabel independen terhadap Pengeluaran Sebulan yang telah distandarisasi, dengan persamaan:

$$\hat{Y} = -2.010 \times 10^6 + 943500X_1 - 119.2X_2 - 195.1X_3 + 47890X_4 + 15650X_5 + 0.6353X_6$$

Secara parsial, variabel Jumlah ART, Konsumsi Lemak, dan Konsumsi Protein berpengaruh signifikan, sedangkan Konsumsi Kalori, Karbohidrat, dan Pengeluaran Seminggu tidak signifikan, kemungkinan akibat multikolinearitas atau kontribusi yang rendah terhadap total pengeluaran.

Uji Multikolinearitas

Uji multikolinearitas mendeteksi hubungan linear antar variabel independen yang memengaruhi stabilitas koefisien regresi. Langkah awal dilakukan dengan melihat matriks korelasi (Tabel 2).

Tabel 2. Korelasi Antar Variabel Independen

Variabel	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆
X ₁	1					
X ₂	-0.448	1				
X ₃	-0.448	0.919	1			
X ₄	-0.309	0.836	0.637	1		
X ₅	-0.401	0.900	0.748	0.852	1	
X ₆	0.029	0.036	0.047	0.004	0.019	1

Hasil penghitungan VIF untuk setiap variabel independen ditunjukkan pada tabel berikut:

Tabel 2. Nilai VIF Setiap Variabel Independen

Variabel	VIF
Jumlah ART	1.284420
Konsumsi Kalori	28.190557
Konsumsi Karbohidrat	11.030746
Konsumsi Lemak	5.495681
Konsumsi Protein	6.937358
Pengeluaran Seminggu	1.006864

Berdasarkan Tabel 2, terdapat korelasi tinggi antara Konsumsi Kalori (X_2), Karbohidrat (X_3), Lemak (X_4), dan Protein (X_5), yang mengindikasikan potensi multikolinearitas. Hasil VIF pada Tabel 3 menunjukkan nilai 28,19 (X_2) dan 11,03 (X_3), melebihi batas umum >10 (Gujarati & Porter, 2009), menandakan multikolinearitas tinggi. Sebaliknya, Jumlah ART (X_1) dan Pengeluaran Seminggu (X_6) memiliki $VIF < 2$, artinya tidak bermasalah.

Karena itu, model regresi klasik dinilai kurang stabil dan metode alternatif seperti LASSO direkomendasikan.

Least Absolute Shrinkage and Selection Operator (LASSO) dengan Algoritma LARS

1. Identifikasi Komponen Gradien antara Variabel Independen dan Residual

$$\hat{c}_j = X^t(y - \hat{\mu}) = \begin{bmatrix} 285.374 \\ 276.120 \\ 163.774 \\ 412.770 \\ 328.177 \\ 22.755 \end{bmatrix}$$

Nilai $\hat{c}_j$ menunjukkan seberapa besar kontribusi awal variabel terhadap residual. Variabel dengan nilai $\hat{c}_j$ terbesar cenderung masuk lebih dulu ke model LASSO. Konsumsi Lemak (412.770) memiliki kontribusi tertinggi, sementara Pengeluaran Seminggu (22.755) terendah.

2. Hasil Identifikasi Variabel dengan Komponen Gradien Terbesar

$$\hat{C} = \max\{|\hat{c}_j|\} = 412.770$$

Konsumsi Lemak memiliki komponen gradien terbesar (412.770), menunjukkan hubungan paling kuat dengan residual awal. Variabel ini dipilih pertama dalam model LASSO karena kontribusinya paling dominan terhadap arah prediksi.

3. Penentuan Himpunan Aktif A dan Matriks Data Independen X_A

Setelah diketahui Konsumsi Lemak memiliki gradien terbesar, variabel ini masuk sebagai satu-satunya anggota awal himpunan aktif A dalam iterasi pertama LARS. Artinya, hanya variabel ini yang mulai membentuk model dengan koefisien berbeda dari nol.

$$X_A = \begin{bmatrix} -0.3440059 \\ -0.5362370 \\ 0.5638666 \\ -0.1736989 \\ \vdots \\ 0.3709765 \end{bmatrix}$$

4. Perhitungan Matriks G_A dan Bobot A_A

Setelah Konsumsi Lemak masuk sebagai variabel aktif pertama, dihitung matriks (skalar) $G_A = X_A^t X_A = 1039$ dengan $G_A^{-1} = 0,000962$ Hasil ini digunakan untuk menghitung bobot arah kontribusi $A_A = (0,000962)^{-1/2} = 32.2$, yang menunjukkan kuatnya pengaruh Konsumsi Lemak dalam menentukan arah awal jalur solusi regresi LASSO.

5. Hasil Vektor Equiangular u_A

Setelah diperoleh $A_A = 32.2$ dan $G_A^{-1} = 0,000962$, dihitung bobot arah $w_A = A_A G_A^{-1} = 0.0310236$. Vektor *equiangular* dihitung dengan $u_A = X_A w_A$, menghasilkan vektor sepanjang 1040 observasi.

$$u_A = \begin{bmatrix} -0.01067 \\ -0.01664 \\ 0.01749 \\ -0.00539 \\ \vdots \\ 0.01151 \end{bmatrix}$$

6. Menghitung vektor *inner product*

$$a = X^t u_A = \begin{bmatrix} -9.9520 \\ 26.9408 \\ 20.5426 \\ 32.2335 \\ 27.4595 \\ 0.1410 \end{bmatrix}$$

Variabel Konsumsi Lemak memiliki *inner product* tertinggi (32.2335), diikuti Konsumsi Protein (27.4595) dan Kalori (26.9408), menunjukkan arah sejalan dengan vektor prediksi. Ketiganya berpotensi menjadi variabel aktif berikutnya. Sebaliknya, Pengeluaran Seminggu (0.1410) dan Jumlah ART (-9.9520) menunjukkan arah tidak sejalan, sehingga kecil kemungkinan terpilih. Nilai-nilai ini menjadi dasar penentuan variabel pada iterasi selanjutnya.

7. Menghitung nilai gamma ($\hat{\gamma}$) dan *update* $\hat{\mu}_A$

Setelah diperoleh arah pergerakan u_A , langkah optimal model ditentukan dengan

$$\hat{\gamma} = \min_{j \in A^c}^+ \left\{ \frac{\hat{C} - \hat{c}_j}{A_A - a_j}, \frac{\hat{C} + \hat{c}_j}{A_A + a_j} \right\} = 3.019889$$

Nilai ini digunakan untuk memperbarui prediksi awal:

$$\hat{\mu}_{A+} = \begin{bmatrix} -0.0322 \\ -0.0502 \\ 0.0528 \\ -0.0163 \\ \vdots \\ 0.0348 \end{bmatrix}$$

Pembaruan ini menunjukkan bahwa model mulai membentuk prediksi linier berdasarkan kontribusi Konsumsi Lemak sebagai variabel aktif pertama.

8. Evaluasi Tanda dan Lanjutkan Seleksi Variabel

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.0937 \\ 0 \\ 0 \end{bmatrix}$$

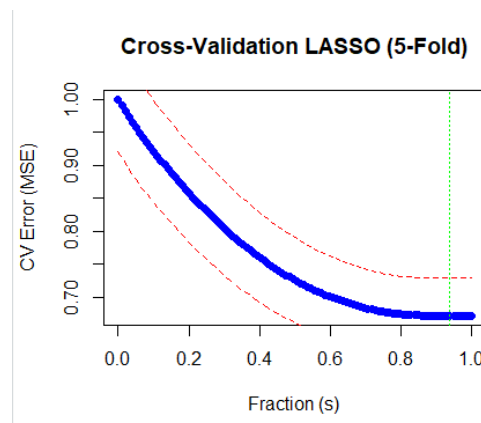
Setelah pembaruan $\hat{\mu}_A$, evaluasi tanda koefisien regresi sementara menunjukkan hanya Konsumsi Lemak yang aktif dengan nilai $\beta_4 = 0,0937$, sedangkan variabel lain masih nol. Evaluasi dilanjutkan dengan mencocokkan tanda β_j dan $\hat{c}_j$ jika sesuai, variabel dipertahankan. Dari tabel, tiga variabel (Konsumsi Lemak, Protein, dan Jumlah ART) memiliki tanda yang sesuai dan dimasukkan ke dalam model. Iterasi dihentikan saat ditemukan ketidaksesuaian tanda pada variabel lainnya.

Tabel 3. Hasil Iterasi Evaluasi Tanda

Iterasi	Variabel yang Diuji	Tanda Koefisien	Tanda Gradien	Hasil Evaluasi
1	Konsumsi Lemak	+	+	Sesuai, lanjut
2	Konsumsi Protein	+	+	Sesuai, lanjut
3	Jumlah ART	+	+	Sesuai, lanjut
4	Konsumsi Kalori, Pengeluaran Seminggu, Konsumsi Karbo	Tanda tidak sesuai	-/+	Tidak sesuai, dihentikan

Validasi Silang

- Menentukan nilai s optimal berdasarkan nilai validasi silang atau *cross-validation* terkecil. Nilai s optimal ditentukan melalui *5-fold Cross-Validation* untuk meminimalkan MSE. Dari hasil validasi, diperoleh $s = 0.939$ sebagai nilai dengan MSE terendah, menunjukkan keseimbangan antara bias dan varians serta menjaga variabel penting tetap aktif tanpa overfitting.



Gambar 1. Visualisasi Hasil Cross-Validation

Titik biru pada grafik menunjukkan rata-rata nilai MSE untuk setiap nilai s , garis merah putus-putus menggambarkan interval kesalahan ($\pm$ standar error), sedangkan garis hijau vertikal menunjukkan titik optimal pada $s=0,939$.

- Pilih model regresi LASSO terbaik berdasarkan nilai s optimal yang telah ditentukan. Model regresi LASSO terbaik diperoleh pada nilai $s=0,939s$, menghasilkan persamaan:

$$\hat{Y} = -2015769 + 937669.5X_1 + 46741.4X_4 + 13706X_5$$

Tiga variabel terpilih oleh LASSO: Jumlah ART, Konsumsi Lemak, dan Konsumsi Protein, sementara variabel lain dieliminasi (koefisien = 0). Pengeluaran meningkat Rp937.669,5 per anggota, Rp46.741,4 per satuan lemak, dan Rp13.706 per satuan protein, menunjukkan efektivitas LASSO dalam menyederhanakan model tanpa mengurangi prediksi.

Perbandingan Ukuran Kebaikan Model

1. Nilai Koefisien Determinasi Model

Tabel 4. Hasil Evaluasi Kebaikan Model

Model	<i>R-squared</i>	<i>Adjusted R-squared</i>
MKT	0.3384315	0.3345888
LASSO	0.3382682 (<i>pseudo</i>)	-

Evaluasi kebaikan model menunjukkan bahwa model MKT memiliki nilai *R-squared* sebesar 0,3384 dan *Adjusted R-squared* 0,3346, sedangkan model LASSO menghasilkan *pseudo R-squared* 0,3383. Ini menunjukkan bahwa LASSO mampu menjelaskan variasi data hampir setara dengan MKT, meskipun hanya menggunakan tiga variabel penting.

2. Perbandingan Koefisien Regresi MKT vs LASSO

Tabel 5. Perbandingan Koefisien Regresi MKT vs LASSO

Variabel	MKT	LASSO
Intersep	-2.010×10^6	-2015769
Jumlah ART	943500	937669.5
Konsumsi Kalori	-119.2	0
Konsumsi Karbo	-195.1	0
Konsumsi Lemak	47890	46741.4
Konsumsi Protein	15650	13706
Pengeluaran Seminggu	0.6353	0

Perbandingan hasil estimasi menunjukkan bahwa model MKT mempertahankan semua variabel, namun hanya Jumlah ART, Konsumsi Lemak, dan Konsumsi Protein yang signifikan. Model LASSO menyederhanakan model dengan mengeliminasi tiga variabel tidak signifikan (Kalori, Karbohidrat, dan Pengeluaran Seminggu) tanpa menurunkan kualitas prediksi secara berarti.

D. Kesimpulan

Hasil penelitian menunjukkan bahwa metode LASSO dengan algoritma LARS efektif mengatasi multikolinearitas dalam model regresi pengeluaran rumah tangga di Kabupaten Indramayu. Uji korelasi dan VIF mengindikasikan multikolinearitas tinggi antar variabel konsumsi ($VIF > 10$), sehingga model klasik (MKT) menjadi tidak stabil. Model LASSO menyederhanakan struktur regresi menjadi tiga variabel signifikan: Jumlah ART, Konsumsi Lemak, dan Konsumsi Protein, dengan nilai $\text{pseudo } R^2$ sebesar 0.3382682, hampir setara dengan R^2 model MKT sebesar 0.3384315. Persamaan regresi LASSO dalam skala asli adalah $\hat{Y} = -2015769 + 937669.5X_1 + 46741.4X_4 + 13706X_5$.

Ucapan Terimakasih

Peneliti mengucapkan terima kasih kepada Ibu Reny Rian Marlina, S.Si., M.Stat. selaku dosen pembimbing atas bimbingan, arahan, dan saran yang telah diberikan selama proses penyusunan penelitian ini. Ucapan terima kasih juga disampaikan kepada keluarga, sahabat, dan rekan-rekan mahasiswa yang telah memberikan dukungan moral dan semangat yang berarti.

Daftar Pustaka

- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2).
- Fadilah, P. M. (2018). Analisis regresi dengan metode Least Absolute Shrinkage and Selection Operator (LASSO) dalam menangani multikolinearitas.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Springer series in statistics: The elements of statistical learning: Data mining, inference, and prediction (2nd ed.).
- Mahalani, A. J., & Rifai, N. A. K. (2022, July). Least Absolute Shrinkage and Selection Operator (LASSO) untuk Mengatasi Multikolinearitas pada Model Regresi Linear Berganda. In *Bandung Conference Series: Statistics* (Vol. 2, No. 2, pp. 119–125).
- Mait, Y. A., Salaki, D. T., & Hanny, H. (2021). Judul artikel. *D'Cartesian*, 10(2), 69–75.
- Pardede, T. T. (2022). Penerapan regresi Least Absolute Shrinkage and Selection Operator (LASSO) untuk mengidentifikasi variabel yang berpengaruh terhadap kejadian stunting di Indonesia (Doctoral dissertation, UNIVERSITAS NEGERI JAKARTA).
- Prabowo, F. K. H., Wilandari, Y., & Rusgiyono, A. (2015). Pemodelan pertumbuhan ekonomi Jawa Tengah menggunakan pendekatan Least Absolute Shrinkage and Selection Operator (LASSO). *Jurnal Gaussian*, 4(4), 855–864.
- Rahmawati, F., & Suratman, R. Y. (2022). Performa Regresi Ridge dan Regresi Lasso pada Data dengan Multikolinearitas. *Leibniz: Jurnal Matematika*, 2(2), 1–10.
- Sartika, I., Debataraja, N. N., & Imro'ah, N. (2020). Analisis regresi dengan metode Least Absolute Shrinkage and Selection Operator (LASSO) dalam mengatasi multikolinearitas. *Bimaster: Buletin Ilmiah Matematika, Statistika Dan Terapannya*, 9(1).
- Sistem Informasi Layanan Statistik (SILASTIK). <https://silastik.bps.go.id/>
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1), 267–288.
- Anisa, L., & Rifai, N. A. K. (2022). Analisis Regresi Logistik Biner dengan Metode Penalized Maximum Likelihood pada Penyakit Covid-19 di RSUD Pringsewu. *Jurnal Riset Statistika*, 129–136. <https://doi.org/10.29313/jrs.v2i2.1425>
- Jody Alwin irawadi, & Sunendiari, S. (2021). Penerapan dan Perbandingan Tiga Metode Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara. *Jurnal Riset Statistika*, 1(1), 19–27. <https://doi.org/10.29313/jrs.v1i1.22>
- Dewi, E. F., & Hajarisman, N. (n.d.). *Penanganan Data Hilang pada Pemodelan Persamaan Terstruktur melalui Metode Full Information Maximum Likelihood (FIML)*. <https://journals.sangkurianpublisher.com/datamath>