Least Absolute Shrinkage And Selection Operator (LASSO) untuk Mengatasi Multikolinearitas pada Model Regresi Linear Berganda

Annisa Juwita Mahalani*, Nur Azizah Komara Rifai

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

*annisajuwita2018@gmail.com, azizah_kr@yahoo.com

Abstract. Linear regression analysis is one of them used to see the intersection between non-free variables and free variables. If there is more than one free variable, multiple linear regression analysis is used. Regression models are said to be good if the assumptions are met. One of them is the absence of multicollinearity, namely a high correlation or relationship between independent variables. In the event of multicholinearity, it can be resolved using the Least Absolute Shrinkage And Selection Operator (LASSO) regression method. The LASSO regression method was first introduced by Robert Tibshirani in 1996. This method can select variables and overcome multicholinearity. To make it easier to computate LASSO, you can use the Least Angle Regression (LARS) algorithm. In this study, we will discuss the application of the LASSO regression method to data on the number of poor people in Indonesia in 2020. By looking at the correlation coefficient and VIF values, there are 3 variables that indicate multicollinearity problems. The gross regional domestic product variable is the variable that is first entered in the model, meaning that the variable is most correlated with error.

Keywords: *LARS, LASSO, Multicollinearity, Poor Population.*

Abstrak. Analisis regresi linear salah satunya digunakan untuk melihat perngaruh antara variabel tak bebas dengan variabel bebas. Apabila variabel bebas lebih dari satu maka digunakan analisis regresi linear berganda. Model regresi dikatakan baik apabila asumsinya terpenuhi. Salah satunya adalah tidak adanya multikolinearitas yaitu korelasi atau hubungan yang tinggi diantara variabel bebas. Apabila terjadi multikolinearitas, dapat diatasi dengan menggunakan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO). Metode regresi LASSO diperkenalkan pertama kali oleh Robert Tibshirani pada tahun 1996. Metode ini dapat menyeleksi variabel dan mengatasi multikolinearitas. Untuk mempermudah komputasi LASSO, dapat menggunakan algoritma *Least Angle Regression* (LARS). Dalam penelitian ini, akan dibahas penerapan metode regresi LASSO pada data jumlah penduduk miskin di Indonesia tahun 2020. Dengan melihat nilai koefisien korelasi dan nilai VIF, terdapat 3 variabel yang terindikasi masalah multikolinearitas. Variabel produk domestik regional bruto adalah variabel yang pertama kali masuk pada model artinya variabel tersebut paling berkorelasi dengan galat.

Kata Kunci: *LARS, LASSO, Multikolinearitas, Penduduk Miskin.*

Corresponding Author
Email: azizah_kr@yahoo.com

A. Pendahuluan

Metode statistik yang digunakan untuk melihat pengaruh variabel bebas dan variabel tak bebas disebut dengan analisis regresi. Metode tersebut dibedakan menjadi regresi sederhana *(simple regression*) yaitu jika hanya terdapat satu variabel bebas dan regresi linier berganda (*multiple linear regression*) yaitu apabila variabel bebas lebih dari satu. Dalam penentuan model analisis regresi, diperlukan adanya estimasi parameter untuk menentukan nilai koefisen regresi. Metode Kuadrat Terkecil (MKT) atau *Ordinary Least Square* (OLS) dapat digunakan untuk menaksir koefisien regresi dengan meminimalkan jumlah kuadrat galat (*error*). Dalam metode kuadrat terkecil (MKT), uji asumsi klasik perlu dilakukan guna mendapatkan model terbaik. Akan tetapi, beberapa pengujian asumsi klasik tersebut sering kali tidak dapat terpenuhi salah satunya adalah adanya multikolinearitas. Multikolinearitas ditunjukkan dengan adanya korelasi yang cukup tinggi antara beberapa variabel bebas dengan melihat nilai VIF (*Variance Inflation Factor*) (Gujarati, 1978). Solusi yang dapat digunakan untuk mengatasi multikolinearitas adalah dengan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO).

LASSO memiliki model regresi yang sederhana dan lebih mudah untuk diinterpretasikan serta merupakan metode yang dapat menyusutkan koefisien regresi mendekati nol atau bahkan sama dengan nol (Tibshirani, 1996). LASSO juga berperan sebagai metode pemilihan variabel sekaligus mengatasi masalah multikolinearitas (Jolliffe, 2002). Dalam mempermudah komputasi LASSO, algoritma Least Angle Regression (LARS) dapat digunakan karena mempunyai modifikasi algoritma yang lebih efisien (Hastie et al, 2008). Penelitian untuk menangani multikolinearitas dengan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO) sudah dilakukan Kusuma (2019) untuk menganalisis kemiskinan dan kerentaan kemiskinan di Jawa Tengah. Selain itu, Wahyuningsih (2021) menerapakan analisis regresi LASSO untuk mengatasi multikolinearitas dan menyeleksi variabel yang memiliki pengaruh paling besar terhadap data persentase kemiskinan di Jawa Tengah.

Kemiskinan merupakan suatu masalah yang hampir dihadapi oleh seluruh negara berkembang, termasuk Indonesia. Masalah tersebut tidak bisa diabaikan karena akan berpengaruh terhadap produktivitas dan sumber daya penduduk. Berdasarkan laporan Survei Sosial Ekonomi Nasional (SUSENAS) pada tahun 2021, penduduk miskin di Indonesia pada bulan Maret tahun 2020 adalah sebanyak 24,7 juta jiwa dan 9,91 juta diantaranya termasuk dalam kategori miskin kronis (kemiskinan ekstrem). Dari total penduduk tersebut, kelompok miskin kronis berada pada kisaran 2,5% hingga 3%. Target Indonesia pada tahun 2024 dapat menurunkan kelompok miskin kronis menjadi 0%. Dengan demikian, kemiskinan merupakan salah satu persoalan mendasar bagi pemerintah yang harus segera dituntaskan.

Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah dalam penelitian ini adalah sebagai berikut: “Apakah terdapat masalah multikolinearitas pada faktor-faktor yang mempengaruhi jumlah penduduk miskin di Indonesia tahun 2020?” dan “Bagaimana penerapan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO) untuk mengatasi masalah multikolinearitas?”. Selanjutnya, adapun tujuan dalam penelitian ini diuraikan dalam pokok-pokok sbb.

1. Mengetahui adanya masalah multikolinearitas pada faktor-faktor yang mempengaruhi jumlah penduduk miskin di Indonesia tahun 2020.
2. Melakukan penerapan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO) untuk mengatasi masalah multikolinearitas.

B. Metodologi Penelitian

Bahan

Peneliti menggunakan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO). Penelitian ini menggunakan data sekunder yang bersumber dari *website* resmi Badan Pusat Statistik (bps.go.id) dan Satu Data Kementerian Ketenagakerjaan Republik Indonesia (satudata.kemnaker.go.id). Data yang digunakan adalah faktor-faktor yang berpengaruh terhadap penduduk miskin berdasarkan banyaknya provinsi di Indonesia pada tahun 2020. Variabel tak bebas (Y) yang digunakan yaitu jumlah penduduk miskin. Terdapat 9 variabel bebas yang diduga berpengaruh terhadap variabel tak bebas yang disajikan pada tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Simbol	Definisi
Garis Kemiskinan	X_1	Nilai rupiah pengeluaran minimum yang diperlukan untuk memenuhi kebutuhan pokok hidupnya selama sebulan
Produk Domestik Regional Bruto	X_2	Jumlah nilai tambah bruto yang timbul dari seluruh sektor perekonomian di daerah tersebut
Pengeluaran Per Kapita	X_3	Biaya yang dikeluarkan untuk konsumsi semua anggota rumah tangga selama sebulan
Tingkat Pengangguran Terbuka	X_4	Persentase jumlah pengangguran terhadap jumlah angkatan kerja
Indeks Pembangunan Manusia	X_5	Bagaimana penduduk dapat mengakses hasil pembangunan dalam memperoleh pendapatan, kesehatan, pendidikan, dan lainnya
Upah Minimum Provinsi	X_6	Upah minimum yang berlaku untuk seluruh kabupaten/kota di satu provinsi yang ditetapkan setiap satu tahun sekali
Rata-Rata Lama Sekolah	X_7	Rata-rata jumlah tahun yang digunakan oleh penduduk dalam menjalani pendidikan formal
Angka Melek Huruf	X_8	Proporsi penduduk usia 15 tahun keatas yang memiliki kemampuan membaca dan menulis kalimat sederhana
Angka Harapan Hidup	X_9	Perkiraan jumlah tahun hidup dari individu di suatu wilayah dari sekelompok makhluk hidup tertentu

Sumber: Badan Pusat Statistik, 2022

Metode Analisis

Peneltian ini akan menerapkan metode regresi *Least Absolute Shrinkage And Selection Operator* (LASSO) dengan algoritma *Least Angle Regression* (LARS) untuk mengatasi masalah multikolineartas pada faktor-faktor yang mempengaruhi jumlah penduduk miskin di Indonesia tahun 2020. Penelitian ini menggunakan *software* RStudio dan *Microsoft Excel*, berikut tahapan yang dilakukan:

1. Membakukan setiap variabel bebas hingga memiliki nilai tengah nol dan varians satu.
2. Memodelkan data dengan menggunakan metode kuadrat terkecil (MKT).
3. Mendeteksi adanya korelasi atau hubungan linear yang tinggi diantara variabel bebas (multikolinearitas) dengan melihat nilai koefisien korelasi dan VIF (*Variance Inflation Factor*). Jika nilai koefisien korelasi > 0,8 dan VIF > 10 maka terdeteksi masalah multikolinearitas.

4. Melakukan penaksiran koefisien LASSO menggunakan algoritma LARS hingga semua variabel bebas masuk ke dalam model.

C. Hasil Penelitian dan Pembahasan

Pembakuan Data

Pembakuan data dilakukan dengan mengubah ke bentuk z-score, yaitu koversi setiap variabel pengamatan ke nilai standar dengan mengurangi nilai dari rata-rata dan membaginya dengan nilai standar deviasi untuk setiap variabel untuk setiap variabel. Hal tersebut dilakukan karena variabel bebas memiliki satuan yang berbeda.

Analisis Regresi Linear Berganda

Dalam melakukan penaksiran koefisien regresi digunakan metode kuadrat terkecil (MKT) untuk mengestimasi koefisien regresi dan melihat pengaruh antar variabel bebas dengan variabel tak bebas. Berdasarkan faktor-faktor yang mempengaruhi jumlah penduduk miskin di Indonesia tahun 2020, model regresi linear berganda yang terbentuk adalah sebagai berikut:

$$Y = 6{,}669e^{-16} - 0{,}0651X_1 + 0{,}773X_2 - 0{,}266X_3 + 0{,}170X_4 + 0{,}018X_5 - 0{,}360X_6 - 0{,}284X_7 - 0{,}0287X_8 + 0{,}137X_9$$

Uji Multikolinearitas

Dalam mendeteksi multikolinearitas, dapat dilihat berdasarkan nilai koefisien korelasi dan *Variance Inflation Factor* (VIF). Jika nilai VIF > 10 dan koefisien korelasi > 0,8 maka terjadi multikolinearitas. Nilai koefisien korelasi untuk setiap variabel bebas disajikan pada tabel 2.

Tabel 2. Nilai Koefisien Korelasi antar Variabel Bebas

	GK	PDRB	PPK	TPT	IPM	UMP	RLS	AMH	AHH
GK	1								
PDRB	0,051	1							
PPK	0,265	0,505	1						
TPT	0,323	0,572	0,505	1					
IPM	0,223	0,412	0,866	0,485	1				
UMP	0,517	-0,021	0,210	0,231	0,050	1			
RLS	0,380	0,223	0,575	0,547	0,781	0,306	1		
AMH	0,224	0,059	0,324	0,361	0,521	0,092	0,644	1	
AHH	0,113	0,452	0,568	0,321	0,772	-0,126	0,415	0,296	1

Sumber: Data penelitian yang sudah diolah, 2022

Berdasarkan Tabel 1, terdapat lebih dari satu variabel bebas yang memiliki nilai korelasi lebih dari 0,8 yaitu pengeluaran per kapita dengan rata-rata lama sekolah, sehingga menjadi indikasi bahwa terdapat masalah multikolinearitas. Selain itu, untuk mendeteksi multikolinearitas dilakukan perhitungan nilai VIF. Nilai VIF dari masing-masing variabel bebas disajikan pada Tabel 3.

Tabel 3. Nilai VIF dari Setiap Variabel Bebas

Variabel Bebas	Nilai VIF
GK(X_1)	1,612174
PDRB (X_2)	2,372031
PPK (X_3)	17,695936

TPT (X_4)	2,299716
IPM (X_5)	56,616007
UMP (X_6)	2,43891
RLS (X_7)	12,769033
AMH (X_8)	1,838418
AHH (X_9)	8,019072

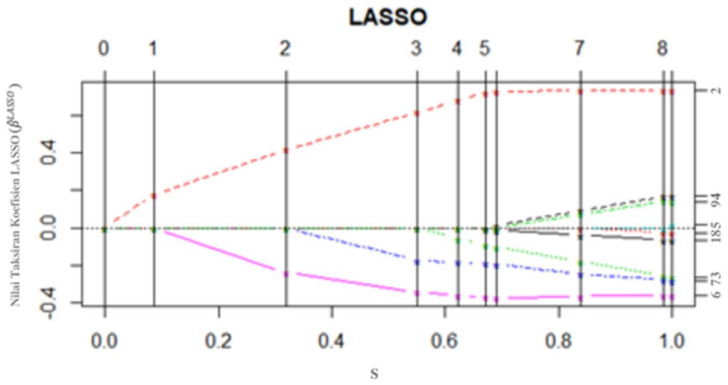
Sumber: Data penelitian yang sudah diolah, 2022

Berdasarkan Tabel 2. diketahui bahwa terdapat 3 variabel bebas yang memiliki nilai VIF > 10, yaitu: pengeluaran per kapita (X_3), indeks pembangunan manusia (X_5) dan rata-rata lama sekolah (X_7) sehingga dapat disimpulkan bahwa terjadi masalah multikolinearitas.

Hal tersebut disebabkan karena ketiga variabel saling berkorelasi. Secara intuitif pengeluaran per kapita akan meningkat, apabila penduduk memiliki taraf pendidikan yang tinggi artinya semakin lama penduduk menempuh pendidikan formal, semakin tinggi pula taraf pendidikan penduduk tersebut. Penduduk yang memiliki taraf pendidikan yang tinggi, akan lebih mudah memasuki lapangan kerja. Hal ini akan mengurangi pengangguran dan akan meningkatkan perekonomian. Apabila perekonomian sudah mulai membaik, maka pengeluaran penduduk untuk konsumsi rumah tangga akan meningkat guna memenuhi kebutuhan hidupnya. Dengan meningkatnya kemampuan untuk memenuhi biaya konsumsi rumah tangga artinya kelayakan hidup penduduk sudah lebih baik, maka akan meningkatkan nilai indeks pembangunan manusia. Karena peningkatan nilai tersebut, salah satunya dapat dilihat berdasarkan standar hidup layak penduduk. Salah satu komponen yang diukur dalam indeks pembangunan manusia adalah pendidikan. Jadi apabila taraf pendidikan meningkat, maka indeks pembangunan manusia akan meningkat karena akses penduduk untuk menggunakan fasilitas pembangunan untuk menempuh pendidikan sudah lebih lebih baik.

Penaksiran Koefisien LASSO dengan algoritma LARS

Dalam penelitian ini, penaksiran dilakukan secara bertahap dengan menetapkan koefisien berawal dari 0. Variabel bebas yang pertama masuk model adalah variabel yang memiliki korelasi tinggi dengan galat (*error*). Selanjutnya, semua variabel bebas secara bertahap masuk ke dalam model. Berikut adalah tahapan variabel bebas yang masuk pada model yang disajikan pada Gambar 1.



Gambar 1. Plot untuk Menduga Kofisien LASSO

Gambar 1. menggambarkan koefisien-koefisien regresi LASSO pada setiap tahapan algoritma LARS. Variabel produk domestik regional bruto adalah variabel bebas pertama yang masuk pada model, artinya variabel tersebut memiliki korelasi tertinggi dengan galat. Variabel upah minimum provinsi adalah variabel kedua yang masuk pada model. Proses berlanjut hingga

semua variabel bebas masuk pada model. Berikut adalah tahapan variabel bebas yang masuk pada model yang disajikan pada Tabel 4.

Tabel 4. Tahapan Variabel yang Masuk pada Model LASSO

Tahapan	Variabel Bebas yang Masuk pada Model
1	PDRB
2	PDRB UMP
3	PDRB UMP RLS
4	PDRB UMP RLS PPK
5	PDRB UMP RLS PPK GK
6	PDRB UMP RLS PPK GK TPT
7	PDRB UMP RLS PPK GK TPT AHH
8	PDRB UMP RLS PPK GK TPT AHH AMH
9	PDRB UMP RLS PPK GK TPT AHH AMH IPM

Sumber: Data penelitian yang sudah diolah, 2022

Berdasarkan Tabel 4, variabel secara bertahap masuk pada model dan terdapat 9 tahapan regresi LASSO dengan algoritma LARS. Selanjutnya, nilai taksiran koefisien LASSO untuk setiap tahapan disajikan pada Tabel 5.

Tabel 5. Nilai Taksiran Koefisien LASSO

Tahapan	GK	PDRB	PPK	TPT	IPM	UMP	RLS	AMH	AHH
1	0	0,177	0	0	0	0	0	0	0
2	0	0,417	0	0	0	-0,239	0	0	0
3	0	0,615	0	0	0	-0,344	-0,174	0	0
4	0	0,681	-0,057	0	0	-0,362	-0,184	0	0
5	-0,008	0,723	-0,094	0	0	-0,369	-0,189	0	0
6	-0,012	0,727	-0,103	0,012	0	-0,372	-0,195	0	0
7	-0,040	0,732	-0,181	0,091	0	-0,365	-0,246	0	0,072
8	-0,065	0,731	-0,255	0,169	0	-0,361	-0,276	-0,028	0,142
9	-0,065	0,732	-0,265	0,170	0,018	-0,359	-0,284	-0,028	0,136

Sumber: Data penelitian yang sudah diolah, 2022

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, peneliti menyimpulkan beberapa hasil penelitian sebagai berikut:

1. Terdapat 3 variabel bebas yang terindikasi masalah multikolinearitas karena memiliki nilai koefsien korelasi > 0,8 dan nilai *Variance Inflation Factor* (VIF) > 10, yaitu: variabel pengeluaran per kapita (X_3), variabel indeks pembangunan manusia (X_5) dan variabel rata-rata lama sekolah (X_7).
2. Variabel bebas pertama yang masuk pada model adalah produk domestik regional bruto (X_2) karena memiliki korelasi paling tinggi dengan galat. Selanjutnya, variabel upah minimum provinsi (X_6) dan variabel bebas terakhir yang masuk pada model adalah indeks pembangunan manusia (X_5).

Acknowledge

Alhamdulillahirabbil'alamin, peneliti mengucapkan terima kasih kepada Allah SWT dan seluruh pihak yang telah memberikan semangat, mendoakan, membantu dan memberi saran kepada peneliti selama penyusunan penelitian ini.

Daftar Pustaka

[1] Andana, A. P., Safitri, D., & Rusgiyono, A. (2017). MODEL REGRESI MENGGUNAKAN *LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR* (LASSO) PADA DATA BANYAKNYA GIZI BURUK KABUPATEN/KOTA DI JAWA TENGAH. *JURNAL GAUSSIAN*, 21-30. Retrieved November 12, 2021, from https://media.neliti.com/media/publications/100864-ID-model-regresi-menggunakan-least-absolute.pdf

[2] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2003). Least Angle Regression. mlpack, 1-44. Retrieved Desember 10, 2021, from http://mlpack.org/papers/lars.pdf

[3] Maulina, Silvia., Karyana, Yayat. (2021). Penerapan *Least Absolute Shrinkage And Selection Operator* Menggunakan Algoritma Least Angle Regression dalam Penanganan Multikolinearitas pada Data Inflasi. Retrieved Februari 13, 2021, from http://repository.unisba.ac.id/handle/123456789/28822

[4] Prabowo, F. K., Wilandari, Y., & Rusgiono, A. (2015). Pemodelan Pertumbuhan Ekonomi Jawa Tengah Menggunakan Pendekatan *Least Absolute Shrinkage And Selection Operator* (LASSO). Jurnal Gaussian, 855-864. Retrieved November 10, 2021, from http://ejournal-s1.undip.ac.id/index.php/gaussian

[5] Pusporini, A., Aunuddin, & Rahman, L. O. (2012). PENERAPAN REGRESI GULUD DAN *LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR* (LASSO) DALAM PENYUSUTAN KOEFISIEN REGRESI. Repository IPB, 23. Retrieved November 18, 2021, from https://repository.ipb.ac.id/handle/123456789/59538

[6] Robbani, M., Agustiani, F., & Herrhyanto, N. (2019). REGRESI *LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR* (LASSO) PADA KASUS INFLASI DI INDONESIA TAHUN 2014-2017. Jurnal EurekaMatika, 7. Retrieved from https://ejournal.upi.edu/index.php/JEM/article/view/22130

[7] Sartika, I., Debataraja, N. N., & Imro'ah, N. (2020). Analisis Regresi dengan Metode *Least Absolute Shrinkage And Selection Operator* (LASSO) Dalam Mengatasi Multikolinearitas. Bimaster, 9, 31-38. Retrieved November 10, 2021, from https://jurnal.untan.ac.id/index.php/jbmstr/article/view/38029

[8] Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society, 267-288. Retrieved Desember 27, 2021, from http://www.jstor.org/stable/2346178

[9] Hasna, Achmad, Anneke Iswani. (2022). *Metode Regresi Probit Biner untuk Pemodelan Faktor-Faktor yang Mempengaruhi Diagnosis Penyakit Jantung*, Jurnal Riset Statistika, 2(1), 28-34.