BAB III
METODOLOGI PENELITIAN

Pada bab ini akan membahas mengenai metodologi penelitian, metode yang digunakan yaitu regresi *Least Absolute Shrinkage and Selection Operator* (LASSO) yang diimplementasikan pada data Inflasi di Negara Indonesia tahun 2014-2017.

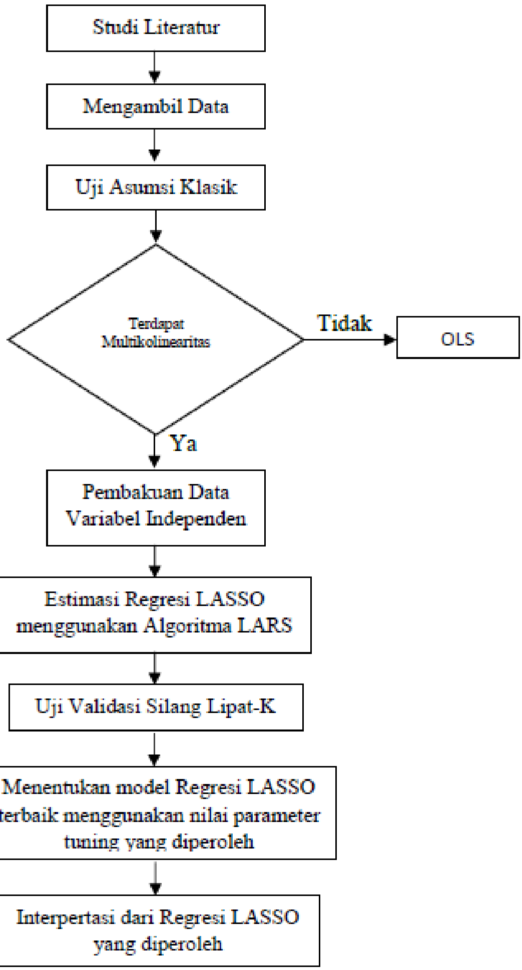
3.1 Metodologi Penelitian

Penelitian ini mengkaji mengenai penerapan regresi berganda pada data yang mengandung multikolinearitas. Analisis regresi berganda yang dapat digunakan pada data yang mengandung multikolinearitas salah satunya yaitu regresi *Least Absolute Shrinkage and Selection Operator* (LASSO). Di dalam peneltian ini akan digunakan Algoritma *Least Angle Regression* (LAR) yang sedikit dimodifikasi untuk memperoleh solusi regresi LASSO secara komputasi. Dan Penelitian ini menggunakan Bahasa pemrograman R untuk menyelesaikan metode regresi LASSO dibantu dengan *package* ‘LARS’ yang disusun oleh Hastie dan Efron pada tahun 2013.

Secara deskriptif paparan metodologi penelitian dalam skripsi ini diawali dengan studi literatur tentang konsep dasar metode regresi *Least Absolute Shrinkage and Selection Operator* (LASSO), lalu mengambil data sekunder dari sumber-sumber terpercaya seperti, Badan Pusat Statistik (BPS), publikasi dari Bank Indonesia, dan *Investing.com*. Kemudian menaksir model dengan variabel–variabel yang ada dengan variabel inflasi sebagai variabel depeden menggunakan metode *Ordinary Least Square* (OLS) untuk melakukan uji asumsi klasik. Dalam uji asumsi klasik, model tidak memenuhi syarat tidak terjadinya multikolinearitas sehingga model regresi dapat diselesaikan dengan regresi LASSO. Setelah itu.menaksir model regresi LASSO dengan menggunakan algoritma *Least Angle Regression* yang dimodifikasi (LARS) dengan menggunakan bahasa pemograman R dengan *package* ‘LARS’ dengan melakukan pembakuan data variabel independen terlebih dahulu sehingga berdistribusi N(0,1). Selanjutnya melakukan uji validasi silang lipat-K untuk memperoleh nilai parameter tuning yang optimal, nilai parameter tuning yang diperoleh digunakan untuk menentukan model regresi

LASSO terbaik. Dan pada akhirnya melakukan interpretasi dari model regresi LASSO terbaik yang telah diperoleh.

Untuk lebih jelasnya metodologi pada penelitian ini dapat dilihat dalam bentuk *flowchart* berikut ini.


Studi Literatur
Mengambil Data
Uji Asumsi Klasik
Terdapat Multikolinearitas
Tidak
OLS
Ya
Pembakuan Data Variabel Independen
Estimasi Regresi LASSO menggunakan Algoritma LARS
Uji Validasi Silang Lipat-K
Menentukan model Regresi LASSO terbaik menggunakan nilai parameter tuning yang diperoleh
Interpertasi dari Regresi LASSO yang diperoleh

3.2 Regresi *Least Absolute Shrinkage and Selection Operator* (LASSO)

Regresi LASSO pertama kali diperkenalkan oleh Robert Tibshirani (1996). Sesuai namanya regresi LASSO merupakan metode regresi berganda yang digunakan untuk *shrinkage* yaitu menyusutkan koefisien taksiran mendekati angka nol dan *selection operator* yaitu menyeleksi variabel-variabel independen sehingga menghasilkan model dengan variabel terbaik. Selain itu, regresi LASSO juga digunakan untuk data yang kontinu dan memerlukan variabel independen yang berdistribusi normal baku. Regresi LASSO ini terinspirasi dari regresi Ridge yang dapat menyelesaikan model regresi yang terdapat multikolinearitas diantara variabel-variabel independen dan dapat menyusutkan koefisien taksiran dari variabel-variabel tersebut hingga mendekati angka nol. Perbedaan antara regresi Ridge dengan LASSO hanya terdapat pada fungsi kendala dari model regresinya. Pada regresi ridge fungsi kendalanya yaitu $\sum_{j=1}^{p} \widehat{\beta}_j^{\,2} \leq t$ atau disebut juga L-2 norm sedangkan pada regresi Lasso fungsi kendalanya berbentuk $\sum_{j=1}^{p} |\beta_j| \leq t$ atau disebut L-1 norm. Penaksir koefisien pada regresi Lasso ($\hat{\beta}^{LASSO}$) diperoleh dengan cara meminimumkan persamaan berikut.

$$\hat{\beta}^{LASSO} = argmin \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \qquad (3.1)$$

dengan fungsi kendala $\sum_{j=1}^{p} |\beta_j| \leq t$,

Persamaan (3.1) dapat ditulis dalam bentuk persamaan pengali *Lagrange* sebagai berikut,

$$\hat{\beta}^{LASSO} = argmin \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (3.2)$$

dimana

y_i = variabel dependen pengamatan ke-i
β_0 = konstanta

β_j = koefisien dari variabel independen
X_{ij} = variabel independen
n = banyaknya observasi
p = banyaknya variabel independen dalam model

Nilai t merupakan suatu besaran yang mengontrol besarnya penyusutan pada koefisien regresi LASSO. Nilai t dinamakan parameter tuning dengan nilai $t \geq 0$.

Misalkan diketahui β_j merupakan penaksir OLS, dengan nilai t_0 didefinisikan $\sum_{j=1}^{p} |\hat{\beta}_j|$:

1. Jika nilai $t < t_0$, maka koefisien OLS akan menyusut ke arah nol, dan memungkinkan untuk menjadi tepat nol.
2. Jika nilai $t \geq t_0$, maka koefisien regresi LASSO memberikan hasil yang sama dengan koefisien OLS.

Menurut Tibshirani (2017, hlm. 4), λ pada persamaan (3.2) disebut sebagai parameter tuning yang berkorespondensi satu-satu dengan t artinya untuk setiap nilai $t \geq 0$ yang menghasilkan solusi $\hat{\beta}^{LASSO}$ terdapat $\lambda \geq 0$ sedemikian sehingga menghasilkan solusi $\hat{\beta}^{LASSO}$ juga.

Solusi regresi LASSO tidak memiliki solusi eksplisit seperti pada regresi Ridge karena pada fungsi kendala regresi LASSO berbentuk fungsi mutlak yang tidak dapat diturunkan. Namun pada kasus matriks dari variabel independen (X) yang orthogonal solusi regresi LASSO dapat diperoleh dengan mudah. Menurut Dwiananda (2015, hlm. 8-9) dalam bentuk matriks persamaan (3.2) dapat dinyatakan sebagai berikut,

$$(Y - X\beta)^T(Y - X\beta) + \lambda|\beta|I \tag{3.3}$$

dimana

Y = matriks pengamatan variabel dependen berukuran $(n \times 1)$
X = matriks variabel independen berukuran $(n \times p)$
β = matriks koefisien variabel independen berukuran $(p \times 1)$
λ = parameter tuning.
$|\beta|$ = matriks diagonal dengan elemen diagonal $|\beta_j|$.

Perhatikan bahwa persamaan (3.3) dapat diuraikan sebagai berikut,

$$Y^TY - 2Y^TX\beta + \beta^TX^TX\beta \ + \lambda|\beta_j|. \qquad (3.4)$$

Dengan adanya komponen $|\beta_j|$, persamaan (3.4) merupakan persamaan yang dapat diturunkan secara analitik jika pada matriks X berlaku sifat ortonormal. Turunan persamaan (3.4) terhadap β didapat,

$$-2Y^TX + 2(X^TX)\beta + \lambda sign(\beta) \qquad (3.5)$$

dimana

$$sign(\beta) = \begin{pmatrix} sign(\beta_1) \\ \vdots \\ sign(\beta_p) \end{pmatrix} \qquad (3.6)$$

Dengan menetapkan persamaan (3.5) sama dengan nol, diperoleh $\hat{\beta}^{LASSO}$. Asumsikan X ortonormal sehingga $X^TX = I$, maka pada kasus ortonormal taksiran OLS,

$$\hat{\beta}^{OLS} = (X^TX)^{-1}X^TY = X^TY.$$

Akibatnya, taksiran $\hat{\beta}^{LASSO}$ dari persamaan (3.5) diperoleh

$$\hat{\beta}^{LASSO} = \ (X^TX)^{-1}X^TY - \frac{\lambda}{2}(X^TX)^{-1}sign(\hat{\beta}^{LASSO})$$

Sehingga diperoleh taksiran koefisien LASSO,

$$\widehat{\beta_J}^{LASSO} = \widehat{\beta_J}^{OLS} - \frac{\lambda}{2}sign\left(\widehat{\beta_J}^{LASSO}\right). \qquad (3.7)$$

Pada persamaan (3.7), $sign\left(\widehat{\beta_J}^{LASSO}\right)$ selalu memiliki tanda yang sama dengan $\widehat{\beta_J}^{OLS}$, artinya $sign\left(\widehat{\beta_J}^{LASSO}\right) = sign\left(\widehat{\beta_J}^{OLS}\right)$. Sehingga persamaan (3.7) menjadi,

$$\widehat{\beta_J}^{LASSO} = \widehat{\beta_J}^{OLS} - \frac{\lambda}{2} sign\left(\widehat{\beta_J}^{OLS}\right)$$

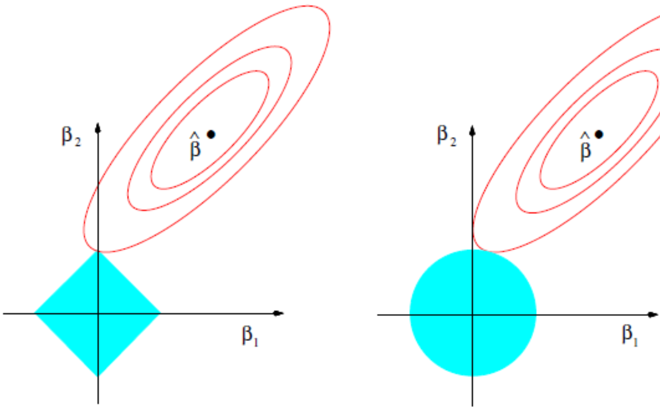
$$= \left(\widehat{\beta_J}^{OLS} - \frac{\lambda}{2}\right) \mathrm{I}_{\left[\widehat{\beta_J}^{OLS} \geq 0\right]} + \left(\widehat{\beta_J}^{OLS} - \frac{\lambda}{2}\right) \mathrm{I}_{\left[\widehat{\beta_J}^{OLS} < 0\right]} \quad (3.8)$$

Sehingga taksiran koefisien β dengan regresi LASSO diperoleh,

$$\widehat{\beta_J}^{LASSO} = sign\left(\widehat{\beta_J}^{OLS}\right)\left(\widehat{\beta_J}^{OLS} - \frac{\lambda}{2}\right)^{+} \quad (3.9)$$

dengan $J = 1,2,3, \ldots, p$

Gambar di bawah ini merupakan ilustrasi dari fungsi kendala antara regresi Lasso dan regresi Ridge dengan 2 variabel independen.



Gambar 3.1 Sebelah kiri area yang berwarna biru merupakan ilustrasi fungsi kendala dari regresi Lasso dengan rumus $|\beta_1| + |\beta_2| \leq t$ sedangkan sebelah kanan area kendala dari regresi Ridge dengan rumus $\beta_1^2 + \beta_2^2 \leq t^2$

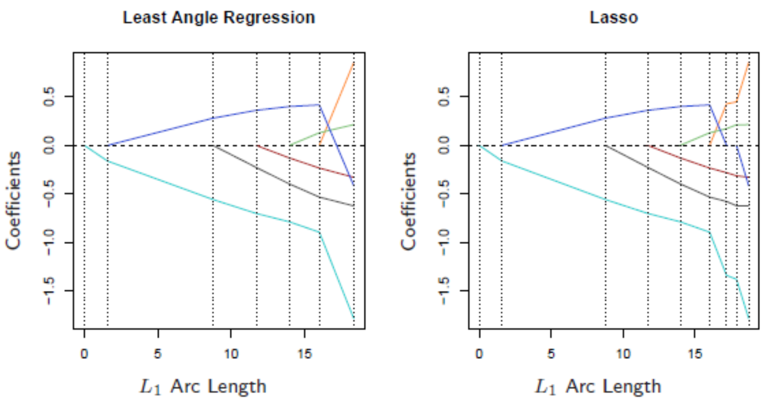
Pada Gambar 3.1, fungsi kendala dalam regresi Ridge berbentuk lingkaran, sedangkan kendala pada regresi Lasso berbentuk belah ketupat yang memiliki sudut. Titik berwarna hitam tersebut merupakan solusi dari OLS, dan kontur elips berwarna merah merupakan galat dari OLS. Titik saat area kendala pertama kali berpotongan dengan elips berwarna merah merupakan hasil taksiran dari regresi Ridge maupun LASSO. Perhatikan bahwa pada kendala regresi LASSO area yang pertama kali berpotongan dengan kontur elips berada di salah satu sudutnya, yang berarti bahwa salah satu koefisien taksiran dari regresi LASSO bernilai nol. Hal inilah yang menyebabkan regresi LASSO disebut sebagai seleksi operator. Sedangkan pada regresi Ridge, salah satu koefisien taksirannya menuju nol tetapi tidak tepat nol. Maka dalam regresi Ridge koefisien taksiran hanya disusutkan mendekati nol. Hal ini pula yang menjadi kelebihan dari regresi LASSO yaitu dapat digunakan sebagai seleksi variabel bebas dalam model, sehingga hanya variabel-variabel yang berpengaruh saja masuk ke dalam model. Hal ini bermanfaat untuk mempermudah dalam menginterpretasikan model regresi LASSO.

3.3 Algoritma *Least Angle Regression* (LAR)

Regresi LASSO merupakan metode regresi modern yang tidak memiliki solusi eksplisit untuk menaksir koefisien regresi ketika matriks $(X^T X)$ tidak berlaku ortonormal dan regresi LASSO juga termasuk permasalahan pemrograman kuadratik. Sehingga untuk menyelesaikan regresi tersebut dibutuhkan perhitungan secara komputasi. Sejak diperkenalkan oleh Robert Tibshirani pada tahun 1996, regresi LASSO kurang diminati karena komputasinya yang sangat lambat jika dibandingkan dengan jaman sekarang. Tetapi regresi LASSO mulai dikenal dan banyak digunakan oleh para ilmuwan untuk melakukan penelitian setelah ditemukan algoritma yang sangat efektif perhitungannya dibanding pemrograman kuadratik serta mampu memberikan solusi LASSO yang baik yaitu algoritma *Least Angle Regression* (LAR). Algoritma LAR merupakan sebuah algoritma untuk menghasilkan model linier yang ditemukan Efron dkk. pada tahun 2002. Algoritma LAR membutuhkan p langkah untuk mendapatkan koefisien taksiran OLS. Dengan memodifikasi algoritma LAR dapat memberikan koefisien taksiran metode LASSO. Algoritma yang dimodifikasi ini memiliki langkah yang lebih efisien dibanding metode LASSO itu sendiri (Hestie dkk, 2011, hlm. 73). Algoritma LAR yang dimodifikasi ini sering disebut juga sebagai algoritma LARS. Algoritma LARS memberikan jalan yang efisien dalam menyelesaikan regresi LASSO (Tibshirani, 2011). Algoritma ini dimulai dengan semua koefisien β sama dengan nol.

Algoritma LAR asli adalah sebagai berikut (Hastie dkk, 2011, hlm. 74).

1. Bakukan variabel independen sehingga memiliki nilai tengah nol dan varians satu. Mulai dengan residual $r = y - \bar{y}, \beta_1, \beta_2, \dots, \beta_p = 0$.
2. Cari variabel independen x_j yang paling berkorelasi dengan r.
3. Ubah nilai β_j dari 0 bergerak menuju koefisien kuadrat terkecil $\langle x_j, r \rangle$, sampai kompetitor lain x_k memiliki korelasi sebesar korelasi x_j dengan sisaan sekarang.
4. Ubah nilai β_j dan β_k bergerak dalam arah yang didefinisikan oleh koefisien kuadrat terkecil bersama dari sisaan sekarang dalam (x_j, x_k) sampai kompetitor x_l lain memiliki korelasi dengan sisaan sekarang dengan besaran yang sama.
5. Teruskan cara ini sampai semua p variabel bebas telah masuk. Setelah min(N-1,p) langkah, solusi model untuk OLS diperoleh.



Gambar 3.2 Gambar sebelah kiri menunjukkan koefisien dari algoritma LAR dalam bentuk fungsi L_1 *Arc Length*, sementara sebelah kanan menunjukkan koefisien regresi LASSO. L_1 *Arc Length* dari kurva $\hat{\beta}(s)\ untuk\ s \in [0, S]$ yang dapat terdiferensialkan, yang didefinisikan sebagai $TV(\beta, S) = \int_0^S \left\|\hat{\beta}(s)\right\|_1 ds$ dimana $\hat{\beta}(s) = \frac{\partial \hat{\beta}(s)}{\partial s}$ (Hastie dkk, 2011)

Pada gambar di atas dapat dilihat bahwa kedua gambar identik sebelum koefisien yang berwarna biru tua memotong garis nol. Sehingga memodifikasi algoritma LAR untuk mendapatkan solusi LASSO adalah dengan memodifikasi dengan menambahkan *statement* langkah ke-4 yaitu.
"Jika koefisien bukan nol mencapai nilai nol, keluarkanlah variabel tersebut dari gugus variabel aktif dan hitung kembali arah OLS bersama".

LAR selalu mengambil p langkah untuk mendapatkan penaksir OLS secara penuh, sedangkan modifikasi LAR untuk metode LASSO dapat memiliki lebih dari p langkah untuk mendapatkannya. Algoritma LASSO dengan memodifikasi LAR adalah suatu cara yang efisien dalam komputasi solusi masalah LASSO khususnya ketika p > N. Pada output algoritma LAR, akan muncul Plot pergerakan variabel-variabel independen dengan parameter tuning bentuk standar (s). Menurut Hastie (2011, hlm. 69) nilai parameter tuning s dapat diperoleh dengan rumus berikut,

$$s = \frac{t}{\sum \left|\hat{\beta}_j^{OLS}\right|}$$

dengan $j = 1,2, \dots, p$. Jika nilai $s = 1$, maka solusi regresi LASSO akan sama dengan solusi OLS.

3.4 Validasi Silang Lipat-K

Validasi silang atau yang disebut juga *K-Fold Cross Validation* (CV) merupakan metode umum yang digunakan dalam pemilihan nilai penalti dari regresi LASSO. Jika diinginkan untuk memilih nilai dari tuning parameter dari bentuk umum persamaan penaksir koefisien regresi LASSO, maka validasi silang merupakan cara standarnya (Tibshirani, 2017). Ide dari validasi silang lipat-K adalah membagi data menjadi K bagian yang sama besar, lalu dari K bagian tersebut dipilih salah satu bagian sebagai data validasi dan bagian yang lain digunakan sebagai data *training*. Data *training* digunakan untuk menghitung nilai $\hat{\beta}$ dan data validasi digunakan untuk menguji kebaikan prediksi dari $X\hat{\beta}$. Proses validasi silang diulang sampai K kali dan setiap bagian digunakan hanya sekali untuk menjadi data validasi. Nilai dari validasi silang ini merupakan penaksir dari galat prediksi atau *prediction galat* $(\widehat{PE})$. Validasi silang lipat-K yang sering digunakan yaitu dengan K=5 atau K=10. Berikut merupakan contoh pengelompokkan data validasi silang lipat-5.

1	2	3	4	5
training	training	validasi	training	training

Gambar 3.3 contoh validasi silang lipat-5

Diberikan training data (x_i, y_i), i=1,2,...,n dan *penaksir* $\hat{f}_t$ yang berhubungan dengan parameter tuning t. Dan berikut merupakan prosedur perhitungan validasi silang lipat-K.

1. Membagi set data menjadi K bagian yang sama besar, $F_1, F_2, \dots, F_K$.
2. Untuk $v = 1,2, \dots, K$, perhatikan bahwa pada data training $(x_i, y_i), i \notin F_K$ dan pada data validasi $(x_i, y_i), i \in F_K$. Untuk setiap nilai t lakukan metode regresi LASSO lalu hitung total galatnya
$$e_v(t) = \sum_{i \in F_K} \left(y_i - \hat{f}_t^{-v}(x_i)\right)^2.$$
3. Untuk setiap parameter tuning, hitung rata-rata galat pada semua bagian lipat

$$CV(t) = \frac{1}{n}\sum_{v=1}^{K} e_v(t) = \frac{1}{n}\sum_{v=1}^{K}\sum_{i\in F_K}\left(y_i - \hat{f}_t^{-v}(x_i)\right)^2.$$

4. Diperoleh taksiran parameter tuning,

$$\hat{t} = min\{CV(t)\}$$

Setelah diperoleh nilai parameter tuning taksiran yang optimal, nilai tersebut digunakan untuk menentukan berhentinya iterasi pada algoritma LAR solusi regresi LASSO terbaik.