PERBANDINGAN ESTIMASI LASSO DAN LASSO *LEAST TRIMMED SQUARES* UNTUK MENGANALISIS DATA DIMENSI BESAR (Studi Kasus: Faktor – Faktor Yang Mempengaruhi Penyebaran Penyakit Tuberkulosis Di Sulawesi Selatan)

Comparison of LASSO and LASSO Least Trimmed Squares Estimation to Analyze High Dimensional Data (*Case Study: Factors Affecting the Spread of Tuberculosis in South Sulawesi*)

TRIGARCIA MALEACHI RANDA

PROGRAM STUDI MAGISTER STATISTIKA
SEKOLAH PASCASARJANA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2022

PERBANDINGAN ESTIMASI LASSO DAN LASSO *LEAST TRIMMED SQUARES* UNTUK MENGANALISIS DATA DIMENSI BESAR (Studi Kasus: Faktor – Faktor Yang Mempengaruhi Penyebaran Penyakit Tuberkulosis Di Sulawesi Selatan)

Tesis

sebagai salah satu syarat untuk mencapai gelar magister

Program Studi Magister Statistika

Disusun dan diajukan oleh

TRIGARCIA MALEACHI RANDA

H062202006

kepada

PROGRAM STUDI MAGISTER STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

TESIS

PERBANDINGAN ESTIMASI LASSO DAN LASSO *LEAST TRIMMED SQUARES* UNTUK MENGANALISIS DATA DIMENSI BESAR (Studi Kasus: Faktor – Faktor Yang Mempengaruhi Penyebaran Penyakit Tuberkulosis Di Sulawesi Selatan)

TRIGARCIA MALEACHI RANDA
H062202006

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Program Studi Magister Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin
pada tanggal 15 November 2022
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping
Dr. Dr. Georgina M. Tinungki, M.Si.	**Dr. Nurtiti Sunusi, S.Si M.Si.**
NIP. 19620926 198702 2 001	NIP. 1972017 199703 2 002
Ketua Program Studi Magister Statistika,	Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin
Dr. Dr. Georgina M. Tinungki, M.Si	**Dr. Eng. Amiruddin, M.Si**
NIP. 19620926 198702 2 001	NIP. 19720515 199702 1 002

PERNYATAAN KEASLIAN TESIS
DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, tesis berjudul "Perbandigan Estimasi LASSO Dan LASSO Least Trimmed Squares Untuk Menganalisis Data Dimensi Besar (Studi Kasus: Faktor – Faktor Yang Mempengaruhi Penyebaran Penyakit Tuberkulosis Di Sulawesi Selatan)" adalah benar karya saya dengan arahan dari komisi pembibing (Dr. Dr. Georgina Maria Tinungki, M.Si sebagai Pembibing Utama dan Dr. Nurtiti Sunusi, S.Si., M.Si sebagai Pembimbing Pendamping). Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasinya yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka tesis ini. Sebagian dari isi tesis ini telah dipublikasikan di Jurnal (EKSAKTA: Journal of Sciences and Data Analysis, ISSN: 2720-9326, Vol. 3, 103-112, DOI: 10.20885/EKSAKTA.vol3.iss2.art6) sebagai artikel dengan judul "Modeling the Proportion of Tuberculosis Cases in South Sulawesi using Sparse Least Trimmed Squares"

Dengan ini saya melimpahkan hak cipta dari karya tulis saya berupa tesis ini kepada Universitas Hasanuddin.

Makassar, 15 November 2022

Yang Menyatakan,

Trigarcia Maleachi Randa

NIM. H062202006

UCAPAN TERIMA KASIH

Puji syukur atas kehadirat Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyusun dan menyelesaikan tesis ini. Penulis menyadari sepenuhnya bahwa apa yang dikemukakan dalam tesis ini masih jauh dari kesempurnaan yang merupakan sebagai akibat dari keterbatasan kemampuan serta berbagai kesulitan yang penulis hadapi dalam menyusun tesis ini.

Penulis memanjatkan doa kepada Tuhan Yang Maha Esa agar memberikan rahmat-Nya kepada pihak yang banyak membantu dalam penyelesaian tesis ini. Penulis juga percaya tesis ini dapat selesai bukan hanya dengan kekuatan pikiran penulis semata akan tetapi karena bantuan dari berbagai pihak juga, baik selama proses perkuliahan bahkan sampai proses pengerjaan tesis di Program Magister Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Namun demikian, penulis dengan senang hati menerima kritik dan saran yang bersifat membangun dari pembaca karya tulis ini demi sempurnanya tesis ini.

Terima kasih yang tak terhingga kepada kedua orang tua tercinta dan saudariku atas doa yang tak pernah putus, dukungan serta segala kebaikan mereka yang sampai kapan pun takkan pernah bisa terbalaskan atas kasih sayang yang tiada henti dalam penyelesaian tesis ini. Selanjutnya, saya ingin menyampaikan juga rasa hormat dan terima kasih kepada:

1. **Prof. Dr. Ir. Jamaluddin Jompa, M.Sc.** selaku Rektor Universitas Hasanuddin.
2. **Dr. Eng. Amiruddin, M.Si**. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam beserta seluruh jajarannya.
3. **Dr. Nurtiti Sanusi, M.Si.** selaku Ketua Departemen Statistika yang menjadi Pembimbing Pertama yang telah bersabar dan bersedia meluangkan banyak waktunya untuk membimbing penulis dan memberikan masukan dalam penyelesaian tesis ini.
4. **Dr. Dr. Georgina Maria Tinungki, M.Si**. selaku Ketua Program Studi Magister Statistika yang menjadi Pembimbing Utama yang telah bersabar dan bersedia meluangkan banyak waktunya untuk membimbing penulis dan memberikan

ilmu, dukungan, dan motivasi serta kemudahan kepada penulis dalam berbagai hal selama menjalani Pendidikan di Departemen Statistika.

5. **Dr. Nirwan Ilyas, M.Si** selaku penguji penulis yang telah bersedia memberikan masukan-masukan dan arahan dalam penyusunan tesis.
6. **Dr. Anna Islamiyati, S.Si., M.Si** selaku penguji penulis yang telah bersedia memberikan masukan-masukan dan arahan dalam penyusunan tesis.
7. **Dr. Erna Tri Herdiani, M.Si** selaku penguji penulis yang telah bersedia memberikan masukan-masukan dan arahan dalam penyusunan tesis.
8. Teman - teman **Lab Statistika** yaitu **Agung Muhammad Takdir, Ratmila, Ainun Utari, Asnidar, Samsir Aditya Ania**, dan **Andi Dulung Laimbong** atas doa, semangat serta kebersamaannya selama ini yang banyak membantu penulis.
9. Teman – teman Mahasiswa Program Magister Statistika angkatan ketiga terkhusus terima kasih atas nasehat dan dukungan luar biasa kepada penulis, terkhusus **Kak Mubasiratul Munawaroh**, **A. Eka Hermia Fitrianingsy**, **Nurul Hidayanti Anggraini**, **Andi Isna Yunita**, **Alimatun Najiha**, **Maktisen Ena**, **Hedi Kuswanto**, **Muh. Qardawi Hamzah**, dan **Mushinil Haq**.

Semoga Allah SWT memberikan pahala yang berlipat ganda atas segala kebaikan yang telah diberikan kepada penulis dan semoga penulisan tesis ini bermanfaat bagi perkembangan ilmu pengetahuan dan teknologi, khususnya dalam dunia statistika dan data sains.

Penulis,

Trigarcia Maleachi Randa

ABSTRAK

TRIGARCIA MAELACHI RANDA. **Perbandingan Estimasi LASSO Dan LASSO Least Trimmed Squares Untuk Menganalisis Data Dimensi Besar (Studi Kasus: Faktor – Faktor Yang Mempengaruhi Penyebaran Penyakit Tuberkulosis Di Sulawesi Selatan)** (dibimbing oleh Georgina Maria Tinungki, dan Nurtiti Sunusi)

Tuberkulosis (TB) merupakan penyebab kematian ke-13 di dunia dan menjadi penyakit menular yang mematikan di Indonesia. Salah satu provinsi penyumbang kasus TB terbanyak di Indonesia pada tahun 2018 adalah Sulawesi Selatan dengan 84 kasus per 100,000 penduduk. Penelitian ini bertujuan untuk mengidentifikasi peubah yang dapat menjelaskan proporsi kasus TB di Sulawesi Selatan. Data yang digunakan memiliki banyak peubah bebas, dan terdapat pencilan. Analisis LASSO *Least Trimmed Squares* (LTS) dapat digunakan untuk menangani data yang memiliki banyak peubah bebas dan pencilan. Analisis data dilakukan pada data simulasi dan data proporsi kasus TB Berdasarkan rata-rata koefisien determinasi (R^2) dan *Root Mean Square Error* (RMSE), LASSO LTS memiliki performa terbaik dibandingkan LASSO pada setiap skenario kondisi ukuran data dan persentase pencilan. Analisis proporsi kasus TB di Sulawesi Selatan menghasilkan kesimpulan bahwa model LASSO LTS berhasil menyeleksi dan menyusutkan variabel menjadi 11 variabel dan menunjukkan hasil yang baik berdasarkan nilai R^2 dan RMSE untuk evaluasi model. Faktor-faktor tersebut bisa menjadi fokus pemerintah jika ingin menurunkan proporsi kasus TB di Sulawesi Selatan.

Kata Kunci: LASSO, pencilan, regresi terpenalti, regresi kekar, tuberkulosis

ABSTRACT

TRIGARCIA MAELACHI RANDA. **Comparison of LASSO and LASSO Least Trimmed Squares Estimation to Analyze High Dimensional Data (Case Study: Factors Affecting the Spread of Tuberculosis in South Sulawesi)** (supervised by Georgina Maria Tinungki, and Nurtiti Sunusi)

Tuberculosis is the 13th leading cause of death in the world and is a deadly infectious disease in Indonesia. One of the provinces that contributed the most TB cases in Indonesia in 2018 was South Sulawesi with 84 cases per 100,000 population. This study aims to identify variables that can explain the proportion of TB cases in South Sulawesi. The data used has many independent variables, and there are outliers. LASSO Least Trimmed Squares (LTS) analysis can be used for handle data that has many independent variables and outliers. Data analysis was performed on simulation data and proportion of TB cases data Based on the average coefficient of determination (R^2) and Root Mean Square Error (RMSE), LASSO LTS has the best performance than LASSO in each scenario condition of data size and outlier percentage. Analysis of proportion of TB cases in South Sulawesi resulted the conclusion that LASSO LTS model successfully select and shrink the variables to 11 variables only and shows satisfying results based on the value of R^2 and RMSE for the model evaluation. The government can focus on these factors if they want to reduce the proportion of TB cases in South Sulawesi.

Kata Kunci: LASSO, outliers, penalized regression, robust regression, tuberculosis

DAFTAR ISI

Halaman

DAFTAR TABEL

Nomor Urut Halaman

DAFTAR GAMBAR

Nomor Urut Halaman

DAFTAR LAMPIRAN

Nomor Urut Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tuberkulosis (TB) adalah penyakit menular yang disebabkan oleh kuman *Mycobacterium tuberculosis* dan merupakan salah satu dari 10 penyebab utama kematian di seluruh dunia (WHO, 2018). Menurut *World Health Organization* (WHO) pada tahun 2018, Indonesia berada pada peringkat ketiga dengan penderita TB tertinggi di Dunia setelah India dan China. Sebanyak 10 juta orang menderita penyakit tuberkulosis, dan 1.2 juta meninggal karena penyakit ini pada tahun 2018. Sedangkan Provinsi Sulawesi Selatan digunakan dalam penelitian ini karena berdasarkan data dari Kementerian Kesehatan Republik Indonesia (Kemenkes RI) pada tahun 2018, Sulawesi Selatan merupakan salah satu Provinsi dengan jumlah semua kasus TB yang diobati dan dilaporkan di antara 100,000 penduduk tertinggi di Indonesia.

Kemudian pada tahun 2020 menurut data Kemenkes RI jumlah kasus tuberkulosis di Sulawesi Selatan sudah mengalami penurunan. Namun, penurunan jumlah kasus TB masih tetap harus diwaspadai karena adanya disparitas (kesenjangan) penyebaran penyakit TB antar daerah di Sulawesi Selatan, misalnya kabupaten Barru dan kabupaten Sidrap pada tahun 2018 memiliki jumlah kasus tuberkulosis berturut-turut sebanyak 182 dan 493 kasus, namun pada tahun 2020 mengalami kenaikkan jumlah kasus secara berturut-turut sebesar 202 dan 591 kasus (BPS, 2020). Hal tersebut diduga karena datanya mengandung pencilan (*outlier*). Oleh sebab itu, perlu dilakukan penelitian lebih lanjut mengenai faktor-faktor yang mempengaruhi penyebaran penyakit tuberkulosis di Provinsi Sulawesi Selatan.

Seluruh data yang digunakan dalam penelitian ini diambil dari publikasi Riskesdas dan BPS pada tahun 2018. Peubah bebas yang dipilih adalah peubah yang berkaitan dengan kesehatan, ekonomi, sumber daya manusia (SDM) dan lingkungan. Faktor-faktor ini dipilih untuk diteliti karena menurut penelitian yang dilakukan oleh Sejati dan Sofiana (2015), secara keseluruhan penelitian tentang tuberkulosis melibatkan faktor-faktor tersebut. Keseluruhan data yang terkumpul memiliki 24 amatan, yaitu kabupaten/kota di Sulawesi Selatan, dengan 25 peubah

bebas yang terkait dengan kesehatan, ekonomi, SDM dan lingkungan. Peubah respon yang digunakan adalah proporsi tuberkulosis di Sulawesi Selatan tahun 2018 untuk masing-masing kabupaten/kota. Jika melihat perbandingan ukuran antara amatan dan jumlah peubah bebas yang digunakan, maka dapat dikatakan bahwa data ini termasuk yang memiliki dimensi besar, karena jumlah peubah bebasnya lebih besar dari jumlah amatannya.

Istilah dimensi besar mengacu pada kasus di mana banyaknya parameter yang tidak diketahui untuk diestimasi, *p*, jauh lebih besar daripada jumlah pengamatan, *n*, yaitu $p >> n$. Saat memodelkan hubungan antara satu kejadian dengan kejadian lainnya, teknik statistika yang lazim digunakan adalah analisis regresi. Analisis regresi akan menghadapi masalah jika jumlah peubah bebas sangat banyak, sehingga diperlukan metode alternatif untuk menyelesaikan persoalan ini. Salah satu cara untuk menyelesaikan masalah ini di dalam literatur statistika dikenal metode regresi terpenalti.

Regresi terpenalti merupakan metode yang pendugaan parameternya didasarkan pada peminimuman jumlah kuadrat galat yang diberi penalti (Hastie *et al*., 2015). Metode regresi terpenalti yang sering digunakan untuk data dengan dimensi besar adalah regresi LASSO (*Least Absolute Shrinkage and Selection Operator*). LASSO merupakan metode komputasi dengan menggunakan pemrograman kuadratik yang dapat mengatasi masalah multikolinearitas dengan menyusutkan koefisien regresi mendekati nol bahkan hingga tepat nol sehingga dapat melakukan seleksi model (Tibshirani, 1996). Tetapi LASSO tidak bersifat *robust* terhadap pencilan (Alfons *et al*., 2011). Terdapat beberapa metode regresi terpenalti yang bersifat *robust* terhadap pencilan dalam literatur dan di antaranya yang sangat direkomendasikan adalah LASSO *least trimmed squares* (LTS).

Menurut Wilems dan Aelst (2005), metode LTS merupakan suatu metode pendugaan parameter regresi *robust* untuk meminimumkan jumlah kuadrat sisaan sebanyak *h*. Akan tetapi, terdapat kekurangan pada metode LTS apabila data memiliki dimensi besar, maka metode ini tidak dapat dilakukan. Salah satu pengembangan metode LTS adalah LASSO LTS yang dapat digunakan mengatasi masalah ini. LASSO LTS diperkenalkan oleh Alfons *et al*. (2013), dimana metode ini merupakan gabungan antara metode LASSO dan LTS yang dapat mengatasi pencilan untuk data dimensi besar.

Beberapa penelitian sebelumnya mengenai metode LASSO dan LTS pernah dilakukan. Rohmawati *et al.* (2018) membandingkan metode estimasi LTS dan S (*Scale)* dalam mengestimasi model orde dua pada metode permukaan respon, dan diperoleh hasil estimasi LTS lebih baik dibandingkan estimasi S berdasarkan nilai R^2. Yanke *et al.* (2022) meneliti tentang penanganan masalah multikolinearitas dalam pemodelan regresi pertumbuhan ekonomi Indonesia berdasarkan teori pertumbuhan ekonomi endogen. Dengan menerapkan metode seleksi peubah (*backward*, *forward*, dan *stepwise*), regresi komponen utama (PCR), *partial least square* (PLS), dan metode regularisasi (*Ridge*, LASSO, dan *Elastic Net*) untuk menyelesaikan masalah multikolinearitas, hasil yang diperoleh bahwa metode LASSO merupakan metode terbaik untuk mengatasi masalah multikolinearitas berdasarkan nilai *mean square error* (MSE).

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, maka permasalahan utama yang ingin dibahas dalam penelitian ini adalah sebagai berikut:

1. Bagaimana perbandingan dari estimasi LASSO dan LASSO LTS?
2. Apa saja faktor-faktor yang mempengaruhi kasus penyebaran penyakit Tuberkulosis di Sulawesi Selatan berdasarkan model terbaik?

1.3 Tujuan Penelitan

Dari permasalahan diatas, maka tujuan penelitian ini adalah sebagai berikut:

1. Menunjukkan perbandingan estimasi LASSO dan LASSO LTS.
2. Mengetahui faktor-faktor yang mempengaruhi kasus penyebaran penyakit Tuberkulosis di Sulawesi Selatan berdasarkan model terbaik.

1.4 Manfaat Penelitan

Adapun manfaat yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Memberikan sumbangsih keilmuan dalam menerapkan estimasi LASSO dan LASSO LTS.
2. Sebagai bahan rujukan kepada Pemerintah dalam hal ini Kementerian Kesehatan, agar perlu memperhatikan faktor-faktor yang berpengaruh secara signifikan terhadap penyebaran penyakit tuberkulosis.

1.5 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Algoritma yang digunakan pada metode LTS adalah algoritma LTS yaitu penggabungan FAST-LTS dan C-*Steps*. Karena dalam proses estimasinya menggunakan algoritma LTS hanya akan memangkas sebaran data berdasarkan jumlah pencilan yang teramati sehingga akan menghasilkan fungsi objektif yang mengecil dan konvergen.
2. Algoritma yang digunakan pada metode LASSO adalah algoritma LARS yang merupakan modifikasi LAR (*Least Angle Regression*) untuk LASSO. Karena LARS merupakan algoritma yang lebih efisien digunakan dan telah dimodifikasi sehingga mempermudah dalam komputasi LASSO

BAB II

TINJAUAN PUSTAKA

2.1 Analisis Regresi

Analisis regresi merupakan suatu teknik statistika untuk memeriksa dan memodelkan hubungan antar peubah (Montgomery *et al*., 2012). Analisis regresi merupakan teknik statistika yang paling banyak digunakan. Dalam analisis regresi terdapat istilah peubah Y dan peubah X. Peubah Y adalah peubah tak bebas atau peubah respon, sedangkan X adalah peubah bebas atau peubah prediktor. Analisis regresi yang hanya melibatkan satu peubah bebas disebut analisis regresi linier sederhana, sedangkan analisis regresi linier berganda melibatkan p peubah bebas. Hal utama dari analisis regresi adalah menduga parameter yang tidak diketahui dari model regresi.

Regresi linier sederhana memiliki beberapa asumsi, yaitu nilai harapan (rataan) sisaan sama dengan nol, ragam sisaan homogen, sisaan saling bebas, sisaan menyebar normal dengan rataan nol dan ragam (σ^2), serta sisaan bebas terhadap peubah bebas. Pada regresi linier berganda terdapat asumsi tambahan bahwa tidak ada multikolinearitas pada peubah bebas.

Model linier artinya linier dalam parameter. Model regresi linier memiliki bentuk sebagai berikut: (Montgomery dan Runger 2003)

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \qquad (2.1)$$

dimana:

i : $1,2,\cdots,n$

j : $1,2,\cdots,p$

y_i : Vektor peubah respon berukuran $n \times 1$

x_{ij} : Matriks peubah bebas berukuran $n \times (p+1)$

β_0 : Intersep

β_j : Slope atau kemiringan

e_i : Vektor sisaan (*error*) berukuran $n \times 1$

2.2 Metode Ordinary Least Square (OLS)

Metode kuadrat terkecil atau disebut juga *Ordinary Least Square* (OLS) merupakan metode yang digunakan untuk menduga koefisien regresi linier (β_0, β_j) dengan cara meminimumkan jumlah kuadrat sisaan (JKS) (Hastie *et al*., 2008). Persamaan yang diminimumkan adalah sebagai berikut:

$$\text{JKS} = \sum_{j=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \right)^2 \tag{2.2}$$

Penduga $\hat{\beta}$ dengan OLS akan menghasilkan penduga yang tak bias serta solusi unik. Pendugaan koefisien regresi dengan OLS memiliki kuadrat tengah sisaan terkecil di antara semua penduga linier yang tak bias. Namun, pada kondisi tertentu (misalnya multikolinieritas dan peubah sangat banyak), metode kuadrat terkecil sering tidak memuaskan. Hal tersebut disebabkan karena adanya masalah keakuratan prediksi yang mengakibatkan penduga kuadrat terkecil memiliki bias rendah tetapi ragam besar. Selain itu, semakin banyak peubah bebas maka model semakin sulit diinterpretasikan (Tibshirani, 1996).

2.3 Metode *Least Trimmed Squares* (LTS)

Menurut Chen (2002), estimasi *Least Trimmed Squares* (LTS) merupakan suatu metode *High Breakdown Value* yang diperkenalkan oleh Rousseeuw pada tahun 1984. Estimator LTS meminimumkan fungsi objektif yaitu jumlah kuadrat sisaan sebanyak h. LTS atau kuadrat terpangkas terkecil adalah salah satu teknik statistik yang digunakan untuk mengestimasi parameter dari model regresi linier yang memberikan alternatif *robust* ke metode regresi klasik berdasarkan meminimalkan jumlah kuadrat galat. Menurut Doornik (2011), penduga untuk LTS adalah sebagai berikut,

$$\hat{\beta}_{LTS} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{h} e_{(i)}^2 \tag{2.3}$$

dimana,

$e_{(i)}^2$: kuadrat galat dari h pengamatan yang telah diurutkan $e_{(1)}^2 \leq e_{(2)}^2 \leq \cdots \leq e_{(n)}^2$

h : konstanta pemotongan

Menurut Wilems dan Aelst (2005), h optimal yang digunakan pada metode LTS adalah

$$h = \left[\frac{n+p+1}{2}\right] \tag{2.4}$$

Nilai h merupakan jumlah pengamatan yang digunakan untuk menduga parameter model regresi. Dengan p merupakan banyaknya peubah bebas dan n merupakan banyaknya pengamatan. Sama halnya dengan penduga lain pada regresi *robust*, pada metode LTS diberikan pembobot (w_{ii}) pada data sehingga data *outlier* tidak mempengaruhi model parameter hasil estimasi. Pembobot w_{ii} dapat disajikan dalam bentuk matriks, yang dinotasikan sebagai $\mathbf{W}$,

$$\mathbf{W} = \begin{pmatrix} w_{11} & 0 & \cdots & 0 \\ 0 & w_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{nn} \end{pmatrix} \tag{2.5}$$

Menurut Chatzinakos dan Zioutas (2014), data pengamatan untuk metode LTS yang diidentifikasi sebagai pencilan/*outlier* diberikan bobot nol dan lainnya diberi bobot satu, dengan pembobotnya adalah sebagai berikut,

$$w_{ii} = \begin{cases} 1\,, & \dfrac{|e_i|}{\hat{\sigma}^*} \\ 0\,, & \text{lainnya} \end{cases} \tag{2.6}$$

dimana,

$$\hat{\sigma}^* = d_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h} e_{(i)}^{\mathbf{2}}} \tag{2.7}$$

$$d_{h,n} = \frac{1}{\sqrt{1 - \dfrac{2n}{hc_{h,n}}\emptyset\left(\dfrac{1}{c_{h,n}}\right)}} \tag{2.8}$$

$$c_{h,n} = \frac{1}{\Phi^{-1}\left(\dfrac{h+n}{2n}\right)} \tag{2. 9}$$

$$W_{Scale} = \sqrt{\frac{\sum w_i e_i^2}{\sum w_i - k}} \tag{2.10}$$

Dengan ∅ merupakan fungsi kepadatan peluang dan distribusi normal standar, Φ merupakan fungsi distribusi kumulatif dan k merupakan banyaknya parameter pada model.

Menurut Willems and Aelst (2005) untuk menduga parameter model dengan menggunakan metode LTS digunakan algoritma LTS yaitu penggabungan *FAST*-LTS dan C-*Steps*, langkah-langkah algoritma LTS adalah sebagai berikut:

1. Menduga parameter model regresi menggunakan OLS
2. Menentukan kuadrat galat $e_{(i)}^2$, kemudian mengurutkan nilai $e_{(i)}^2$ dari yang terkecil ke terbesar
3. Menentukan nilai h yaitu pada persamaan (2.4) dan menghitung $\sum_{i=1}^{h} e_{(i)}^2$
4. Menghitung estimasi parameter $\hat{\beta}_{new}$ melalui OLS dari h pengamatan (nilai h awal)
5. Menentukan kuadrat residual $e_{(i)\,new}^{\mathbf{2}}$ yang didapatkan dari estimasi parameter $\hat{\beta}_{new}$, kemudian menghitung h_{new} observasi dengan mengurutkan nilai $e_{(i)\,new}^{\mathbf{2}}$ dari yang terkecil ke terbesar
6. Menghitung $\sum_{i=1}^{h_{new}} e_{(i)\,new}^2$
7. Ulangi langkah 4 sampai 6, iterasi dihentikan hingga dihasilkan hasil yang konvergen yaitu jika nilai $Q_{LTS}^{m+1} \leq Q_{LTS}^{m}$, m adalah jumlah iterasi
8. Memberikan bobot pada data dengan menerapkan *Final Weighted Least Square*.

2.4 *Least Absolute Shrinkage and Selection Operator* (LASSO)

Tibshirani pertama kali memperkenalkan metode Least Absolute Shrinkage and Selection Operator (LASSO) pada tahun 1996. Metode LASSO merupakan metode yang dapat digunakan untuk menyusutkan koefisien regresi dari peubah bebas yang memiliki korelasi tinggi dengan galat menjadi tepat pada nol atau mendekati nol. Sehingga dapat dikatakan bahwa metode LASSO berfungsi sebagai metode untuk melakukan seleksi variabel sekaligus hal tersebut dapat mengatasi masalah multikolinearitas (Tibshirani 1996). Penduga koefisien LASSO diperoleh dengan menggunakan pemrograman kuadratik karena tidak dapat diperoleh dalam bentuk tertutup seperti pada OLS atau regresi gulud. Metode LASSO mulai dikenal

setelah ditemukannya algoritma Least Angle Regression (LAR) pada tahun 2004 oleh Efron (Hastie *et al*., 2008).

Menurut Zhao dan Yu (2006), persamaan secara umum LASSO dinyatakan sebagai berikut:

$$\mathbf{y}^{**} = \mathbf{X}^{**} + \boldsymbol{\beta} + \mathbf{e}^{**} \tag{2.11}$$

dimana:

$\mathbf{y}^{**}$: Vektor peubah respon berukuran $n \times 1$

$\mathbf{X}^{**}$: Matriks peubah bebas berukuran $(n \times p)$

$\boldsymbol{\beta}$: Vektor dari koefisien LASSO berukuran $(k+1) \times 1$

$\mathbf{e}^{**}$: Vektor sisaan (*error*) berukuran $n \times 1$

Menurut Tibshirani (1996), penduga koefisien LASSO dapat diperoleh dengan meminimumkan jumlah kuadrat sisaan (JKS) dengan persamaan sebagai berikut:

$$\hat{\beta}_{\text{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{j=1}^{n} \left(y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \right)^2 \right\} \tag{2.12}$$

dengan syarat $\sum_{j=1}^{k} |\beta_j| \leq \lambda$. Nilai λ merupakan parameter *tuning* yang mengontrol penyusutan koefisien LASSO dengan $t \geq 0$.

Menurut Tibshirani (1996), jika $\lambda < \lambda_0$ dengan $\lambda_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$ maka akan menyebabkan koefisien menyusut mendekati nol atau tepat pada nol atau tepat pada nol, sehingga LASSO akan berperan sebagai seleksi variabel. Akan tetapi jika $\lambda > \lambda_0$ maka penduga koefisien LASSO memberikan hasil yang sama dengan penduga kuadrat terkecil. Koefisien regresi LASSO ditentukan berdasarkan parameter *tuning* yang sudah dibakukan $s = \frac{\lambda}{\sum_{j=1}^{k} |\hat{\beta}_j^0|}$ dengan $\lambda = \sum_{j=1}^{p} |\hat{\beta}_j|$, $\hat{\beta}_j^0$ adalah penduga kuadrat terkecil untuk model penuh, nilai λ optimal diperoleh melalui validasi silang (Randa *et al*., 2022).

2.5 Algoritma *Least Angle Regression* (LAR)

Least Angle Regression (LAR) merupakan suatu metode regresi yang algoritmanya dapat dimodifikasi menjadi algoritma komputasi untuk metode LASSO. LAR menghasilkan efisiensi algoritma dalam menduga koefisien LASSO dengan

komputasi yang lebih cepat dibandingkan pemrograman kuadratik Hastie *et al*. (2008).

LAR mempunyai kesamaan pada regresi dengan metode *forward stepwise*. Regresi dengan metode *forward stepwise* membentuk model secara bertahap dengan menambahkan sebuah peubah bebas pada satu waktu. Pada setiap tahapannya akan diidentifikasi peubah bebas terbaik untuk diikutsertakan ke dalam model dan dihitung kembali nilai penduga kuadrat terkecilnya. Least Angle Regression menggunakan cara kerja yang sama tetapi hanya memasukkan sebanyak peubah bebas sebagaimana mestinya. Pada langkah pertama dilakukan identifikasi peubah bebas yang paling berkorelasi dengan peubah respon. Berikut merupakan tahapan algoritma LAR asli: (Hastie *et al*., 2008):

1. Membakukan peubah bebas yang digunakan sehingga memiliki nilai tengah nol dan ragam satu. Tujuan dilakukan pembakuan ini agar dapat membandingkan dugaan koefisien regresi yang memiliki ragam yang berbeda dalam suatu model. Dimulai dengan sisaan $\boldsymbol{r} = \mathbf{y} - \bar{\mathbf{y}}, \beta_1, \cdots, \beta_p = 0$
2. Mencari peubah bebas x_j yang paling berkorelasi dengan $\boldsymbol{r}$.
3. Mengubah nilai β_j yang pada awalnya bernilai 0 bergerak menuju koefisien kuadrat terkecil $(x_j, \boldsymbol{r})$, sampai x_k yang lain memiliki korelasi yang kurang lebih sama dengan sisaan x_j.
4. Mengubah nilai β_j dan β_k bergerak ke arah koefisien gabungan kuadrat terkecil dengan sisaan sekarang dengan (x_j, x_k), sampai suatu peubah bebas yang lain misal xl memiliki korelasi yang cukup dengan sisaan (x_j, x_k).

   Modifikasi algoritma LAR untuk mendapatkan solusi LASSO adalah dengan memodifikasi langkah ke-4 menjadi:

Jika koefisien bukan nol mencapai nilai nol, keluarkan peubah tersebut dari gugus peubah aktif dan hitung kembali arah kuadrat terkecil bersama.

5. Mengulang langkah nomor 4 sampai semua p peubah bebas dimasukkan. Setelah $\min(N-1, p)$ langkah, solusi model penuh untuk kuadrat terkecil diperoleh.

LAR selalu mengambil p (banyaknya peubah) langkah untuk mendapatkan penduga kuadrat terkecil secara penuh, sedangkan modifikasi LAR untuk LASSO dapat memiliki lebih dari p langkah untuk mendapatkannya. Algoritma LASSO

dengan memodifikasi LAR merupakan cara yang efisien dalam komputasi solusi masalah LASSO, terutama ketika jumlah peubah bebas yang digunakan jauh lebih banyak daripada data amatannya (Hastie *et al*., 2008).

2.6 Validasi Silang (CV)

Salah satu metode pemilihan model terbaik bisa dilakukan menggunakan nilai validasi silang atau *Cross Validation* (CV). Menurut Hastie *et al*. (2008) validasi silang merupakan metode paling sederhana dan banyak dipakai secara luas untuk menduga galat prediksi. Idealnya, ketika data yang dimiliki memadai, akan dapat ditentukan suatu anak gugus data validasi dan digunakan untuk mengukur ketepatan model yang dimiliki. Namun seringkali data yang dimiliki terlalu sedikit sehingga tidak memungkinkan untuk dilakukan validasi secara langsung. Solusinya adalah melakukan validasi silang yang menggunakan sebagian data yang tersedia untuk membangun model dan sebagian data yang lain digunakan sebagai data pengujian model.

Validasi silang membagi data menjadi dua bagian, yaitu data training dan data test. Data training digunakan untuk membangun nilai $\widehat{\boldsymbol{\beta}}$ sedangkan data test digunakan untuk menguji kebaikan prediksi dari $\mathbf{X}\widehat{\boldsymbol{\beta}}$. Nilai validasi silang yang diperoleh merupakan penduga bagi sisaan prediksi (Izenman, 2008). James *et al*. (2013) mendefinisikan data latih sebagai data pemodelan dan data uji sebagai data validasi. Pengelompokkan data pemodelan dan data validasi dilakukan secara acak, dan setiap pengamatan memiliki peluang yang sama menjadi data validasi.

Salah satu metode tipe validasi silang adalah k*-fold*. Metode ini memiliki kelebihan ketika jumlah data amatan yang digunakan sedikit. Dalam validasi silang k-fold, semua observasi dipartisi secara acak ke dalam k subcontoh. Setiap sub-contoh digunakan sebagai data test dan sisanya digunakan sebagai data training. Proses validasi silang diulang sampai k kali, dan setiap satu sub-contoh digunakan hanya sekali dalam data test (Izenman, 2008).

Nilai kuadrat tengah sisaan hasil validasi silang (CV MSE) dihitung dengan menggunakan persamaan sebagai berikut:

$$\text{CV MSE} = \frac{1}{k}\sum_{k=1}^{K}\sum_{(x_i,y_i)\epsilon T}\left(y_i - \hat{y}_{-k}(x_i)\right)^2 \qquad (2.13)$$

dengan $y_i - \hat{y}_{-k}(x_i)$ adalah dugaan y untuk x_i pada saat *fold* ke- k tidak digunakan dalam menduga model, dan y_i adalah elemen peubah respon ke-i pada data uji T. Menurut Izenman (2008), validasi silang yang sebaiknya digunakan adalah validasi silang 5-*fold* atau 10-*fold* karena menghasilkan nilai sisaan validasi silang dengan bias tinggi tetapi ragam rendah.

2.7 Estimasi Parameter Regresi Dengan Metode LASSO

Regresi digunakan untuk menduga nilai suatu respon dari peubah bebas yang sudah diketahui atau diasumsikan ada hubungan dengan respon. Nilai dugaan parameter dari model regresi linier dapat ditaksir dengan metode kuadrat terkecil atau *ordinary least squares* (OLS). OLS pada prinsipnya adalah meminimumkan jumlah kuadrat *error* (residual). Model umum persamaan regresi dengan sampel *n* dan jumlah prediktor *p* dinyatakan sebagai berikut

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.14}$$

dimana

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{2.15}$$

Selanjutnya akan ditunjukan penduga untuk β dengan meminimumkan jumlah kuadrat galat

$$\sum_{i=1}^{n} e_i^2 \tag{2.16}$$

Diketahui bahwa

$$e_i^2 = \left(y_1 - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right)^2 \tag{2.17}$$

maka turunan parsial terhadap $\beta_0, \beta_1, \cdots, \beta_p$ dan menyamakan dengan nol sehingga diperoleh nilai estimasi model regresi linier adalah sebagai berikut:

$$\begin{aligned}
\frac{\partial s}{\partial \beta_0} &= -2\sum_{i-1}^{n}\left(y_1 - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right) &= 0\\
\frac{\partial s}{\partial \beta_1} &= -2\sum_{i-1}^{n}\left(y_1 - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right)x_{1i} &= 0\\
\frac{\partial s}{\partial \beta_2} &= -2\sum_{i-1}^{n}\left(y_1 - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}\right)x_{2i} &= 0\\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots\\
\frac{\partial s}{\partial \beta_p} &= -2\sum_{i-1}^{n}\left(y_1 - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right)x_{pi} &= 0
\end{aligned} \tag{2.18}$$

Dengan demikian diperoleh persamaan berikut:

$$\begin{aligned}
n\hat{\beta}_0 \;+\; \hat{\beta}_1\sum_{i-1}^{n}x_{1i} + \;\; \hat{\beta}_2\sum_{i-1}^{n}x_{2i} + \cdots + \hat{\beta}_p\sum_{i-1}^{n}x_{pi} \;\; &= \sum_{i-1}^{n}y_i\\
\hat{\beta}_0\sum_{i-1}^{n}x_{1i} + \hat{\beta}_1\sum_{i-1}^{n}x_{1i}^2 + \hat{\beta}_2\sum_{i-1}^{n}x_{1i}x_{2i} + \cdots + \hat{\beta}_p\sum_{i-1}^{n}x_{1i}x_{pi} &= \sum_{i-1}^{n}x_{1i}y_i\\
\hat{\beta}_0\sum_{i-1}^{n}x_{2i} + \hat{\beta}_1\sum_{i-1}^{n}x_{1i}x_{2i} + \hat{\beta}_2\sum_{i-1}^{n}x_{2i}^2 + \cdots + \hat{\beta}_p\sum_{i-1}^{n}x_{2i}x_{pi} &= \sum_{i-1}^{n}x_{2i}y_i\\
\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &\vdots\\
\hat{\beta}_0\sum_{i-1}^{n}x_{pi} + \hat{\beta}_1\sum_{i-1}^{n}x_{1i}x_{pi} + \hat{\beta}_2\sum_{i-1}^{n}x_{2i}x_{pi} + \cdots + \hat{\beta}_p\sum_{i-1}^{n}x_{pi}^2 &= \sum_{i-1}^{n}x_{pi}y_i
\end{aligned} \tag{2.19}$$

Berdasarkan persamaan (2.39) dapat dibuat dalam persamaan berikut:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{2.20}$$

untuk menyelesaikan persamaan (2.40), maka dikalikan kedua sisinya dengan invers dari $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$**.** Maka diperoleh penduga dari $\boldsymbol{\beta}$ adalah sebagai berikut:

$$\begin{aligned}
\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\\
\mathbf{I}\boldsymbol{\beta} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\\
\boldsymbol{\beta} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

Sehingga solusi estimasi parameter regresi dengan OLS adalah

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{2.21}$$

Suatu matriks khusus di mana invers-nya dapat diperoleh dengan mentransposkan disebut matriks orthogonal. Pada kasus matriks ortonormal dimana $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$, solusi estimasi parameter regresi LASSO memiliki solusi eksplisit dan dapat dengan mudah dibandingkan dengan solusi OLS. Jika diasumsikan tidak ada intersep persamaan dapat ditulis sebagai berikut

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\\ &= (\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\\ &= \mathbf{I}\mathbf{X}^{\mathrm{T}}\mathbf{y}\end{aligned}$$

Maka solusi OLS untuk kasus matriks ortonormal adalah

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{2.22}$$

Melalui bentuk Lagrangian, persamaan (2.12) dapat ditulis dalam persamaan berikut

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1) \tag{2.23}$$

Kemudian persamaan (2.49) dapat diuraikan sebagai berikut

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\left(\sqrt{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2}\right)^2 + \lambda|\boldsymbol{\beta}|\right)\\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}|)\\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \lambda|\boldsymbol{\beta}|\right)\\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{I}\boldsymbol{\beta} + \lambda|\boldsymbol{\beta}|\right)\end{aligned}$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{2}\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} + \lambda|\boldsymbol{\beta}|\right) \tag{2.24}$$

Karena $\mathbf{y}^{\mathrm{T}}\mathbf{y}$ tidak memiliki parameter $\boldsymbol{\beta}$, sehingga dapat dibuang dan diperoleh persamaan sebagai berikut

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(-2\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} + \lambda|\boldsymbol{\beta}|\right)$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\sum_{j=1}^{d}\left(-2\boldsymbol{\beta}_j\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}_j} + \boldsymbol{\beta}_j^2 + \lambda\left|\boldsymbol{\beta}_j\right|\right) \tag{2.25}$$

Di mana X_j adalah kolom ke- j dari kolom matriks X. Dan karena terpisah maka dapat dipecahkan masalah optimasi secara terpisah untuk setiap peubah dalam penjumlahan. Kemudian untuk setiap $|\boldsymbol{\beta}_j| = \boldsymbol{\beta}$, sehingga perlu diselesaikan persamaan berikut

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(-2\boldsymbol{\beta}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^2 + \lambda\boldsymbol{\beta}\right) \tag{2.26}$$

Meskipun persamaan tersebut tidak dapat diturunkan, masalah ini dapat dipecahkan menjadi tiga kasus. Dalam kasus pertama, untuk $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} > 0$, maka solusi $\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} > 0$ sehingga perlu dicari solusi persamaan berikut

$$\underset{\boldsymbol{\beta}>0}{\operatorname{argmin}}\left(-2\boldsymbol{\beta}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^2 + \lambda\boldsymbol{\beta}\right) \tag{2.27}$$

Langkah berikut dicari turunan terhadap $\boldsymbol{\beta}$ dan menyamakan dengan nol sehingga diperoleh

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(-2\boldsymbol{\beta}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^2 + \lambda\boldsymbol{\beta}\right) = 0$$

$$-2\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + 2\boldsymbol{\beta} + \lambda = 0$$

$$2\boldsymbol{\beta} = 2\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \lambda$$

$$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \frac{\lambda}{2} \tag{2.29}$$

Dalam kasus kedua, untuk $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} < 0$, maka solusi $\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} < 0$ sehingga perlu dicari solusi persamaan berikut

$$\underset{\boldsymbol{\beta}<0}{\operatorname{argmin}}\left(-2\boldsymbol{\beta}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^2 - \lambda\boldsymbol{\beta}\right) \tag{2.30}$$

Kemudian dicari turunan terhadap β dan menyamakan dengan nol sehingga diperoleh

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(-2\boldsymbol{\beta}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \boldsymbol{\beta}^2 - \lambda\boldsymbol{\beta}\right) = 0$$

$$-2\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + 2\boldsymbol{\beta} - \lambda = 0$$

$$2\boldsymbol{\beta} = 2\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \lambda$$

$$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} + \frac{\lambda}{2} \tag{2.31}$$

Dalam kasus ketiga untuk $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = 0$, maka solusi $\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}} = 0$.

Jadi, ketika kolom matriks X ortogonal, solusi untuk estimasi parameter regresi LASSO adalah:

$$\widehat{\boldsymbol{\beta}}_{\text{LASSO}} = \begin{cases} \widehat{\boldsymbol{\beta}}_{\text{OLS}} - \dfrac{\lambda}{2}, & \text{untuk } \widehat{\boldsymbol{\beta}}_{\text{OLS}} > 0 \\ \widehat{\boldsymbol{\beta}}_{\text{OLS}} + \dfrac{\lambda}{2}, & \text{untuk } \widehat{\boldsymbol{\beta}}_{\text{OLS}} < 0 \\ 0, & \text{lainnya} \end{cases} \tag{2.32}$$

Secara umum regresi LASSO tidak memiliki solusi secara eksplisit dalam menentukan koefisien taksirannya karena pada fungsi kendala regresi LASSO berbentuk fungsi mutlak yang tidak dapat diturunkan karena turunan kirinya berbeda dengan turunan kanannya, dengan demikian dibutuhkan pemrograman komputasi untuk menyelesaikannya. Algoritma LARS merupakan algoritma yang sangat efektif dalam membantu menyelesaikan solusi regresi LASSO secara komputasi.

Menurut Efron *et al*. (2004), LARS melakukan estimasi $\widehat{\boldsymbol{\mu}} = \mathbf{X}^{*}\widehat{\boldsymbol{\beta}}$, dengan langkah-langkah yang berurutan, dan di setiap langkah akan menambah satu kovariat ke dalam model. LARS merupakan modifikasi LAR untuk LASSO. Nilai $\widehat{\boldsymbol{\mu}}$ diperoleh dari teknik iterasi dengan nilai awal $\widehat{\boldsymbol{\mu}}^{(0)} = 0$. Langkah-langkah estimasi koefisien LASSO dengan algoritma LARS sebagai berikut:

Mencari vektor yang sebanding dengan vektor korelasi antara peubah prediktor dan galat dari setiap peubah bebas

$$\hat{\boldsymbol{c}} = \mathbf{X}^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) \tag{2.33}$$

Menentukan korelasi saat mutlak terbesar

$$\hat{C} = \max\{|\hat{c}_j|\} \tag{2.34}$$

maka diperoleh $s_j = sign\{\hat{c}_j\}$ untuk $j \epsilon \mathrm{A}$

Menentukan X_A, Himpunan A merupakan himpunan indeks aktif dari peubah bebas {1, 2, 3, ..., m}. Himpunan indeks aktif A ditentukan berdasarkan nilai korelasi mutlak terbesar. Didefinisikan matriks:

$$\mathrm{X}_A = \left(\dots s_j \mathrm{X}_j^{*} \dots\right)_{j \in A} \tag{2.35}$$

dengan X^{*} merupakan matrix X yang dinormalisasi dan tanda s_j bernilai ±1, maka

$$\mathrm{G_A} = \mathrm{X}_A^{\mathrm{T}}\mathrm{X}_A \text{ dan } \mathrm{P}_A = \left(\mathbf{1}^{\mathrm{T}}\mathrm{G}_A^{-1}\mathbf{1}_A\right)^{-1/2} \tag{2.36}$$

Menghitung nilai vektor *equiangular*, vektor *equiangular* adalah suatu vektor yang membagi sudut dari kolom-kolom menjadi sama besar dengan besar sudutnya kurang dari 90^o. Nilai vektor *equiangular* dicari menggunakan rumus sebagai berikut:

$$\boldsymbol{u}_A = \mathrm{X}_A\omega_A \text{ dengan } \omega_A = \mathrm{P}_A\mathrm{G}_A^{-1}\mathbf{1}_A \tag{2.37}$$

Menghitung vektor *inner product*:

$$\boldsymbol{a} \equiv \mathbf{X}^{\mathrm{T}}\boldsymbol{u}_A \tag{2.38}$$

Menghitung $\widehat{\boldsymbol{\mu}}_A$ dengan $\widehat{\boldsymbol{\mu}}_{A+} = \widehat{\boldsymbol{\mu}}_A + \hat{\gamma}\boldsymbol{u}_A$

$$\hat{\gamma} = \min_{j\epsilon A^{\mathrm{c}}}^{+}\left\{\frac{\hat{C}-\hat{c}_j}{\mathrm{P}_A - a_j}, \frac{\hat{C}-\hat{c}_j}{\mathrm{P}_A - a_j}\right\} \tag{2.39}$$

$\min_{j\epsilon A^{\mathrm{c}}}^{+}$ menunjukkan bahwa yang dipilih adalah nilai minimum positif dari j yang bukan merupakan himpunan A. Pada tahap akhir dalam memperoleh nilai $\hat{\gamma}$ menggunakan rumus

$$\hat{\gamma}_m = \frac{\hat{C}_m}{A_m} \tag{2.40}$$

Tanda dari koordinat bukan nol $\hat{\beta}_j$ sama dengan tanda $\hat{c}_j$, dengan rumus (Efron *et al*., 2004):

$$\mathrm{sign}\left(\hat{\beta}_j\right) = \mathrm{sign}\left(\hat{c}_j\right) = s_j \tag{2.41}$$

didefinisikan bahwa $\hat{d} = s_j\boldsymbol{\omega}_{Aj}$ sehingga $\widehat{\boldsymbol{\mu}}$ menjadi:

$$\boldsymbol{\mu}(\gamma) = \mathbf{X}\boldsymbol{\beta}(\gamma) \text{ dengan } \beta_j(\gamma) = \hat{\beta}_j + \gamma\hat{d}_j \tag{2.42}$$

$\beta_j(\gamma)$ akan berubah tanda pada saat

$$\gamma_j = \frac{-\hat{\beta}_j}{\hat{d}_j} \tag{2.43}$$

perubahan pertama yang terjadi pada

$$\tilde{\gamma} = \min_{\gamma_j>0}\{\gamma_j\} \tag{2.44}$$

Jika $\tilde{\gamma} < \hat{\gamma}$ maka $\beta_j(\gamma)$ bukan merupakan solusi LASSO karena pada $\beta_j(\gamma)$ telah berubah tanda melainkan pada c_j tidak berubah tanda dan proses LARS berhenti pada $\gamma = \tilde{\gamma}$ dan menghapus j dari perhitungan vektor *equiangular* berikutnya dan peubah j dimasukkan kembali pada tahap perhitungan LARS selanjutnya, maka:

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{A+}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{A}} + \hat{\gamma}\boldsymbol{u}_{\boldsymbol{A}} \text{ dan } A_+ = A - \{j\} \tag{2.45}$$

2.8 Metode LASSO LTS

Dalam analisis data terapan, ada peningkatan ketersediaan kumpulan data yang berisi sejumlah besar peubah. Model linier yang menyertakan lengkap peubah respon seringkali memiliki kinerja prediksi yang buruk karena cenderung memiliki varians yang besar. Selain itu, model besar pada umumnya sulit untuk diinterpretasikan. Dalam banyak kasus, jumlah peubah bahkan lebih besar dari jumlah pengamatan. Metode tradisional seperti *least squares* kemudian tidak dapat lagi diterapkan karena kekurangan rank dari matriks desain.

Untuk meningkatkan akurasi prediksi dan sebagai solusi untuk masalah komputasi dengan data berdimensi tinggi, istilah penalti pada koefisien regresi dapat ditambahkan ke fungsi objektif. Pendekatan ini mengecilkan koefisien dan mengurangi varians dengan harga bias yang meningkat. Tibshirani (1996) memperkenalkan *Least Absolute Shrinkage and Selection Operator* (LASSO), di mana fungsi penalti adalah norma L_1. Misalkan $\mathbf{y} = (y_1, \cdots, y_n)'$adalah respon dan $\mathbf{X} = \left(x_{ij}\right)_{1 \le i \le n, 1 \le j \le p}$ matriks peubah bebas, di mana n menyatakan banyaknya pengamatan dan p menyatakan banyaknya peubah. Selain itu, misalkan $x_1, \cdots, x_n$ adalah pengamatan dimensi-p, yaitu baris $\mathbf{X}$. Kita asumsikan model regresi standar

$$\mathbf{y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + e_i \tag{2.46}$$

di mana parameter regresi adalah $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_p\right)'$, dan *error* e_i memiliki nilai harapan nol. Dengan parameter penalti λ , penduga LASSO dari $\boldsymbol{\beta}$ adalah

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + n\lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{2.47}$$

LASSO sering digunakan dalam praktik karena penalti L_1 memungkinkan untuk mengecilkan beberapa koefisien hingga tepat nol, yaitu, untuk menghasilkan

estimasi model *sparse* (model di mana hanya sebagian kecil parameter yang tidak nol) yang sangat dapat diinterpretasikan. Selain itu, algoritma cepat untuk menghitung LASSO tersedia melalui kerangka *Least Angle Regression* (LARS). Namun, LASSO tidak *robust* untuk *outlier*. Oleh karena itu, diperlukan alternatif *robust*.

Penduga model s*parse* adalah topik yang sangat penting dalam analisis data modern karena semakin tersedianya kumpulan data dengan sejumlah besar peubah. Metode *Least Trimmed Squares* (LTS) digunakan untuk menemukan himpunan bagian (*subset*) dari h pengamatan yang kuadrat terkecilnya menghasilkan jumlah kuadrat galat terkecil. Ukuran *subset* h dapat dilihat sebagai tebakan awal dari jumlah observasi yang baik dalam data. Sementara LTS bersifat sangat *robust*, akan tetapi tidak menghasilkan estimasi model *sparse*. Selanjutnya, jika $h < p$, estimator LTS tidak dapat dihitung. Versi *sparse* dari LTS diperoleh dengan menambahkan penalti L_1 dengan parameter penalti λ ke persamaan (2.3), sehingga penduga untuk LASSO LTS adalah sebagai berikut,

$$\widehat{\boldsymbol{\beta}}_{\text{LASSO}-\text{LTS}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} e_{(i)}^2 + h\lambda \sum_{j=1}^{p} |\beta_j| \tag{2.48}$$

LASSO LTS memiliki *breakdown point* yang tinggi dan tahan terhadap *outlier* regresi berganda, termasuk *leverage point*. Selain bersifat sangat robust, dan mirip dengan penduga LASSO, LASSO LTS memiliki kelebihan

1. Meningkatkan kinerja prediksi melalui pengurangan varians jika ukuran sampel relatif kecil terhadap dimensi
2. Memastikan interpretasi yang lebih tinggi karena pemilihan model simultan
3. Menghindari masalah komputasi metode regresi *robust* tradisional dalam kasus data berdimensi tinggi.

2.8.1 *Breakdown Point*

Ukuran yang paling populer untuk kekekaran (*robustness*) estimator adalah *replacement finite-sample breakdown* (FBP). Misalkan $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$ menyatakan sampel. Untuk estimator regresi $\hat{\beta}$, *breakdown point* didefinisikan sebagai

$$e^*(\hat{\beta}; \mathbf{Z}) = \min\left\{\frac{m}{n} : \sup \left\|\hat{\beta}(\bar{\mathbf{Z}})\right\|_2 = \infty\right\} \tag{2.49}$$

di mana $\bar{\mathbf{Z}}$ adalah data rusak yang diperoleh dari $\mathbf{Z}$ dengan mengganti m dari n titik data asli dengan nilai *arbitrary*.

Misalkan $\rho(x)$ adalah *loss function* cembung dan simetris dengan $\rho(0) = 0$ dan $\rho(x) > 0$ untuk $x \neq 0$, dan didefinisikan $\boldsymbol{\rho}(\boldsymbol{x}) = \big(\rho(x_1), \cdots, \rho(x_n)\big)^{\mathrm{T}}$. Dengan ukuran himpunan bagian $h \leq n$, maka estimator regresi sebagai berikut:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} \boldsymbol{\rho}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})_{i:n} + h\lambda \sum_{j=1}^{p} |\beta_j| \tag{2.50}$$

di mana $\big(\boldsymbol{\rho}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\big)_{i:n} \leq \cdots \big(\boldsymbol{\rho}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\big)_{n:n}$ adalah *regression loss* yang diurutkan. Kemudian *breakdown point* penduga $\hat{\beta}$ diberikan sebagai berikut

$$e^*\big(\hat{\beta}; \mathbf{Z}\big) = \frac{n - h + 1}{n} \tag{2.51}$$

Breakdown point sama untuk setiap *loss function* p yang memenuhi asumsi. Secara khusus, *breakdown point* untuk estimator LASSO LTS $\hat{\beta}_{\mathrm{LASSO-LTS}}$ dengan ukuran *subset* $h \leq n$, di mana $\rho(x) = x^2$, adalah $(n - h + 1)/n$. Semakin kecil nilai h, semakin tinggi *breakdown point*-nya. Dengan mengambil h yang cukup kecil, bahkan dimungkinkan untuk memiliki *breakdown point* yang lebih besar dari 50%. Namun, sementara ini secara matematis mungkin, tidak disarankan untuk menggunakan $h < n/2$ karena statistik yang kekar (*robust*) bertujuan untuk model yang sesuai dengan sebagian besar data. Sebagai gantinya, disarankan untuk mengambil nilai h yang sama dengan sebagian kecil dari ukuran sampel, dengan $\alpha = 0.75$, sehingga perkiraan akhir didasarkan pada jumlah pengamatan yang cukup besar.

2.8.2 Algoritma LASSO LTS

Untuk parameter penalti tetap λ berdasarkan persamaan (2.16), maka ditentukan *objective function*

$$Q(H, \beta) = \sum_{i \epsilon H} (y_i - \mathbf{X}_i'\boldsymbol{\beta})^2 + h\lambda \sum_{j=1}^{p} |\beta_j| \tag{2.52}$$

yang merupakan jumlah kuadrat galat L_1 terpenalti berdasarkan subsample $H \subseteq \{1, \cdots, n\}$ dengan $|H| = h$. Dengan

$$\hat{\beta}_H = \underset{\beta}{\operatorname{argmin}}(H, \beta) \tag{2.53}$$

penduga LASSO LTS diberikan oleh $\hat{\beta}_{H_{opt}}$, dimana

$$H_{opt} = \underset{H \subseteq \{1,\cdots,n\}:|H|=h}{\operatorname{argmin}} \left(H, \hat{\beta}_H\right) \tag{2.54}$$

Oleh karena itu *Sparse* LTS digunakan untuk menemukan himpunan bagian dari $h \leq n$ pengamatan yang *Least Absolute Shrinkage and Selection Operator* (LASSO) *fit*-nya menghasilkan jumlah kuadrat galat terkecil. Untuk menemukan *subset* optimal ini, digunakan analog dari algoritma FAST-LTS yang dikembangkan oleh Rousseeuw dan Van Driessen (2006).

Algoritma ini didasarkan pada *concentration steps* atau C-*steps*. C-*steps* pada iterasi k terdiri dari menghitung solusi LASSO berdasarkan *subset* H_k, dengan $|H_k| = h$, dan membuat *subset* berikutnya H_{k+1} dari pengamatan yang sesuai dengan h kuadrat galat terkecil. Misalkan H_k menyatakan subsampel tertentu yang diturunkan pada iterasi k dan misalkan $\hat{\beta}_{H_k}$ adalah koefisien dari LASSO *fit* yang sesuai. Setelah menghitung kuadrat galat $e_k^2 = \left(e_{k,1}^2, \cdots, e_{k,n}^2\right)'$ dengan $e_{k,i}^2 = \left(y_i - \mathbf{X}_i' \hat{\beta}_{H_k}\right)^{\mathbf{2}}$, subsampel H_{k+1} untuk iterasi $k+1$ didefinisikan sebagai himpunan indeks yang sesuai dengan h kuadrat galat terkecil. Dalam istilah matematika, ini dapat ditulis sebagai

$$H_{k+1} = \left\{ i \in \{1, \cdots, n\} : e_{k,i}^2 \in \left\{ \left(e_k^2\right)_{j:n} : j = 1, \cdots, h \right\} \right\} \tag{2.56}$$

Dimana $\left(e_k^2\right)_{1:n} \leq \cdots \leq \left(e_k^2\right)_{n:n}$ menyatakan kuadrat galat yang diurutkan. Misalkan $\hat{\beta}_{H_{k+1}}$ menyatakan koefisien LASSO *fit* berdasarkan H_{k+1}. Maka

$$Q\left(H_{k+1}, \hat{\beta}_{H_{k+1}}\right) \leq Q\left(H_{k+1}, \hat{\beta}_{H_k}\right) \leq Q\left(H_k, \hat{\beta}_{H_k}\right) \tag{2.57}$$

dimana pertidaksamaan pertama mengikuti definisi $\hat{\beta}_{H_{k+1}}$, dan pertidaksamaan kedua dari definisi H_k. Dari persamaan (2.58) didapat bahwa C-*steps* menghasilkan penurunan fungsi objektif LTS yang *sparse*, dan bahwa urutan dari C-*steps* menghasilkan konvergensi ke minimum lokal dalam sejumlah langkah yang terbatas. Untuk meningkatkan peluang mencapai minimum global, jumlah s yang cukup besar dari subsampel awal H_0 harus digunakan, masing-masing digunakan sebagai titik awal untuk urutan C-*steps*. Daripada memilih h titik data secara acak, setiap *subset* awal H_0 dengan ukuran h disusun dari sebuah *elemental subset* berukuran 3. Ambil

tiga pengamatan dari data secara acak, misalkan, x_{i_1}, x_{i_2}, dan x_{i_3}. Maka LASSO *fit* untuk *elemental subset* ini sebagai berikut,

$$\widehat{\boldsymbol{\beta}}_{\{i_1,i_2,i_3\}} = \underset{\beta}{\operatorname{argmin}}\, Q(\{i_1,i_2,i_3\},\beta) \tag{2.58}$$

dan *subset* awal H_0 kemudian diberikan oleh indeks h pengamatan dengan kuadrat galat terkecil berdasarkan pada persamaan (2.58). Algoritma FAST-LTS *nonsparse* menggunakan himpunan bagian elemen berukuran p, karena setiap regresi OLS memerlukan setidaknya observasi sebanyak dimensi p. Hal ini akan membuat algoritma tidak berlaku jika $p > n$. Untungnya LASSO sudah didefinisikan dengan benar untuk sampel ukuran 3, bahkan untuk nilai p yang besar. Selain itu, dari sudut pandang kekekaran (*robustness*), hanya menggunakan tiga pengamatan yang optimal, karena memastikan probabilitas tertinggi untuk tidak menyertakan *outlier* dalam himpunan *elemental*. Penting untuk dicatat bahwa *elemental subset* berukuran 3 hanya digunakan untuk membangun himpunan bagian awal berukuran h untuk algoritma C-*steps*. Semua C-*steps* dilakukan pada himpunan bagian dari ukuran h.

2.8.3 *Reweighted LASSO LTS Estimator*

Misalkan α menyatakan proporsi pengamatan dari sampel penuh yang akan dipertahankan di setiap subsampel, dan $h = (n+1)\alpha$. Kemudian $(1-\alpha)$ dapat diartikan sebagai tebakan awal dari proporsi *outlier* dalam data. Tebakan awal ini biasanya agak konservatif untuk memastikan bahwa *outlier* tidak mempengaruhi hasil, dan karena itu dapat mengakibatkan hilangnya efisiensi statistik. Untuk meningkatkan efisiensi, langkah *reweighted* yang menurunkan bobot *outlier* yang terdeteksi oleh estimator LASSO LTS dapat dilakukan.

Di bawah model *error* normal, pengamatan dengan galat terstandarisasi yang lebih besar dari kuantil tertentu dari distribusi normal baku dapat dinyatakan sebagai *outlier*. Karena estimasi LASSO LTS seperti LASSO adalah bias, kita perlu memusatkan galatnya. Estimasi natural untuk pusat galat adalah

$$\hat{\mu}_{raw} = \frac{1}{h}\sum_{i\epsilon H_{opt}} e_i \tag{2.59}$$

di mana $e_i = y_i - \mathrm{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{LASSO-LTS}$ dan H_{opt} adalah himpunan bagian optimal dari (2.54). Kemudian estimasi skala galat yang terkait dengan penduga *raw* LASSO LTS diberikan oleh

$$\hat{\sigma}_{raw} = k_\alpha \sqrt{\frac{1}{h}\sum_{i\epsilon H_{opt}} (e_c^2)_{i:n}} \tag{2.60}$$

dengan kuadrat galat terpusat $e_c^2 = \left((e_1 - \hat{\mu}_{raw})^2, \cdots, (e_n - \hat{\mu}_{raw})\right)^{\mathrm{T}}$, dan

$$k_\alpha = \left(\frac{1}{\alpha}\int_{-\Phi^{-1}((\alpha+1)/2)}^{\Phi^{-1}((\alpha+1)/2)} u^2 d\Phi(u)\right)^{-1/2} \tag{2.61}$$

k_α merupakan sebuah faktor untuk memastikan bahwa $\hat{\sigma}_{raw}$ adalah estimasi yang konsisten dari standar deviasi pada model normal. Formulasi ini memungkinkan untuk menentukan bobot biner

$$w_i \begin{cases} 1, & \text{jika } |(e_i - \hat{\mu}_{raw})/\hat{\sigma}_{raw}| \le \Phi^{-1}(1-\delta), \\ 0, & \text{jika } |(e_i - \hat{\mu}_{raw})/\hat{\sigma}_{raw}| \le \Phi^{-1}(1-\delta), \end{cases} \quad i = 1, \cdots, n \tag{2.62}$$

Reweighted estimator LASSO LTS diberikan dengan LASSO *fit* terbobot

$$\hat{\beta}_{reweighted} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(y_i - x_i^{\mathrm{T}}\beta\right)^2 + \lambda n_w \sum_{j=1}^{p} |\beta_j| \tag{2.63}$$

dengan $n_w = \sum_{i=1}^{n} w_i$ jumlah bobot. Dengan pilihan bobot yang diberikan pada (2.62), *Reweighted* penduga LASSO LTS adalah LASSO *fit* berdasarkan pengamatan tidak ditandai sebagai pencilan. Tentu saja, skema pembobotan lain dapat dipertimbangkan. Menggunakan estimasi pusat galat

$$\hat{\mu}_{reweighted} = \frac{1}{n_w}\sum_{i=1}^{n} w_i \left(y_i - x_i^{\mathrm{T}}\hat{\beta}_{reweighted}\right) \tag{2.64}$$

estimasi skala galat dari *Reweighted* penduga LASSO LTS diberikan oleh

$$\hat{\sigma}_{reweighted} = k_{\alpha_w} \sqrt{\frac{1}{n_w}\sum_{i=1}^{n} w_i \left(y_i - x_i^{\mathrm{T}}\hat{\beta}_{reweighted}\right)} \tag{2.65}$$

di mana k_{α_w} adalah faktor konsistensi dari persamaan (2.28) dengan $\alpha_w = n_w/n$.

Perhatikan bahwa langkah pembobotan ulang ini secara konseptual berbeda dari *adaptive* LASSO oleh Zou (2006). Sementara *adaptive* LASSO mendapatkan penalti secara individu pada prediktor dari koefisien penduga awal, *Reweighted*

penduga LASSO LTS bertujuan untuk memasukkan semua pengamatan *nonoutlying* ke dalam pencocokan model.

2.8.4 Pemilihan Parameter Penalti

Dalam analisis data praktis, nilai parameter penalti λ yang sesuai tidak diketahui sebelumnya. Disarankan untuk memilih λ dengan mengoptimalkan *Bayes Information Criterion* (BIC), atau perkiraan kinerja prediksi melalui *cross validation* (CV). Dalam penelitian ini digunakan BIC karena membutuhkan lebih sedikit upaya komputasi. BIC dari model yang diestimasi dengan parameter penyusutan λ diberikan oleh

$$\mathrm{BIC}(\lambda) = \log(\hat{\sigma}) + df(\lambda)\frac{\log(n)}{n} \tag{2.66}$$

di mana $\hat{\sigma}$ menunjukkan perkiraan skala galat yang sesuai, (2.60) atau (2.65), dan $df(\lambda)$ adalah derajat kebebasan model. Derajat kebebasan diberikan oleh jumlah parameter estimasi bukan nol di $\widehat{\boldsymbol{\beta}}$ (Zou *et al*., 2007).

Sebagai alternatif BIC, validasi silang dapat digunakan. Untuk mencegah *outlier* mempengaruhi pemilihan λ, prediksi *robust loss function* harus digunakan. Pemilihan natural adalah *root trimmed mean squared prediction error* (RTMSPE) dengan proporsi pemangkasan yang sama seperti untuk menghitung LASSO LTS. Dalam validasi silang k*-fold*, data dibagi secara acak dalam k blok dengan ukuran yang kira-kira sama. Setiap blok dikeluarkan sekali untuk kecocokan model, dan blok yang dikeluarkan digunakan sebagai data uji. Dengan cara ini, dan untuk nilai λ tertentu, sebuah prediksi diperoleh untuk setiap pengamatan dalam sampel. Dinyatakan vektor kuadrat *error* dugaan $\mathbf{e}^2 = (e_1^2, \cdots, e_n^2)$. Maka

$$\mathrm{RTMSPE}(\lambda) = \sqrt{\frac{1}{h}\sum_{i \epsilon H_{opt}} (\mathbf{e}^2)_{i:n}} \tag{2.67}$$

Untuk mengurangi variabilitas, RTMSE dirata-ratakan pada lebih dari 500 pemisahan acak data yang berbeda.

Parameter λ yang dipilih kemudian meminimalkan $\mathrm{BIC}(\lambda)$ atau $\mathrm{RTMSPE}(\lambda)$ pada *grid* nilai di interval $[0, \hat{\lambda}_0]$. Diambil *grid* dengan langkah ukuran $0.025\hat{\lambda}_0$, di mana $\hat{\lambda}_0$ adalah perkiraan parameter penyusutan λ_0 yang akan menyusutkan semua

parameter menjadi nol. Jika $p > n$, 0 tentu saja dikeluarkan dari *grid*. Untuk solusi LASSO diperoleh

$$\hat{\lambda}_0 = \frac{2}{n} \max_{j \in \{1,,\cdots p\}} \mathrm{Cor}\left(\mathbf{y}, \mathbf{x}_j\right) \tag{2.68}$$

$\mathrm{Cor}(\mathbf{y}, \mathbf{x}_j)$ merupakan korelasi Pearson antara $\mathbf{y}$ dan kolom ke- j dari matriks $\mathbf{X}$. Untuk LASSO LTS, diperlukan estimasi *robust* $\hat{\lambda}_0$. Disarankan untuk mengganti korelasi Pearson pada persamaan (2.68) dengan korelasi kuat berdasarkan *bivariate winorization* data (Khan *et al*., 2007).

2.9 Tuberkulosis

Di seluruh dunia, sekitar 10 juta orang jatuh sakit tuberkulosis (TB) setiap tahun. TB adalah salah satu dari 10 penyebab kematian teratas, dan penyebab utama dari agen infeksi tunggal (*Mycobacterium tuberculosis*), peringkat di atas HIV/AIDS. Penyakit ini dapat menyerang siapa saja di mana saja, tetapi kebanyakan orang yang mengembangkan TB (sekitar 90%) adalah orang dewasa, rasio pria dan wanita adalah 2:1, dan tingkat kasus di tingkat nasional bervariasi dari kurang dari 50 hingga lebih dari 5000 per 1 juta penduduk. per tahun. Hampir 90% kasus setiap tahun berada di 30 negara dengan beban TB tinggi. Secara global, diperkirakan 1.7 miliar orang terinfeksi *Mycobacterium tuberculosis* dan dengan demikian berisiko terkena penyakit tersebut (WHO, 2018).

Dengan diagnosis dan pengobatan yang tepat waktu serta antibiotik, kebanyakan orang yang terkena TB dapat disembuhkan dan penularan selanjutnya dibatasi. Jumlah kasus yang terjadi setiap tahun (dan dengan demikian jumlah kematian terkait TB) juga dapat diturunkan dengan mengurangi prevalensi faktor risiko terkait kesehatan untuk TB, misalnya merokok, diabetes dan infeksi HIV, memberikan pengobatan pencegahan kepada orang dengan infeksi TB laten, dan tindakan pada determinan infeksi dan penyakit TB yang lebih luas misalnya kemiskinan, kualitas perumahan dan kekurangan gizi (WHO, 2018).

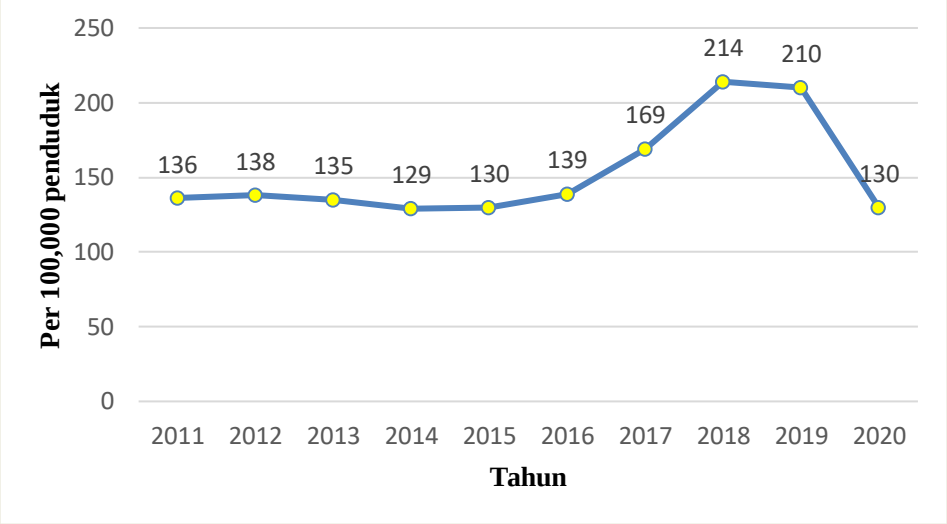
TB saat ini masih merupakan masalah kesehatan masyarakat baik di Indonesia maupun internasional sehingga menjadi salah satu tujuan pembangunan Kesehatan berkelanjutan (SDGs). TB merupakan salah satu dari 10 penyebab utama kematian di seluruh dunia (WHO, 2018).

Indonesia berada pada peringkat ke-3 dengan penderita TB tertinggi di Dunia setelah India dan China. Meskipun terjadi penurunan kasus baru TB, tetapi tidak cukup cepat untuk mencapai target Strategi END TB tahun 2020, yaitu pengurangan kasus TB sebesar 20% antara tahun 2015 – 2020. Pada tahun 2015 – 2019 penurunan kumulatif kasus TB hanya sebesar 9% (WHO, 2018).

Begitu juga dengan kematian akibat TB, jumlah kematian pada tahun 2018 sebesar 1,2 juta. Secara global kematian akibat TB per tahun menurun secara global, tetapi tidak mencapai target Strategi END TB tahun 2020 sebesar 35% antara tahun 2015 – 2020. Jumlah kematian kumulatif antara tahun 2015 – 2019 sebesar 14%, yaitu kurang dari setengah dari target yang ditentukan (WHO, 2018). Beban penyakit yang disebabkan oleh TB dapat diukur salah satunya dengan angka notifikasi semua kasus tuberkulosis atau *Case Notification Rate* (CNR).

2.9.1 *Case Notification Rate* (CNR)

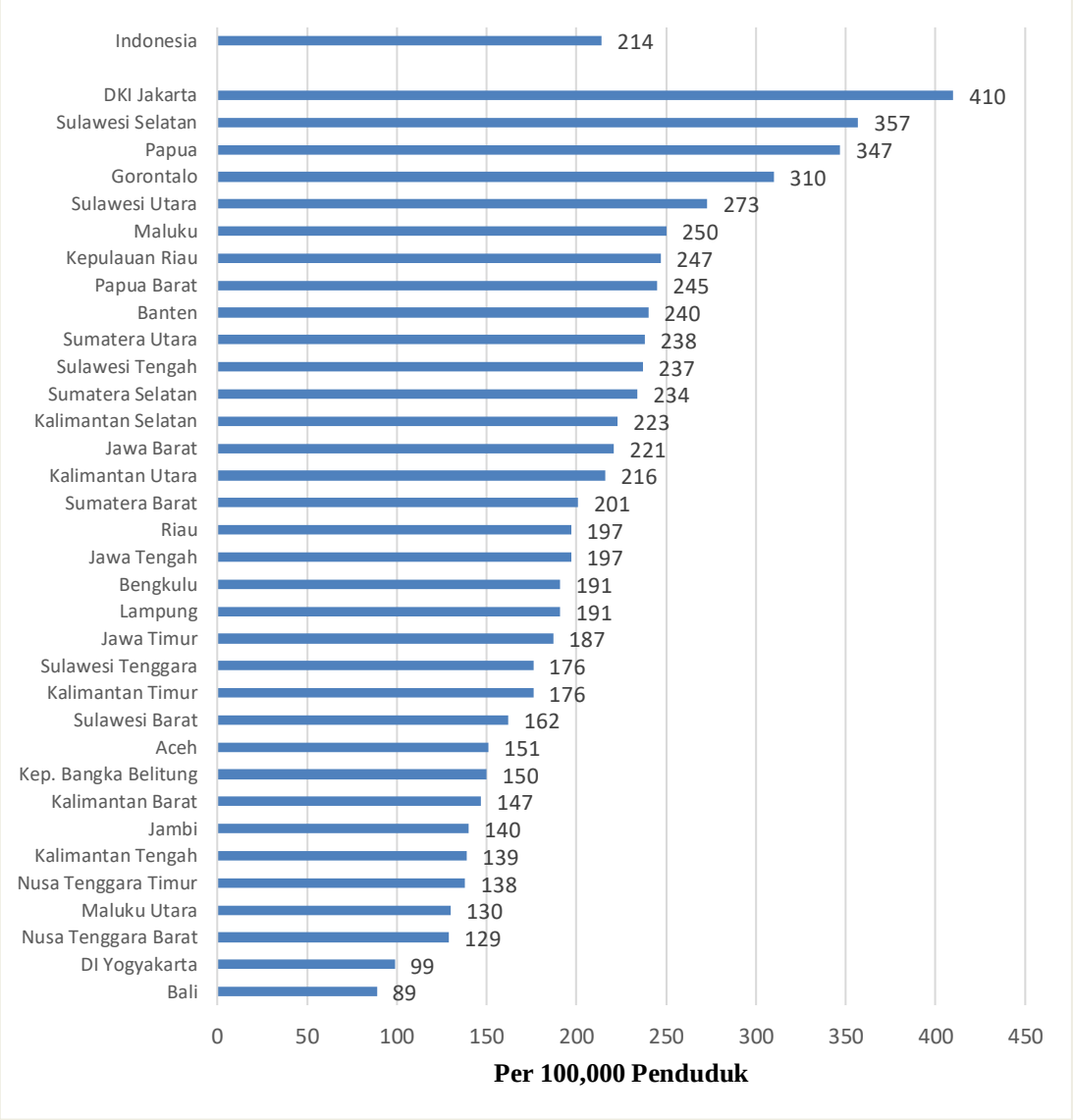
Angka notifikasi Semua Kasus Tuberkulosis atau *Case Notification Rate* (CNR) adalah jumlah semua kasus TB yang diobati dan dilaporkan di antara 100,000 penduduk yang ada di suatu wilayah tertentu. Angka ini apabila dikumpulkan serial, akan menggambarkan kecenderungan (*tren*) meningkat atau menurunnya penemuan kasus dari tahun ke tahun di suatu wilayah (Kemenkes RI, 2021).



Gambar 1. Angka Notifikasi Semua Kasus TB Per 100,000 Penduduk Indonesia Tahun 2011-2020

Sumber: Kemenkes RI, 2021

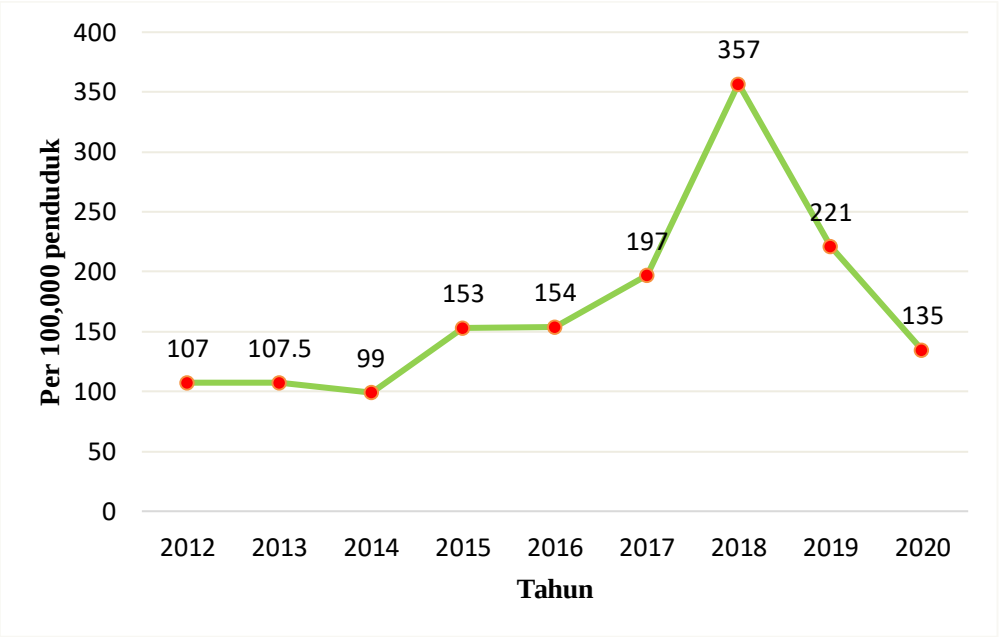
Gambar 1. menunjukkan angka notifikasi semua kasus TB per 100,000 penduduk dari tahun 2010-2020 yang secara nasional memperlihatkan kecenderungan peningkatan CNR sampai tahun 2018 dan menurun pada tahun 2019 dan 2020.



Gambar 2. Angka Notifikasi Semua Kasus TB Per 100,000 Penduduk Menurut Provinsi Tahun 2018

Sumber: Kemenkes RI, 2018

Gambar 2. menunjukan cakupan semua kasus TB (per 100,000 penduduk) menurut Provinsi pada tahun 2018. Salah satu Provinsi dengan CNR yang tertinggi di Indonesia adalah Provinsi Sulawesi Selatan yaiut sebesar 84 per 100,000 penduduk (Kemenkes RI, 2018).

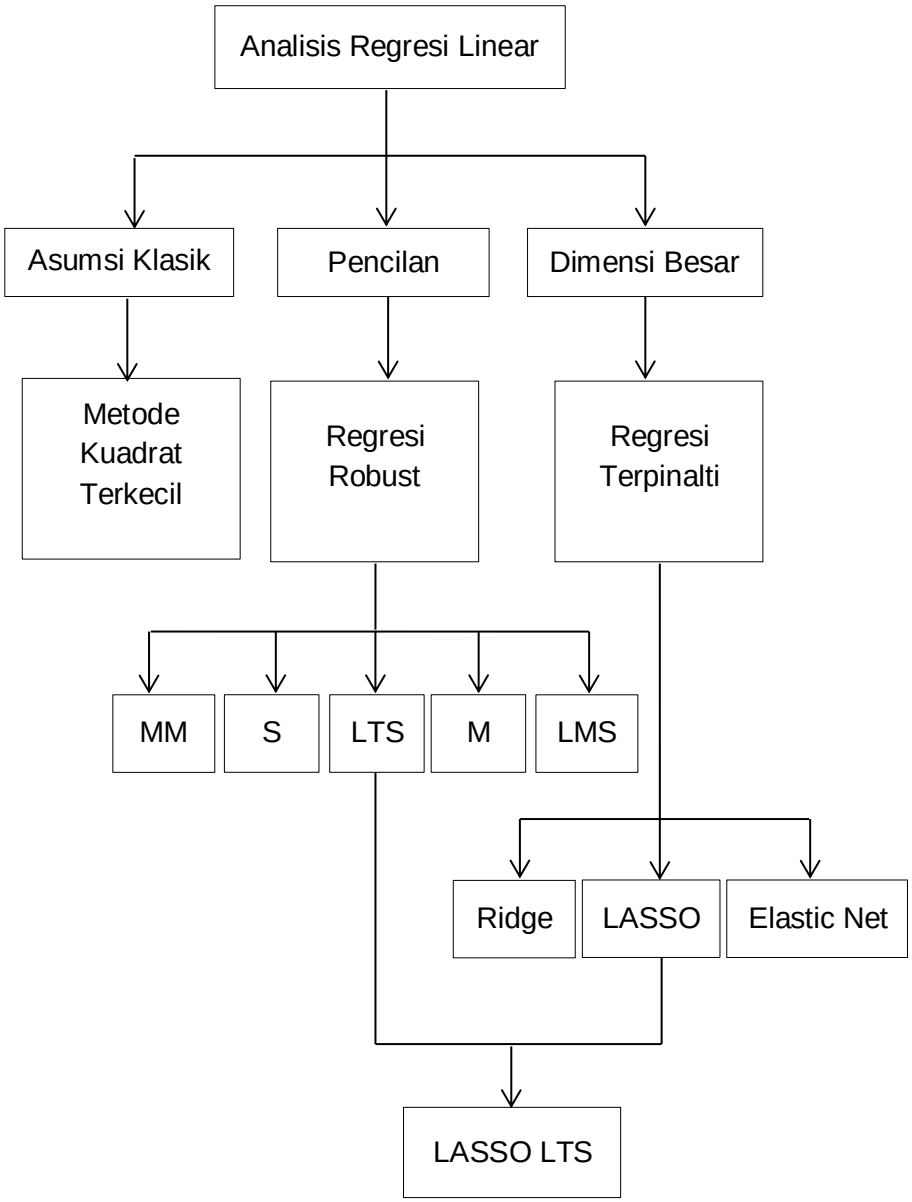


Gambar 3. Angka Notifikasi Semua Kasus TB Per 100.000 Penduduk Provinsi Sulawesi Selatan Tahun 2012-2020

Sumber: Kemenkes RI, 2020

Gambar 3. menunjukkan angka notifikasi semua kasus TB per 100,000 penduduk dari tahun 2012-2020 untuk Provinsi Sulawesi Selatan yang juga memperlihatkan kecenderungan peningkatan CNR sampai tahun 2018 dan menurun pada tahun 2019 dan 2020. Namun, penurunan jumlah kasus tuberkulosis masih tetap harus diwaspadai karena adanya disparitas (kesenjangan) penyebaran penyakit tuberkulosis antar daerah di Sulawesi Selatan, misalnya kabupaten Sidrap pada tahun 2018 memiliki jumlah kasus tuberkulosis sebanyak 493 kasus, namun pada tahun 2020 mengalami kenaikkan jumlah kasus sebesar 591 kasus. Hal tersebut diduga karena datanya mengandung pencilan (*outliers*). Oleh sebab itu, perlu dilakukan penelitian lebih lanjut mengenai faktor-faktor yang mempengaruhi penyebaran penyakit tuberkulosis di Provinsi Sulawesi Selatan.

2.10 Kerangka Konseptual



Gambar 4. Diagram Kerangka Konseptual Metode Penelitian