

**EFEKTIVITAS ANALISIS REGRESI KOMPONEN UTAMA *ROBUST*  
DENGAN METODE MCD-LTS PADA DATA INDEKS PEMBANGUNAN  
MANUSIA (IPM) PROVINSI SUMATERA UTARA**

**(Skripsi)**

**Oleh**

**NURJANAH**



**JURUSAN MATEMATIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMPUNG  
BANDAR LAMPUNG  
2023**

## **ABSTRACT**

### **EFFECTIVENESS OF ROBUST MAIN COMPONENT REGRESSION ANALYSIS USING THE MCD-LTS METHOD ON HUMAN DEVELOPMENT INDEX (HDI) DATA OF NORTH SUMATRA PROVINCE**

**By**

**NURJANAH**

Principal Component Regression (PCR) is one of the methods used in overcoming multicollinearity problems with the stages of conducting Principal Component Analysis (PCA) and then regressing with Ordinary Least Squares (OLS). If there are outliers, the PCR robust method is used, namely using the Minimum Covariance Determinant (MCD) method which is then regressed using the Least Trimmed Squares (LTS) method which in this study used data from the Human Development Index (HDI) of North Sumatra province which is influenced by seven independent variables, including labor force participation rates, percentage of poor population, population aged 15 years and over who are employed, school enrollment rates, number of school facilities, number of health facilities, and life expectancy. The purpose of this study was to measure the effectiveness of the application of the MCD and LTS methods in robust principal component regression for modeling the Human Development Index (HDI) of North Sumatra province. The results of this study indicate that the MCD-LTS method is considered effective in overcoming multicollinearity and outlier problems more than the classical AKU-OLS method with an Adjusted  $R^2$  value of 0.5573 and an RMSE value of 0.4339.

**Keyword :** Principal Component Regression, Multicollinearity, Outlier, Robust, Minimum Covariance Determinant, Least Trimmed Squares, Human Development Index

## ABSTRAK

### EFEKTIVITAS ANALISIS REGRESI KOMPONEN UTAMA *ROBUST* DENGAN METODE MCD-LTS PADA DATA INDEKS PEMBANGUNAN MANUSIA (IPM) PROVINSI SUMATERA UTARA

Oleh

NURJANAH

Regresi Komponen Utama (RKU) merupakan salah satu metode yang digunakan dalam mengatasi masalah multikolinearitas dengan tahapan melakukan Analisis Komponen Utama (AKU) kemudian diregresikan dengan *Ordinary Least Squares* (OLS). Apabila terdapat *outlier*, digunakan metode *robust* pada RKU, yaitu menggunakan metode *Minimum Covariance Determinant* (MCD) yang kemudian diregresikan dengan metode *Least Trimmed Squares* (LTS) yang dalam penelitian ini digunakan data Indeks Pembangunan Manusia (IPM) provinsi Sumatera Utara yang dipengaruhi tujuh variabel bebas yaitu tingkat partisipasi angkatan kerja, presentase penduduk miskin, penduduk umur 15 tahun keatas yang bekerja, angka partisipasi sekolah, jumlah fasilitas sekolah, banyaknya sarana kesehatan dan angka harapan hidup. Tujuan penelitian ini yaitu mengukur efektivitas penerapan metode MCD dan LTS dalam regresi komponen utama *robust* untuk pemodelan Indeks Pembangunan Manusia (IPM) provinsi Sumatera Utara. Hasil dari penelitian ini menunjukkan bahwa metode *robust* MCD-LTS dinilai lebih efektif dalam mengatasi masalah multikolinearitas dan *outlier* daripada metode klasik AKU-OLS dengan nilai *Adjusted R*<sup>2</sup> 0.5573 dan nilai RMSE 0.4339.

**Kata kunci** : Regresi Komponen Utama, Multikolinearitas, *Outlier*, *Robust*, *Minimum Covariance Determinant*, *Least Trimmed Squares*, Indeks Pembangunan Manusia

**EFEKTIVITAS ANALISIS REGRESI KOMPONEN UTAMA *ROBUST*  
DENGAN METODE MCD-LTS PADA DATA INDEKS PEMBANGUNAN  
MANUSIA (IPM) PROVINSI SUMATERA UTARA**

**Oleh**

**Nurjanah**

**Skripsi**

**Sebagai Salah Satu Syarat untuk Mencapai Gelar  
SARJANA MATEMATIKA**

**Pada**

**Jurusan Matematika  
Fakultas Matematika dan Ilmu Pengetahuan Alam  
Universitas Lampung**



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMPUNG  
BANDAR LAMPUNG  
2023**

Judul Skripsi : **EFEKTIVITAS ANALISIS REGRESI  
KOMPONEN UTAMA *ROBUST* DENGAN  
METODE MCD-LTS PADA DATA INDEKS  
PEMBANGUNAN MANUSIA (IPM) PROVINSI  
SUMATERA UTARA**

Nama Mahasiswa : **Nurjanah**

Nomor Pokok Mahasiswa : **1917031043**

Program Studi : **Matematika**

Fakultas : **Matematika dan Ilmu Pengetahuan Alam**



1. Komisi Pembimbing

**Dr. Khoirin Nisa, M.Si**  
**NIP. 19740726 200003 2 001**

**Drs. Eri Setiawan, M.Si.**  
**NIP. 19581101 198803 1 002**

2. Ketua Jurusan Matematika

**Dr. Aang Nuryaman, S.Si., M.Si**  
**NIP. 19740316 200501 1 001**

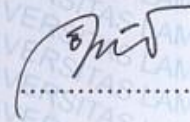
**MENGESAHKAN**

1. Tim Penguji

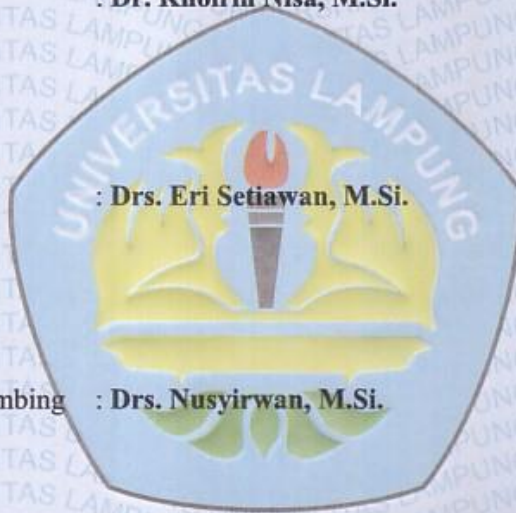
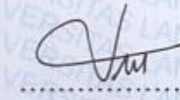
Ketua : **Dr. Khoirin Nisa, M.Si.**



Sekretaris : **Drs. Eri Setiawan, M.Si.**



Penguji  
Bukan Pembimbing : **Drs. Nusyirwan, M.Si.**



2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam  
Universitas Lampung



**Dr. Eng Heri Satria, S.Si., M.Si.**  
**NIP. 197110012005011002**

Tanggal Lulus Ujian Skripsi : 23 Agustus 2023

## PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : **Nurjanah**  
Nomor Pokok Mahasiswa : **1917031043**  
Jurusan : **Matematika**  
Judul Skripsi : **EFEKTIVITAS ANALISIS REGRESI  
KOMPONEN UTAMA *ROBUST* DENGAN  
METODE MCD-LTS PADA DATA INDEKS  
PEMBANGUNAN MANUSIA (IPM) PROVINSI  
SUMATERA UTARA**

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 23 Agustus 2023  
Penulis,



**Nurjanah**  
NPM. 1917031043

## **RIWAYAT HIDUP**

Penulis bernama lengkap Nurjanah. Lahir di Kalianda pada tanggal 27 April 2002, merupakan anak kedua dari tiga bersaudara, pasangan Bapak Tukijan dan Ibu Lestari. Penulis mempunyai kakak bernama Arif Abdullah, S.Pd. dan adik bernama Saidah Rahmawati.

Penulis mengawali Pendidikan Sekolah Dasar di SD Negeri 1 Sidorejo pada tahun 2007-2013. Selanjutnya penulis melanjutkan Pendidikan Sekolah Menengah Pertama di SMP Negeri 1 Sidomulyo pada tahun 2013-2016 dan melanjutkan pendidikan Sekolah Menengah Atas di SMA Negeri 1 Sidomulyo jurusan MIPA pada tahun 2016-2019.

Pada tahun 2019 penulis melanjutkan pendidikan Strata Satu (S1) di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Lampung. melalui jalur SBMPTN. Selama menjadi mahasiswa penulis aktif di beberapa organisasi yaitu Generasi Muda HIMATIKA (GEMATIKA) 2019, Himpunan Mahasiswa Matematika (HIMATIKA) FMIPA Unila sebagai Anggota Biro Dana dan Usaha periode 2020, dan Anggota Aktif Koperasi Mahasiswa (KOPMA) Universitas Lampung sejak tahun 2020.

Pada bulan Januari hingga Februari 2022 penulis melaksanakan Kerja Praktik (KP) di Badan Pusat Statistik (BPS) Kabupaten Lampung Selatan sebagai bentuk pengembangan diri serta menerapkan ilmu yang telah didapat selama perkuliahan. Selanjutnya pada bulan Juni hingga Agustus 2022 penulis melaksanakan Kuliah Kerja Nyata (KKN) Periode II di Desa Belimbing Sari, Kecamatan Jabung, Kabupaten Lampung Timur sebagai bentuk pengabdian kepada masyarakat.



## KATA INSPIRASI

*“...dan berbuat baiklah. Sungguh, Allah menyukai orang-orang yang berbuat baik”*  
(Q.S Al-Baqarah: 195)

*“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya...”*  
(Q.S Al-Baqarah: 286)

*“Karena Sesungguhnya bersama kesulitan ada kemudahan, Sesungguhnya bersama kesulitan ada kemudahan”*  
(Q.S Al-Insyirah: 5-6)

*“Cukuplah Allah (menjadi penolong) bagi kami dan Dia sebaik-baiknya pelindung”*  
(Q.S Ali Imran : 173)

## **PERSEMBAHAN**

Dengan mengucapkan puji dan syukur saya haturkan kepada Allah SWT. Yang telah memberikan rahmat, hidayah, dan karunia-Nya kepada saya. Saya persembahkan karya sederhana dengan penuh ketulusan hati sebagai rasa cinta dan sayang saya kepada :

### **Mama, Bapak, Kakak dan Adikku**

Terima kasih telah memberikan kasih sayang, semangat dan doa yang tiada henti untuk kelancaran setiap langkahku. Terima kasih atas segala pengorbanan, nasihat dan dukungan yang membuatku bersemangat dan selalu merasa bersyukur.

### **Dosen Pembimbing dan Pembahas**

Terima kasih kepada dosen pembimbing dan pembahas yang telah membantu, memberikan motivasi, arahan serta ilmu yang berharga kepada penulis.

### **Almamater Tercinta Universitas Lampung**

## SANWACANA

Puji dan syukur penulis haturkan kepada Allah SWT., yang telah memberikan rahmat, hidayah, serta karunia-Nya kepada penulis, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Efektivitas Analisis Regresi Komponen Utama *Robust* dengan Metode MCD-LTS pada Data Indeks Pembangunan Manusia (IPM) Provinsi Sumatera Utara”. terselesaikannya skripsi ini tidak lepas dari dukungan, bimbingan, saran, serta do’a dari berbagai pihak. Dengan segala kerendahan hati penulis mengucapkan terima kasih kepada:

1. Ibu Dr. Khoirin Nisa, M.Si., selaku Dosen Pembimbing 1 atas kesabaran dan kesediaannya untuk memberikan bimbingan, kritik, dan saran dalam proses penyelesaian skripsi ini serta selalu meluangkan waktunya untuk bimbingan.
2. Bapak Drs. Eri Setiawan, M.Si., selaku Dosen Pembimbing 2 yang telah memberikan saran serta arahan kepada penulis dan meluangkan waktunya untuk bimbingan.
3. Bapak Drs. Nusyirwan, M.Si., selaku Dosen Pembahas skripsi yang telah memberikan kritik, saran dan masukan penulis dalam penyelesaian skripsi.
4. Ibu Dorrah Azis, M.Si., selaku Dosen Pembimbing Akademik
5. Bapak Dr. Aang Nuryaman, S.Si, M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Bapak Dr. Eng Heri Satria, S.Si., M.Si, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Seluruh dosen, Staf, dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang telah banyak membantu selama perkuliahan.

8. Bapak Tukijan, Mama Lestari selaku kedua orang tuaku tercinta telah memberikan kasih sayang, semangat dan doa yang tiada henti untuk kelancaran setiap langkahku, semoga bapak dan mama diberi umur panjang, sehat selalu, lancar rezekinya dan diberi kemudahan dalam hidupnya.
9. Kakakku Arif Abdullah, S.Pd., Kakak iparku Vera Yulyanti, S.Pd., dan Adikku Saidah Rahmawati yang selalu memberi doa dan dukungan selama perkuliahan serta penulisan skripsi ini.
10. Kucingku Zoey yang telah menemani dan selalu menghiburku selama perkuliahan hingga selesainya skripsi ini.
11. Sahabat-sahabatku tersayang, yaitu Arifki, Clara, Rizqa dan Vista yang senantiasa membantu dan menemaniku dalam suka maupun duka selama perkuliahan serta penulisan skripsi ini.
12. Teman-teman satu bimbingan, yaitu Azizah, Melis, Widya dan Yusril yang selalu memberikan semangat dalam pengerjaan skripsi ini.
13. Teman-teman Matematika 2019, terima kasih atas kebersamaannya.
14. Seluruh pihak yang telah membantu dan terlibat dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa banyak terdapat kekurangan dalam penulisan skripsi ini masih. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun guna penelitian selanjutnya agar lebih baik.

Bandar Lampung, 23 Agustus 2023  
Penulis

Nurjanah  
NPM. 1917031043

## DAFTAR ISI

	Halaman
<b>DAFTAR TABEL .....</b>	<b>xv</b>
<b>DAFTAR GAMBAR.....</b>	<b>xvi</b>
<b>I. PENDAHULUAN.....</b>	<b>1</b>
1.1 Latar Belakang dan Masalah .....	1
1.2 Tujuan Penelitian.....	3
1.3 Manfaat Penelitian.....	3
<b>II. TINJAUAN PUSTAKA.....</b>	<b>4</b>
2.1 Analisis Regresi.....	4
2.2 Asumsi Analisis Regresi .....	5
2.3 <i>Ordinary Least Square</i> (OLS).....	6
2.4 Pengujian Signifikansi Regresi Berganda .....	8
2.5 Koefisien Determinasi ( $R^2$ ) .....	9
2.5.1 Koefisien Determinasi yang Disesuaikan ( <i>Adjusted R<sup>2</sup></i> ).....	10
2.6 <i>Root Mean Square Error</i> (RMSE) .....	11
2.7 Multikolinearitas .....	11
2.8 <i>Outlier</i> .....	13
2.9 Analisis Komponen Utama .....	14
2.10 Regresi Komponen Utama .....	18
2.11 <i>Minimum Covariance Determinant</i> (MCD).....	19
2.12 <i>Least Trimmed Square</i> (LTS).....	21
<b>III. METODOLOGI PENELITIAN .....</b>	<b>23</b>
3.1 Waktu dan Tempat Penelitian .....	23
3.2 Data Penelitian .....	23
3.3 Metodologi Penelitian .....	23
3.4 Diagram Alir ( <i>Flowchart</i> ) .....	26
<b>IV. HASIL DAN PEMBAHASAN.....</b>	<b>30</b>
4.1 Analisis Deskriptif Data .....	30
4.2 Analisis Regresi Berganda .....	31
4.2.1 Pengujian Signifikansi Regresi Berganda.....	31
4.2.2 Ukuran Kecocokan Model .....	33

4.3	Deteksi Multikolinearitas .....	33
4.3	Deteksi <i>Outlier</i> .....	33
4.5	Analisis Regresi Komponen Utama Klasik.....	36
4.5.1	Analisis Komponen Utama Klasik (AKU Klasik).....	36
4.5.2	Regresi Komponen Utama Klasik-OLS (RKU Klasik-OLS) .....	37
4.6	Analisis Regresi Komponen Utama <i>Robust</i> .....	40
4.6.1	Analisis Komponen Utama <i>Robust</i> (AKU <i>Robust</i> ).....	40
4.6.2	Regresi Komponen Utama <i>Robust</i> -LTS (RKU <i>Robust</i> -LTS).....	42
4.7.	Perbandingan Model Regresi .....	45
<b>V.</b>	<b>KESIMPULAN.....</b>	<b>46</b>

## **DAFTAR PUSTAKA**

## **LAMPIRAN**

## DAFTAR TABEL

Tabel	Halaman
1. Analisis Deskriptif Data.....	30
2. Uji Signifikansi Koefisien Regresi Secara Individual .....	32
3. Deteksi Multikolinearitas dengan Nilai VIF .....	34
4. Deteksi Multikolinearitas dengan Nilai Koefisien Korelasi Parsial .....	34
5. Pendeteksian <i>outlier</i> pada variabel bebas .....	35
6. Uji t RKU Klasik-OLS dengan Dua Komponen Utama.....	38
7. Uji t RKU <i>Robust</i> -LTS dengan Dua Komponen Utama.....	43
8. Perbandingan Efektivitas Model Regresi.....	45

## DAFTAR GAMBAR

Gambar	Halaman
1. Diagram Alir .....	26
2. Diagram Alir Analisis Komponen Utama Klasik .....	28
3. Diagram Alir Analisis Komponen Utama <i>Robust</i> MCD .....	29



## I. PENDAHULUAN

### 1.1 Latar Belakang dan Masalah

Analisis regresi merupakan suatu metode analisis statistik yang digunakan untuk menyelidiki model hubungan antara sebuah variabel respon  $Y$  dengan satu atau lebih variabel prediktor  $X_1, X_2, \dots, X_k$ . Analisis regresi dapat diklasifikasikan menjadi dua macam, yakni analisis regresi sederhana dan analisis regresi berganda. Analisis regresi yang menjelaskan hubungan satu variabel respon dengan satu variabel prediktor dinamakan analisis regresi sederhana. Sedangkan, untuk analisis regresi yang menjelaskan hubungan satu variabel respon dengan lebih dari satu variabel prediktor dinamai dengan analisis regresi berganda.

Pada ilmu statistik, suatu model regresi dikatakan baik, jika asumsi-asumsi klasiknya terpenuhi, yaitu residual berdistribusi normal, tidak terjadi autokorelasi, heteroskedastisitas, dan multikolinieritas. Oleh karena itu, proses kontrol terhadap model perlu dilakukan untuk menelaah apakah asumsi-asumsi tersebut terpenuhi atau tidak.

Dalam analisis regresi linear berganda kerap muncul masalah, salah satunya adalah adanya multikolinieritas yang terjadi karena adanya hubungan (korelasi) antar variabel prediktor dalam model regresi (Notiragayu & Nisa, 2008). Adanya multikolinieritas ini dapat menyebabkan nilai varian besar sehingga sulit untuk mendapatkan hasil taksiran yang tepat dan akurat. Akibat dari hal tersebut, interval pendugaan akan cenderung lebih besar dan nilai statistik uji  $t$  akan kecil, sehingganya membuat variabel bebas statistik tidak signifikan mempengaruhi

variabel tidak bebas, walaupun nilai koefisien determinasi ( $R^2$ ) masih relatif tinggi (Widarjono, 2005). Maka sebab itu, untuk memperoleh model regresi yang baik dibutuhkan metode untuk mengatasi masalah multikolinearitas tersebut.

Menurut Montgomery, *et al.* (2012) salah satu metode statistik yang berguna dalam mengatasi masalah multikolinearitas yaitu metode Regresi Komponen Utama (RKU). Terdapat dua tahapan dalam menangani masalah multikolinearitas ini, tahap kesatu adalah melakukan Analisis Komponen Utama (AKU) dengan vektor eigen dari matriks kovarian, kemudian untuk tahap kedua adalah meregresikan hasil dari tahap pertama tersebut menggunakan metode *Ordinary Least Squares* (OLS). Akan tetapi, analisis pada kedua tahap tersebut menurut Notiragayu dan Nisa (2008) sangat sensitif terhadap *outlier*. *Outlier* ialah pengamatan yang berada jauh (ekstrem) dari pengamatan-pengamatan lainnya. Maka sebab dari itu, Regresi Komponen Utama (RKU) telah berkembang, sehingga menjadikan Regresi Komponen Utama (RKU) *robust*, yang menggunakan metode *robust* dalam kedua tahapan tersebut.

Dalam penelitian ini, metode analisis komponen utama *robust* dilakukan menggunakan matriks peragam *robust*, yaitu dengan metode *Minimum Covariance Determinant* (MCD) dan analisis regresi komponen utamanya dilakukan dengan menggunakan metode *Least Trimmed Squares* (LTS). Metode MCD dipilih karena dinilai lebih efektif dibanding metode lainnya, metode tersebut juga dinilai sangat *robust* dalam penentuan lokasi dan sebaran multivariat (Rousseeuw & Driessen, 1999). Sedangkan metode LTS dipilih karena Nisa (2006) menyatakan bahwa metode tersebut *robust* pada nilai *outlier* yang memiliki nilai *breakdown point* yang lebih besar bila disandingkan dengan metode-metode lainnya.

Sebelumnya penelitian analisis regresi komponen utama *robust* dengan metode MCD-LTS telah dilakukan oleh Larasati, *et al.* (2020) yang dibandingkan dengan regresi komponen utama klasik berdasarkan nilai bias dan *Mean Square Error* (MSE) menggunakan data simulasi dan diperoleh bahwa metode MCD-LTS lebih efektif dan efisien dalam mengatasi masalah multikolinearitas.

Indeks Pembangunan Manusia (IPM) ialah indikator penting yang digunakan untuk mengukur tingkat keberhasilan pada upaya membangun kualitas hidup manusia. IPM terdiri dari tiga unsur yakni standar kehidupan atau ekonomi, pendidikan, dan kesehatan. Faktor-faktor dalam setiap unsur pembentuk Indeks Pembangunan Manusia (IPM) cenderung mempunyai korelasi yang kuat satu dengan yang lain dikarenakan faktor-faktor tersebut saling berinteraksi sehingganya menyebabkan masalah multikolinearitas (Putra & Ratnasari, 2015). Dalam penelitian ini, peneliti menjadikan Indeks Pembangunan Manusia (IPM) provinsi Sumatera Utara tahun 2021 dan faktor-faktor yang mempengaruhinya menjadi objek dalam penelitian.

Berdasarkan hal tersebut, maka peneliti mengangkat judul “Efektivitas Analisis Regresi Komponen Utama *Robust* dengan Metode MCD-LTS pada Data Indeks Pembangunan Manusia (IPM) Provinsi Sumatera Utara”

## **1.2 Tujuan Penelitian**

Adapun tujuan penelitian ini yaitu mengukur efektivitas penerapan metode MCD dan LTS dalam regresi komponen utama *robust* untuk pemodelan Indeks Pembangunan Manusia (IPM) provinsi Sumatera Utara.

## **1.3 Manfaat Penelitian**

Adapun manfaat penelitian ini ialah menjadikan masukan bagi peneliti, mahasiswa serta para pembacanya guna menentukan metode yang baik dalam menganalisis data yang mengandung multikolinearitas.

## II. TINJAUAN PUSTAKA

### 2.1 Analisis Regresi

Analisis regresi merupakan metode statistika yang digunakan dalam menentukan korelasi fungsional suatu variabel yakni variabel terikat dengan satu atau lebih variabel bebas (Gujarati & Porter, 2011). Variabel terikat atau variabel respon dinotasikan dengan  $Y$  yakni ialah variabel yang telah dipengaruhi oleh variabel lain. Sedangkan, variabel bebas atau variabel prediktor dinotasikan dengan  $X$  yaitu variabel yang tidak dipengaruhi variabel lain.

Menurut Sembiring (2003), model regresi artinya model yang menggambarkan tentang korelasi diantara variabel prediktor dengan variabel respon.

Bentuk paling umum dari persamaan model regresi yaitu sebagai berikut ini:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (2.1)$$

dengan :

$Y_i$  = variabel respon (terikat)

$X_{ji}$  = variabel prediktor (bebas)

$\beta_0$  = konstanta (intersep)

$\beta_j$  = koefisien regresi

$\varepsilon_i$  = galat / residual

Persamaan (2.1) jika dilambangkan dalam bentuk matriks maka dapat dituliskan sebagai berikut :

$$\mathbf{Y}_{(nx1)} = \mathbf{X}_{(nxk)}\boldsymbol{\beta}_{(kx1)} + \boldsymbol{\varepsilon}_{(nx1)} \quad (2.2)$$

dimana,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Analisis regresi dapat diklasifikasikan menjadi dua macam, yakni analisis regresi sederhana serta analisis regresi berganda. Analisis regresi yang dapat menjelaskan korelasi atau hubungan satu variabel terikat dengan satu variabel bebas dinamakan analisis regresi sederhana. Sementara itu, analisis regresi yang menjelaskan korelasi atau hubungan satu variabel terikat dengan lebih dari satu variabel bebas dinamakan dengan analisis regresi berganda.

## 2.2 Asumsi Analisis Regresi

Menurut Gujarati & Porter (2011), asumsi analisis regresi terdiri dari:

1. Model regresi linear dalam parameter.
2. Galat berdistribusi normal dengan rata-rata nol dan variansi  $\sigma^2$ ,  $\varepsilon \sim N(0, \sigma^2)$
3. Distribusi dari galat memiliki varians yang konstan.
4. Tidak terdapat autokorelasi antar galat,  $cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
5. Tidak terjadi multikolinearitas, maka dapat diartikan tidak terdapat hubungan atau korelasi linear yang dinilai kuat diantara dua atau lebih variabel bebas dalam model regresi.

### 2.3 Ordinary Least Square (OLS)

Menurut Montgomery, *et al.* (2012), OLS ialah suatu metode yang dipergunakan saat memperkirakan atau mengestimasi parameter regresi dengan cara meminimalkan jumlah kuadrat residual. OLS adalah metode yang paling banyak digunakan dan dibandingkan dengan metode lain pada pembentukan model regresi atau mengestimasi parameter regresi. Estimasi koefisien regresi dengan metode OLS dapat dihitung dengan meminimalkan:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$S(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Karena  $\mathbf{y}'\mathbf{X}\boldsymbol{\beta}$  adalah skalar dan *transposenya* yaitu  $(\mathbf{y}'\mathbf{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$  juga skalar, sehingganya  $\mathbf{y}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$ , maka  $S(\boldsymbol{\beta})$  dapat dinyatakan sebagai berikut:

$$S(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \quad (2.3)$$

Untuk memperoleh estimator OLS ( $\hat{\boldsymbol{\beta}}$ ), yang meminimumkan  $S(\boldsymbol{\beta})$  disyaratkan bahwa

$$\left. \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \mathbf{0}$$

Turunan pertama dari  $S(\boldsymbol{\beta})$  terhadap  $\hat{\boldsymbol{\beta}}$  adalah:

$$\left. \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \quad (2.4)$$

Dari persamaan (2.4), karena  $\left. \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \mathbf{0}$ , maka

$$-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad (2.5)$$

Persamaan (2.5) jika dituliskan dalam bentuk matriks hasilnya:

$$\begin{bmatrix} n & \sum_{i=1}^n X_{1i} & \sum_{i=1}^n X_{2i} & \cdots & \sum_{i=1}^n X_{ki} \\ \sum_{i=1}^n X_{1i} & \sum_{i=1}^n X_{1i}^2 & \sum_{i=1}^n X_{1i}X_{2i} & \cdots & \sum_{i=1}^n X_{1i}X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ki} & \sum_{i=1}^n X_{ki}X_{1i} & \sum_{i=1}^n X_{ki}X_{2i} & \cdots & \sum_{i=1}^n X_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{1i}Y_i \\ \vdots \\ \sum_{i=1}^n X_{ki}Y_i \end{bmatrix}$$

Apabila dijabarkan menjadi:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_{1i} + \hat{\beta}_2 \sum_{i=1}^n X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^n X_{ki} = \sum_{i=1}^n Y_i$$

$$\hat{\beta}_0 \sum_{i=1}^n X_{1i} + \hat{\beta}_1 \sum_{i=1}^n X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^n X_{1i}X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^n X_{1i}X_{ki} = \sum_{i=1}^n X_{1i}Y_i$$

⋮

$$\hat{\beta}_0 \sum_{i=1}^n X_{ki} + \hat{\beta}_1 \sum_{i=1}^n X_{ki}X_{1i} + \hat{\beta}_2 \sum_{i=1}^n X_{ki}X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^n X_{ki}^2 = \sum_{i=1}^n X_{ki}Y_i$$

Untuk mendapatkan solusi dari persamaan (2.5), estimator parameter  $\hat{\beta}$  dapat dirumuskan sebagai berikut:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) \quad (2.6)$$

Ketika asumsi-asumsi analisis regresi terpenuhi, maka dugaan metode kuadrat terkecil atau OLS akan menghasilkan model regresi yang linear, tak bias, serta memiliki varian minimum atau disebut *Best Linear Unbiased Estimator* (BLUE)

Menurut Kutner *et al.* (2005), koefisien regresi pada unit satuannya yang berbeda dalam model regresi yang tak distandarisi bisa mengakibatkan koefisien regresi tak dapat diperbandingkan. Oleh sebab itu, diperlukan standarisasi data menggunakan rumus sebagai berikut:

$$Y_i^* = \frac{Y_i - \bar{Y}}{S_Y} \text{ serta } X_{ji}^* = \frac{X_{ji} - \bar{X}_j}{S_j} \quad (2.7)$$

$$\text{dimana } S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}} \text{ dan } S_j = \sqrt{\frac{\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2}{n-1}}$$

dengan:

$$Y_i^* = \text{variabel terikat dalam bentuk standar}$$

$X_{ji}^*$  = variabel bebas dalam bentuk standar

$\bar{Y}$  = rata-rata variabel terikat

$\bar{X}_j$  = rata-rata variabel bebas

$S_Y$  = standar deviasi variabel terikat

$S_j$  = standar deviasi variabel bebas

$j = 1, 2, \dots, k$

Sehingga diperoleh model regresi standar sebagai berikut:

$$Y_i^* = \beta_1^* X_{1i}^* + \beta_2^* X_{2i}^* + \dots + \beta_k^* X_{ki}^* + \varepsilon_i^* \quad (2.8)$$

Alasan mengapa tidak ada parameter intersep ( $\beta_0^*$ ) dalam model regresi standar adalah perhitungan OLS akan selalu menyebabkan estimasi parameter intersep mendekati nilai nol (Kutner *et al.*, 2005). Hubungan antara estimator regresi standar dengan estimator regresi bentuk asli adalah sebagai berikut:

$$\beta_j = \left( \frac{S_Y}{S_j} \right) \beta_j^*$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \dots - \beta_k \bar{X}_k$$

## 2.4 Pengujian Signifikansi Regresi Berganda

Menurut Montgomery, *et al.* (2012), pada analisis regresi berganda terdapat sejumlah uji signifikansi yang berfungsi guna menguji ketepatan model, uji-uji yang dilakukan pada model maupun individu yaitu:

### 1. Uji F Simultan

Uji tersebut dipergunakan untuk memilih/menguji apakah terdapat korelasi linier antara variabel terikat  $Y$  menggunakan variabel bebas  $X_1, X_2, \dots, X_k$ .

Jika terdapat korelasi linier di antara variabel terikat dengan variabel bebas maka mengindikasikan bahwa model regresi yang terbentuk sesuai.

Hipotesis

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  (Model regresi tidak sesuai)

$H_1$ : terdapat minimal satu  $\beta_j \neq 0$ , dengan  $j = 1, 2, \dots, k$  (Model regresi sesuai)



statistik uji

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}; SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2; SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2.9)$$

Kriteria uji

$H_0$  ditolak jika  $F_0 > F_{\text{tabel}} = F_{(\alpha; k; n-k-1)}$  atau  $p\text{-value} < \alpha$ .

Penolakan  $H_0$  menunjukkan bahwa model sesuai yang berarti terdapat korelasi linier antara variabel terikat dengan variabel bebas.

## 2. Uji t parsial

Uji ini dipergunakan untuk menguji apakah terdapat pengaruh signifikan pada tiap-tiap variabel bebas  $X_1, X_2, \dots, X_k$  terhadap variabel terikat  $Y$ .

Hipotesis

$H_0: \beta_j = 0$  (koefisien regresi ke- $j$  tidak signifikan)

$H_1: \beta_j \neq 0, j = 1, 2, \dots, k$  (koefisien regresi ke- $j$  signifikan)

statistik uji

$$t_0 = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \text{ dengan } \text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \quad (2.10)$$

dimana  $C_{jj}$  adalah elemen diagonal dari  $(\mathbf{X}'\mathbf{X})^{-1}$  dan  $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-k-1}$

Kriteria uji

$H_0$  ditolak jika  $|t_0| > t_{\text{tabel}} = t_{(\alpha/2; n-k-1)}$  atau  $p\text{-value} < \alpha$ .

Penolakan  $H_0$  menunjukkan bahwa variabel bebas  $X_1, X_2, \dots, X_k$  berpengaruh atau signifikan terhadap variabel terikat  $Y$ .

## 2.5 Koefisien Determinasi ( $R^2$ )

Koefisien determinasi ( $R^2$ ) ialah suatu ukuran nilai ataupun nilai yang bisa digunakan guna melihat seberapa jauhkah kecocokan dari model regresi. Koefisien determinasi mengukur proporsi atau persentase keseluruhan variasi dalam variabel terikat yang dijelaskan oleh model regresi (Gujarati & Porter, 2011).

$$R^2 = \frac{SSR}{SST} \quad (2.11)$$

dengan  $SST = SSR + SSE$

Sifat-sifat koefisien determinasi adalah sebagai berikut:

1. Koefisien determinasi merupakan besaran non-negatif.
2. Batasannya adalah  $0 \leq R^2 \leq 1$ . Nilai  $R^2$  sebesar 1 menyatakan kecocokan sempurna, sedangkan nilai  $R^2$  sebesar 0 menyatakan tak terdapat hubungan antara variabel terikat dengan variabel bebas.

Semakin dekat nilai  $R^2$  mendekati 1 maka garis regresi akan semakin baik karena mampu merepresentasikan data aktualnya. Namun, sebaliknya jika semakin mendekati 0 maka garis regresi kurang baik (Sembiring, 2003).

### 2.5.1 Koefisien Determinasi yang Disesuaikan (*Adjusted R<sup>2</sup>*)

Masalah sering muncul ketika menggunakan koefisien determinasi ( $R^2$ ). Artinya, menambahkan variabel independen ke dalam model selalu meningkatkan nilainya, terlepas dari apakah variabel independen tambahan itu terkait dengan variabel dependen (Widarjono, 2005).

Dengan demikian, banyak peneliti menyarankan menggunakan nilai *Adjusted R<sup>2</sup>*. Penggunaan *Adjusted R<sup>2</sup>* merupakan alternatif yang lebih baik dibandingkan dengan  $R^2$  yang kerap menimbulkan masalah (Tobin, 1958)

Interpretasinya sama dengan  $R^2$ , namun demikian nilai *Adjusted R<sup>2</sup>* dapat naik atau turun apabila terdapatnya penambahan variabel independen baru, tergantung dari korelasi di antara variabel independen tambahan tersebut dengan variabel terikatnya. Bila nilainya tidak positif, maka nilai tersebut diasumsikan 0, atau variabel independen tak dapat merepresentasikan variabel dependen.

Menurut Widarjono (2005), nilai *Adjusted R<sup>2</sup>* dapat dihitung dengan rumus:

$$R_{adj}^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n-k-1)}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)} \quad (2.12)$$

## 2.6 Root Mean Square Error (RMSE)

*Root Mean Square Error* (RMSE) merupakan akar kuadrat dari MSE. Keakuratan metodenya ditandai dengan nilai RMSE yang kecil. Metode estimasi yang mempunyai nilai RMSE lebih kecil dikatakan lebih akurat daripada metode yang nilai RMSE lebih besar. RMSE didefinisikan sebagai berikut (Kutner *et al.*, 2005):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\text{MSE}} \quad (2.13)$$

## 2.7 Multikolinearitas

Menurut Widarjono (2005), Multikolinearitas diartikan adanya hubungan antara variabel prediktor dalam regresi berganda. Adanya multikolinearitas ini masih dapat menghasilkan estimator yang BLUE namun memiliki varian yang besar. Akibat terdapatnya multikolinieritas, dalam model regresi linear berganda ialah (Gujarati & Porter, 2011):

1. Estimator OLS bersifat BLUE, namun memiliki varians serta juga kovarians yang cukup tinggi sehingganya tidak mudah memperoleh estimasi yang tepat.
2. Dampak penduga OLS memiliki varians serta kovarians yang besar, maka mengakibatkan interval estimasi akan cenderung lebar dan nilai yang dihitung statistik uji t akan kecil, sehingganya dapat berakibat variabel prediktor tidak lagi signifikan mempengaruhi variabel respon.
3. Meskipun secara individunya variabel prediktor tidak mempengaruhi variabel respon dalam uji statistik t, akan tetapi nilai koefisien determinasi ( $R^2$ ) masih relatif tinggi.

Terdapat beberapa cara yang bisa dipergunakan dalam mendeteksi masalah multikolinieritas pada suatu model regresi, berganda yaitu:

1. Bila nilai  $R^2$  tinggi, namun sekedar sedikit variabel prediktor yang signifikan Satu diantara ciri terdapatnya gejala multikolinieritas ialah model memiliki koefisien determinasi ( $R^2$ ) yang tinggi (misalnya: antara 0,7 dan 1) namun

sekedar sedikit variabel prediktor yang signifikan mempengaruhi variabel respon melalui uji t parsial. Tetapi, berdasarkan uji F secara statistik signifikan yang diartikan bahwa semua variabel prediktor secara bersama-sama mempengaruhi variabel respon. Dalam kasus ini terdapat suatu kontradiktif dimana berdasarkan uji t parsial, variabel prediktor tidak berpengaruh terhadap variabel respon, tetapi secara bersama-sama variabel prediktor mempengaruhi variabel respon (Gujarati & Porter, 2011).

2. Melihat koefisien korelasi parsial antar variabel prediktor

Penggunaan koefisien korelasi parsial membantu dalam mengidentifikasi hubungan antara variabel prediktor yang unik, setelah menghilangkan efek variabel prediktor lainnya. Ini membantu mengidentifikasi masalah multikolinearitas dan memahami kontribusi masing-masing variabel dalam analisis regresi (Hair *et al.*, 2014). Jika koefisien korelasi lebih besar dari 0,75 atau mendekati 1 maka dapat diduga terdapat multikolinieritas pada model regresi tersebut. Koefisien korelasi parsial antar variabel prediktor dapat dihitung dengan rumus korelasi *pearson* sebagai berikut:

$$r_{X_j X_l} = \frac{n \sum X_j X_l - \sum X_j \sum X_l}{\sqrt{n \sum X_j^2 - (\sum X_j)^2} \sqrt{n \sum X_l^2 - (\sum X_l)^2}} \quad (2.14)$$

dengan  $j = 1, 2, \dots, k$  dan  $l = 1, 2, \dots, k$ .

3. Melihat nilai *Variance Inflation Factors* (VIF)

*Variance Inflation Factors* (VIF) yang merupakan salah satu indikator guna mengukur besarnya multikolinieritas. VIF menunjukkan peningkatan ragam dari koefisien regresi yang disebabkan oleh terdapatnya ketergantungan linier antar variabel prediktor dalam model regresi.

*Variance Inflation Factors* (VIF) bisa dihitung menggunakan rumus sebagai berikut:

$$VIF = \frac{1}{1 - R_j^2} \quad (2.15)$$

dengan  $R_j^2$  ialah koefisien determinasi ke- $j$  dimana  $j = 1, 2, \dots, k$ .

Nilai VIF diatas 5 atau 10 telah menandakan adanya multikolinieritas dalam model regresi yang harus diatasi (Field., 2013).

## 2.8 Outlier

Menurut Sembiring (2003), umumnya, *outlier* bisa ditafsirkan sebagai data yang tak mengikuti pola umum model atau data yang keluar dari model lain dan tak terdapat dalam selang kepercayaan.

Menurut Fauzia *et al.*(2019), ada beberapa peluang data *outlier*, yaitu *outlier* pada variabel bebas, *outlier* pada variabel terikat, maupun keduanya. Adanya *outlier* dalam variabel bebas dapat dideteksi dengan menghitung jarak mahalanobis. Jarak mahalanobis merupakan metode yang kuat dalam mendeteksi *outlier* karena dapat memperhitungkan korelasi yang lebih tinggi (Ghorbani, 2019). Untuk menghitung jarak mahalanobis dipergunakan vektor rata-rata dan matriks kovarian.

Seuatu pengamatan  $X_i$  dideteksi menjadi *outlier* bila jarak mahalanobisnya:

$$d_{MD}^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) > \chi_{(k,\alpha)}^2 \quad (2.16)$$

dengan  $\bar{\mathbf{X}}$  dan  $\mathbf{S}$  merupakan vektor rata-rata dan matriks kovarian dari data.

Kemudian, guna mendeteksi ada tidaknya *outlier* dalam variabel terikat bisa diatasi dengan melihat residual (*error*) yang berada di model regresi tersebut. Salah satu metode yang bisa dipergunakan yaitu metode *DFFITS* (*Difference in Fit Standardized*).

Menurut Kutner *et al.* (2005), metode *DFFITS* digunakan untuk mendeteksi diagnosis yang terhapus (*deletion diagnostic*) dari pengamatan ke- $i$  pada nilai prediksi atau *fitted*-nya. Perhitungan  $DFFITS_i$  adalah sebagai berikut:

$$DFFITS_i = t_i \left( \frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}} \quad (2.17)$$

dimana  $t_i$  adalah *studentized deleted residual* untuk kasus ke- $i$  yang dapat dihitung dengan rumus:

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}} \quad (2.18)$$

dimana:

$e_i$  = residual ke- $i$

$h_{ii}$  = elemen baris ke- $i$  kolom ke- $i$  dari matriks  $\mathbf{H}$ .

Menurut Montgomery *et al.* (2012),  $\mathbf{H}$  disebut matriks *hat* karena mentransformasi vektor respon  $Y$  ke dalam vektor respon kecocokan  $\hat{Y}$ . Dengan:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

Sehingga  $\mathbf{H}$  adalah matriks berukuran  $n \times n$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (2.19)$$

Data disebut *outlier* jika nilai  $|DFFITs_i| > 2\sqrt{\frac{p}{n}}$ ;  $p$  merupakan banyaknya parameter dan  $n$  ialah banyaknya pengamatan (Perihatini, 2018).

## 2.9 Analisis Komponen Utama

Analisis komponen utama merupakan analisis multivariat yang mentransformasi variabel-variabel asal yang saling berhubungan sebagai variabel-variabel baru yang tak saling berhubungan, dengan mereduksi sejumlah variabel tersebut sehingga memiliki dimensi yang lebih kecil, tetapi tetap bisa menjelaskan sebagian besar keragaman variabel aslinya (Jolliffe, 2002).

Dalam analisis komponen utama, diacukan bahwa skala pengukuran berasal dari  $X_1, X_2, \dots, X_k$  sama, setelah itu dibuat variabel baru  $U$  yang diklaim menjadi komponen utama yang artinya kombinasi linier dari  $X_1, X_2, \dots, X_k$  dengan bentuk sebagai berikut:

$$\begin{aligned}U_1 &= \mathbf{a}_1' \mathbf{X} = a_{11} X_1 + a_{12} X_2 + \dots + a_{1k} X_k \\ U_2 &= \mathbf{a}_2' \mathbf{X} = a_{21} X_1 + a_{22} X_2 + \dots + a_{2k} X_k \\ &\vdots \\ U_k &= \mathbf{a}_k' \mathbf{X} = a_{k1} X_1 + a_{k2} X_2 + \dots + a_{kk} X_k\end{aligned} \quad (2.20)$$

dengan

$$\begin{aligned}\text{Cov}(U_j, U_i) &= \mathbf{a}_j' \boldsymbol{\Sigma} \mathbf{a}_i && \text{dan} \\ \text{Var}(U_j) &= \mathbf{a}_j' \boldsymbol{\Sigma} \mathbf{a}_j && j = 1, 2, \dots, k\end{aligned}$$

Karena  $\text{Cov}(\mathbf{X}) = \mathbf{\Sigma}$  nilainya tidak diketahui, maka  $\mathbf{\Sigma}$  dapat diduga dari sampel yaitu  $\hat{\mathbf{\Sigma}} = \mathbf{S}$  yang didefinisikan sebagai berikut:

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1k} \\ S_{21} & S_{22} & \dots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \dots & S_{kk} \end{bmatrix}$$

dengan

$$S_{ij} = \frac{1}{n-1} \sum_{r=1}^n (X_{ri} - \bar{X}_i)(X_{rj} - \bar{X}_j) \text{ dan } S_{ij} = S_{ji}, i, j = 1, 2, \dots, k \quad (2.21)$$

Sehingga dapat dirumuskan bahwa

$$\begin{aligned} \text{Cov}(U_j, U_i) &= \mathbf{a}_j' \mathbf{S} \mathbf{a}_i & \text{dan} \\ \text{Var}(U_j) &= \mathbf{a}_j' \mathbf{S} \mathbf{a}_j & j = 1, 2, \dots, k \end{aligned}$$

Ketentuan dalam membuat komponen utama yang menggambarkan kombinasi linier yang berasal dari  $X_1, X_2, \dots, X_k$  supaya memiliki varian maksimum ialah dengan menentukan *eigenvector* yaitu  $\mathbf{a}'_j = [\mathbf{a}_{1j}, \mathbf{a}_{2j}, \dots, \mathbf{a}_{kj}]$  menggunakan  $j = 1, 2, \dots, k$ , sedemikian sehingga  $\text{Var}(U_j) = \mathbf{a}'_j \mathbf{S} \mathbf{a}_j$  maksimum berkendala  $\mathbf{a}'_j \mathbf{a}_j = 1$ .

Untuk mencari *eigenvector* yaitu  $\mathbf{a}'_j = [\mathbf{a}_{1j}, \mathbf{a}_{2j}, \dots, \mathbf{a}_{kj}]$  dari  $\mathbf{S}$  berlaku  $(\mathbf{S} - \lambda \mathbf{I}) \mathbf{a}_j = \mathbf{0}$ . Namun, sebelum mencari *eigenvector*, terlebih dahulu mencari *eigenvalue* ( $\lambda$ ) berasal dari  $\mathbf{S}$  dengan syarat  $|\mathbf{S} - \lambda \mathbf{I}| = 0$ .

Keragaman yang mampu digambarkan oleh komponen utama ke- $j$  terhadap keragaman total adalah

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \dots + \lambda_k} \times 100\% ; j = 1, 2, \dots, k \quad (2.22)$$

Sementara itu, secara kumulatif, keragaman total yang digambarkan oleh  $m$  komponen utama adalah

$$\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^k \lambda_j} \times 100\% \text{ dengan } m \leq k$$

$m$  merupakan banyaknya komponen utama yang terpilih.

Menurut Widhiharih (2010) dalam Fauzia *et al.* (2019), dalam persoalan regresi skala pengukuran pada variabel-variabel bebas  $X_1, X_2, \dots, X_k$  umumnya belum sama, sehingganya perlu disamakan menggunakan transformasi dalam variabel baku  $Z$ .

Variabel baku  $Z$  didapatkan dari hasil transformasi terhadap variabel asal yang dirumuskan sebagai berikut:

$$\begin{aligned} Z_1 &= \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \\ Z_2 &= \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}} \\ &\vdots \\ Z_k &= \frac{(X_k - \mu_k)}{\sqrt{\sigma_{kk}}} \end{aligned} \quad (2.23)$$

Komponen utama yang dibentuk menjadi kombinasi linier berasal dari variabel yang dibakukan yaitu  $Z_1, Z_2, \dots, Z_k$  adalah:

$$\begin{aligned} U_1 &= \mathbf{a}_1' \mathbf{Z} = a_{11} Z_1 + a_{12} Z_2 + \dots + a_{1k} Z_k \\ U_2 &= \mathbf{a}_2' \mathbf{Z} = a_{21} Z_1 + a_{22} Z_2 + \dots + a_{2k} Z_k \\ &\vdots \\ U_k &= \mathbf{a}_k' \mathbf{Z} = a_{k1} Z_1 + a_{k2} Z_2 + \dots + a_{kk} Z_k \end{aligned} \quad (2.24)$$

dengan  $\text{Cov}(\mathbf{Z})$  merupakan matriks korelasi  $\mathbf{R}$  yang dituliskan sebagai berikut:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{bmatrix}$$

Nilai  $r_{ij}$  dapat dihitung dengan rumus korelasi *pearson* berikut:

$$r_{ij} = r_{Z_i Z_j} = \frac{n \sum Z_i Z_j - \sum Z_i \sum Z_j}{\sqrt{n \sum Z_i^2 - (\sum Z_i)^2} \sqrt{n \sum Z_j^2 - (\sum Z_j)^2}} \quad (2.25)$$

Atau dapat diperoleh dari matriks kovarian standar ( $\mathbf{S}^*$ ) yaitu

$$\mathbf{S}^* = \begin{bmatrix} S_{11}^* & S_{12}^* & \dots & S_{1k}^* \\ S_{21}^* & S_{22}^* & \dots & S_{2k}^* \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1}^* & S_{k2}^* & \dots & S_{kk}^* \end{bmatrix}$$

dimana

$$S_{ij}^* = \frac{1}{n-1} \sum_{r=1}^n (Z_{ri} - \bar{Z}_i)(Z_{rj} - \bar{Z}_j) \text{ dan } S_{ij}^* = S_{ji}^* \text{ untuk } i, j = 1, 2, \dots, k \quad (2.26)$$

dengan rumus sebagai berikut:

$$r_{ij} = \frac{S_{ij}^*}{\sqrt{S_{ii}^* S_{jj}^*}} \quad (2.27)$$



Persamaan yang dijelaskan sesuai variabel-variabel  $X_1, X_2, \dots, X_k$  menggunakan matriks  $\mathbf{S}$  maka berlaku pada variabel  $Z_1, Z_2, \dots, Z_k$  dengan matriks  $\mathbf{R}$ .

Ketentuan dalam membentuk komponen utama yang merupakan kombinasi linier dari  $Z_1, Z_2, \dots, Z_k$  supaya memiliki varian maksimum ialah menggunakan *eigenvector* yaitu  $\mathbf{a}'_j = [\mathbf{a}_{1j}, \mathbf{a}_{2j}, \dots, \mathbf{a}_{kj}]$  dengan  $j = 1, 2, \dots, k$ , sedemikian sehingga  $\text{Var}(W_j) = \mathbf{a}'_j \mathbf{R} \mathbf{a}_j$  maksimum dengan kendala  $\mathbf{a}'_j \mathbf{a}_j = \mathbf{1}$ .

Dalam pemilihan *eigenvector* yaitu  $\mathbf{a}'_j = [\mathbf{a}_{1j}, \mathbf{a}_{2j}, \dots, \mathbf{a}_{kj}]$  dari  $\mathbf{R}$  berlaku  $(\mathbf{R} - \lambda \mathbf{I}) \mathbf{a}_j = \mathbf{0}$ . Namun, diperlukan nilai *eigenvalue* ( $\lambda$ ) dari  $\mathbf{R}$  dengan syarat  $|\mathbf{R} - \lambda \mathbf{I}| = 0$ .

Menurut Johnson & Wichern (2007), variabel dibakukan yaitu  $Z_1, Z_2, \dots, Z_k$  mampu menjelaskan keragaman total oleh komponen utama ke- $j$  sebagai berikut:

$$\frac{\lambda_j}{k} \times 100\% ; j = 1, 2, \dots, k \quad (2.28)$$

Terdapat beberapa cara untuk memilih komponen utama yang dipergunakan dalam regresi komponen utama yaitu:

1. Menentukan komponen-komponen utama yang memiliki kumulatif proporsi keragaman total 75%.
2. Memilih *eigenvalue* yang memiliki nilai  $\geq 1$ .
3. Mengamati *scree plot* yaitu plot antara *eigenvalue*  $\lambda_j$  dengan  $j$ .

Kemudian jika komponen utama telah diperoleh, maka langkah selanjutnya yaitu dengan menghitung skor komponen utama dari tiap-tiap pengamatan yaitu:

$$\text{SK-U}_{ji} = \mathbf{a}'_j \mathbf{Z}_i \text{ untuk } i = 1, 2, \dots, n \text{ dan } j = 1, 2, \dots, m \quad (2.29)$$

dimana

$\text{SK-U}_{ji}$  = skor komponen utama ke- $j$  untuk pengamatan ke- $i$

$\mathbf{a}'_j$  = *eigenvector* komponen utama ke- $j$

$\mathbf{Z}_i$  = vektor skor variabel baku yang diamati pada pengamatan ke- $i$

## 2.10 Regresi Komponen Utama

Menurut Notiragayu & Nisa (2008), Regresi Komponen Utama (RKU) adalah satu diantara metode yang bisa dipergunakan dalam mengatasi masalah multikolinearitas dengan dua tahapan, tahap pertama melakukan analisis komponen utama terhadap peubah bebas  $X$  yang kemudian untuk tahap selanjutnya ialah meregresikan komponen-komponen utama tersebut dengan peubah tak bebas  $Y$ .

Sesudah melakukan analisis komponen utama tahapan berikutnya yaitu meregresikan skor komponen utama yang telah terpilih dengan variabel terikat  $Y$  dengan metode *Ordinary Least Squares* (OLS).

Sehingga, persamaan regresi komponen utama dapat dituliskan sebagai berikut:

$$Y_i = \alpha_0 + \alpha_1 U_{1i} + \alpha_2 U_{2i} + \dots + \alpha_m U_{mi} + \gamma_i \quad (2.30)$$

dengan:

$Y_i$  = variabel terikat

$U_{ji}$  = variabel bebas, yaitu komponen utama terpilih

$\alpha_0$  = konstanta (intersep)

$\alpha_j$  = koefisien regresi

$\gamma_i$  = *error* random

$n$  = banyaknya pengamatan

$m$  = banyaknya komponen utama yang terpilih

Ataupun dapat dimuat dalam bentuk matriks sebagai berikut:

$$\mathbf{y} = \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\gamma} \quad (2.31)$$

dengan

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \mathbf{U} = \begin{bmatrix} 1 & U_{11} & U_{21} & \dots & U_{m1} \\ 1 & U_{12} & U_{22} & \dots & U_{m2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & U_{1n} & U_{2n} & \dots & U_{mn} \end{bmatrix}; \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{bmatrix}$$

dimana:

$\boldsymbol{\alpha}$  = vektor  $q \times 1$  dari koefisien regresi, dengan  $q = m+1$

$\mathbf{y}$  = vektor  $n \times 1$  dari variabel respon

$\mathbf{U}$  = matriks  $n \times q$  dari variabel prediktor (komponen utama terpilih)

$\boldsymbol{\gamma}$  = vektor  $n \times 1$  dari *error* random

Parameter  $\hat{\boldsymbol{\alpha}}$  diestimasi menggunakan metode *Ordinary Least Squares* (OLS) dengan rumus:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{U}'\mathbf{U})^{-1}(\mathbf{U}'\mathbf{y}) \quad (2.32)$$

### 2.11 *Minimum Covariance Determinant* (MCD)

Metode *Minimum Covariance Determinant* (MCD) ialah metode yang efisien dan *robust* dalam mengevaluasi dan mengatasi data yang terkontaminasi oleh *outlier* (Maronna *et al.*, 2006). MCD digunakan guna memperoleh penaksir yang *robust* dari rata-rata dan kovarian beberapa pengamatan yang mempunyai determinan matriks kovarian yang terkecil.

Menurut Rousseuw & Driessen (1999), penduga MCD adalah pasangan  $(\bar{\mathbf{X}}, \mathbf{S})$ , dimana  $\bar{\mathbf{X}}$  adalah vektor rata-rata dan  $\mathbf{S}$  adalah matriks kovarian yang meminimumkan nilai determinan  $\mathbf{S}$  pada subsampel yang berisikan tepat sebanyak  $h$  anggota dari  $n$  pengamatan, dimana nilai standar dari  $h = [(n+k+1)/2]$ .

Penduga MCD cukup mudah ditemukan jika  $n$  kecil. Tetapi, jika  $n$  besar, kombinasi subsampel yang harus ditemukan untuk mendapatkan penaksir MCD menjadi sangat banyak. Sehingga, untuk mengatasi kelemahan tersebut, Rousseuw & Driessen (1999) menemukan suatu algoritma yang baru untuk metode MCD yang disebut *fast-MCD*.

Metode *fast-MCD* merupakan metode yang digunakan untuk mengatasi kelemahan pada penaksir *robust* MCD yang telah disebutkan sebelumnya. Metode *fast-MCD* dilakukan dengan mengganti  $\bar{\mathbf{X}}$  dan  $\mathbf{S}$  dengan  $\bar{\mathbf{X}}_{MCD}$  dan  $\mathbf{S}_{MCD}$  yang merupakan vektor rata-rata dan matriks kovarian dengan metode *fast-MCD* pada penaksiran parameter untuk model umum analisis komponen utama. Salah satu teorema penting dalam *fast-MCD* adalah *C-steps*.

Misalkan  $n$  merupakan jumlah pengamatan dan  $k$  merupakan jumlah variabel, sehingga algoritma FAST-MCD menurut Rousseeuw & Driessen (1999) adalah sebagai berikut:

1. Memilih  $h = [(n+k+1)/2]$ , namun diperkenankan pula memilih  $h$  dengan  $[(n + k + 1)/2] \leq h \leq n$ .
2. Bila  $h = n$ , maka estimasi lokasi  $\bar{\mathbf{X}}_{MCD}$  ialah rata-rata dari himpunan data dan perkiraan  $\mathbf{S}_{MCD}$  ialah matriks kovariannya.
3. Pada  $h < n$  dan  $k \geq 2$ . Jika  $n$  kecil (misal  $n \leq 500$ ), maka
  - A. Bentuk  $h$  himpunan bagian  $H_1$  awal menggunakan cara:
    - 1) Mengambil acak  $(k+1)$  himpunan bagian  $J$  lalu menghitung  $\bar{\mathbf{X}}_0 =$  rata-rata ( $J$ ) dan  $\mathbf{S}_0 = \text{cov}(J)$  [Bila  $\det(\mathbf{S}_0) = 0$ , maka perluas  $J$  menggunakan penambahan satu per satu pengamatan acak lain sehingganya  $\det(\mathbf{S}_0) > 0$ ].
    - 2) Menghitung jarak  $d_0^2(i) = (\mathbf{X}_i - \bar{\mathbf{X}}_0)' \mathbf{S}_0^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}_0)$  untuk  $i = 1, 2, \dots, n$ . Susun ke dalam  $d_0(\pi(1)) \leq \dots \leq d_0(\pi(n))$  kemudian menempatkannya pada  $H_1 = \{ \pi(1), \dots, \pi(h) \}$
  - B. Melakukan *C-Steps* yakni menggunakan langkah sebagai berikut:
    - 1) Menghitung  $\bar{\mathbf{X}}_1 = \frac{1}{h} \sum_{i \in H_1} \mathbf{X}_i$
    - 2) Menghitung  $\mathbf{S}_1 = \frac{1}{h} \sum_{i \in H_1} (\mathbf{X}_i - \bar{\mathbf{X}}_1)(\mathbf{X}_i - \bar{\mathbf{X}}_1)'$
    - 3) Apabila  $\det(\mathbf{S}_1) \neq 0$ , maka jarak relatif dihitung menggunakan rumus jarak mahalanobis, sebagai berikut:
 
$$d_1^2(i) = (\mathbf{X}_i - \bar{\mathbf{X}}_1)' \mathbf{S}_1^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}_1) \text{ untuk } i = 1, 2, \dots, n$$
    - 4) Susun secara runtut ke dalam  $d_1(\pi(1)) \leq \dots \leq d_1(\pi(n))$
    - 5) Letakkan dalam  $H_2 = \{ \pi(1), \dots, \pi(h) \}$
    - 6) Menghitung  $\bar{\mathbf{X}}_2 =$  rata-rata ( $H_2$ ) dan  $\mathbf{S}_2 = \text{cov}(H_2)$
- C. Mengulangi *C-Steps* sehingganya didapatkan  $\det(\mathbf{S}_1) \geq \det(\mathbf{S}_2) \geq \dots \geq \det(\mathbf{S}_{m-1}) \geq \det(\mathbf{S}_m)$  yang bernilai nonnegatif.
- D. Perulangan berhenti jika  $\det(\mathbf{S}_m) = 0$  atau  $\det(\mathbf{S}_m)$  konvergen dengan  $\det(\mathbf{S}_{m-1})$ .
- E. Dapatkan solusi  $(\bar{\mathbf{X}}, \mathbf{S})$  dengan  $\det(\mathbf{S})$  terkecil.

## 2.12 Least Trimmed Square (LTS)

Metode *Least Trimmed Squares* (LTS) pertama kali diperkenalkan oleh Rousseeuw pada tahun 1984 sebagai metode alternatif untuk mengatasi kelemahan metode *Ordinary Least Squares* (OLS). LTS adalah metode robust yang berguna dalam analisis data multivariat yang sensitif terhadap adanya data *outlier*. Metode ini mengurangi pengaruh *outlier* dalam proses estimasi dan memberikan hasil yang lebih stabil (Hubert *et al.*, 2018)

Menurut Rousseeuw & Leroy (1987), LTS merupakan metode pendugaan parameter regresi *robust* yang bertujuan meminimumkan jumlah kuadrat  $h$  residual (fungsi objektif). Dengan fungsi objektif dari metode LTS adalah sebagai berikut:

$$Q = \sum_{i=1}^h e_{(i:n)}^2 \quad (2.33)$$

dengan

$$h = [(n+k+1)/2] \text{ atau } h = \gamma * n \text{ dimana } \gamma = (1 - \alpha)$$

$\gamma$  = persentase besar data yang akan dipangkas (*trimmed*)

$\alpha$  = nilai breakdown (persentase banyaknya pencilan terhadap seluruh data)

$$e_i = (Y_i - \hat{Y}_i)$$

$e_1^2 \leq e_2^2 \leq \dots \leq e_n^2$  : kuadrat residual disusun dari terkecil ke terbesar

$n$  = banyaknya pengamatan

$p$  = banyaknya parameter

Jumlah  $h$  menunjukkan sejumlah subset data dengan kuadrat residual terkecil. Menurut Rousseeuw & Driessen (1999), estimasi parameter LTS dilakukan dengan menggabungkan algoritma FAST-LTS dan *C-Steps* seperti berikut:

1. Menghitung estimasi awal parameter regresi yaitu  $\hat{\beta}^0$ .
2. Menghitung residual  $e_0(i)$  untuk  $i = 1, 2, \dots, n$ .
3. Menghitung residual  $e_0^2(i)$  untuk  $i = 1, 2, \dots, n$ .
4. Susun secara runtut ke dalam  $e_0^2(\pi(1)) \leq \dots \leq e_0^2(\pi(n))$  dan letakkan pada  $H_1 = \{\pi(1), \dots, \pi(h)\}$  dengan  $h = \gamma * n$
5. Melakukan *C-Steps* yaitu dengan cara:
  - a. Menghitung  $\hat{\beta}^1$ , yaitu estimasi parameter regresi dari  $H_1$ .
  - b. Menghitung residual  $e_1(i)$  untuk  $i = 1, 2, \dots, n$ .

- c. Menghitung residual  $e_1^2(i)$  untuk  $i = 1, 2, \dots, n$ .
- d. Menghitung  $Q_1 = \sum_{i \in H_1} e_1^2(i)$ .
- e. Susun secara runtut ke dalam  $e_1^2(\pi(1)) \leq \dots \leq e_1^2(\pi(n))$  dan letakkan dalam  $H_2 = \{\pi(1), \dots, \pi(h)\}$  dengan  $h = \gamma * n$ .
- f. Menghitung  $\hat{\beta}^2$ , yaitu estimasi parameter regresi dari  $H_2$ .
6. Mengulagi *C-Steps* sehingga didapatkan  $Q_1 \geq Q_2 \geq \dots \geq Q_{m-1} \geq Q_m$  yang bernilai nonnegatif.
7. Perulangan berhenti ketika  $Q_m \leq Q_{m-1}$  dan nilai estimasi parameter regresi  $\hat{\beta}^m$  konvergen dengan  $\hat{\beta}^{m-1}$ .
8. Dapatkan  $\hat{\beta}$  dengan fungsi objektif  $Q$  terkecil.

### **III. METODOLOGI PENELITIAN**

#### **3.1 Waktu dan Tempat Penelitian**

Penelitian ini dilakukan pada semester ganjil 2021/2022, di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

#### **3.2 Data Penelitian**

Data yang digunakan dalam penelitian ini merupakan data real yakni data sekunder yang didapatkan melalui publikasi Badan Pusat Statistik Provinsi Sumatera Utara tahun 2021. Data sekunder yang digunakan adalah data Indeks Pembangunan Manusia (Y) yang dipengaruhi tujuh variabel bebas yaitu tingkat partisipasi angkatan kerja ( $X_1$ ), presentase penduduk miskin ( $X_2$ ), penduduk umur 15 tahun keatas yang bekerja ( $X_3$ ), angka partisipasi sekolah ( $X_4$ ), jumlah fasilitas sekolah ( $X_5$ ), banyaknya sarana kesehatan ( $X_6$ ) dan angka harapan hidup ( $X_7$ ).

#### **3.3 Metodologi Penelitian**

Penelitian ini dilakukan secara studi pustaka yang diperoleh dari buku-buku penunjang, jurnal dan juga media lain seperti internet. Guna memudahkan perhitungan dan didapatkan hasil penelitian yang tepat dan akurat maka digunakan

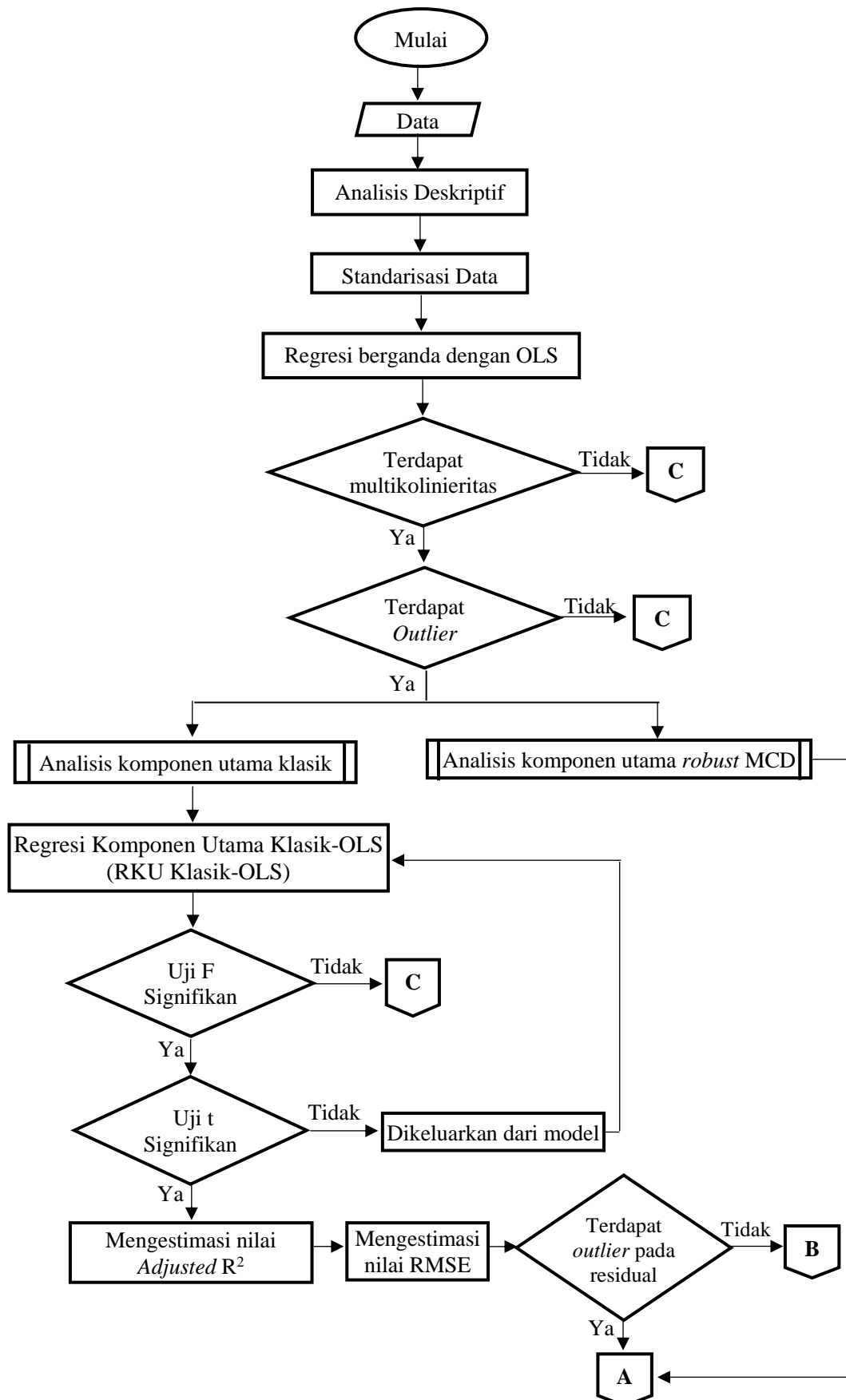
*software* R.studio 4.2.0. Adapun langkah-langkah yang dilakukan adalah sebagai berikut:

1. Melakukan analisis deskriptif pada data.
2. Melakukan standarisasi data.
3. Melakukan uji asumsi non-multikolinearitas dengan cara memeriksa nilai VIF serta juga melihat nilai koefisien korelasi pearson antar variabel bebas.
4. Melakukan estimasi parameter regresi menggunakan metode OLS.
5. Melakukan pendeteksian outlier pada variabel bebas menggunakan jarak mahalanobis.
6. Melakukan analisis komponen utama klasik pada variabel bebas dengan langkah sebagai berikut:
  - a) Menghitung penaksir klasik, yaitu matriks korelasi antar variabel bebas yang telah distandarisasikan.
  - b) Mengestimasi nilai *eigenvalue* dan *eigenvector*.
  - c) Menetapkan komponen-komponen utama yang nantinya akan dipakai, yakni komponen-komponen utama yang memiliki  $eigenvalue \geq 1$
  - d) Mengestimasi proporsi kumulatif varian yang bisa dijelaskan melalui komponen-komponen utama yang telah terpilih.
  - e) Menghitung skor komponen utama.
7. Jika terdapat *outlier*, maka dilakukan analisis komponen utama *robust*, dengan langkah sebagai berikut:
  - a) Menghitung penaksir *robust*, yaitu matriks kovarian antar variabel bebas yang telah distandarisasikan menggunakan metode *C-Step* dengan algoritma FAST-MCD.
  - b) Mengestimasi nilai *eigenvalue* dan *eigenvector*.
  - c) Menetapkan komponen-komponen utama yang akan dipakai, yakni komponen utama yang memiliki  $eigenvalue \geq 1$
  - d) Menghitung proporsi kumulatif varian yang dapat dijelaskan melalui komponen-komponen utama yang telah terpilih sebelumnya.
  - e) Mengestimasi skor komponen utama.

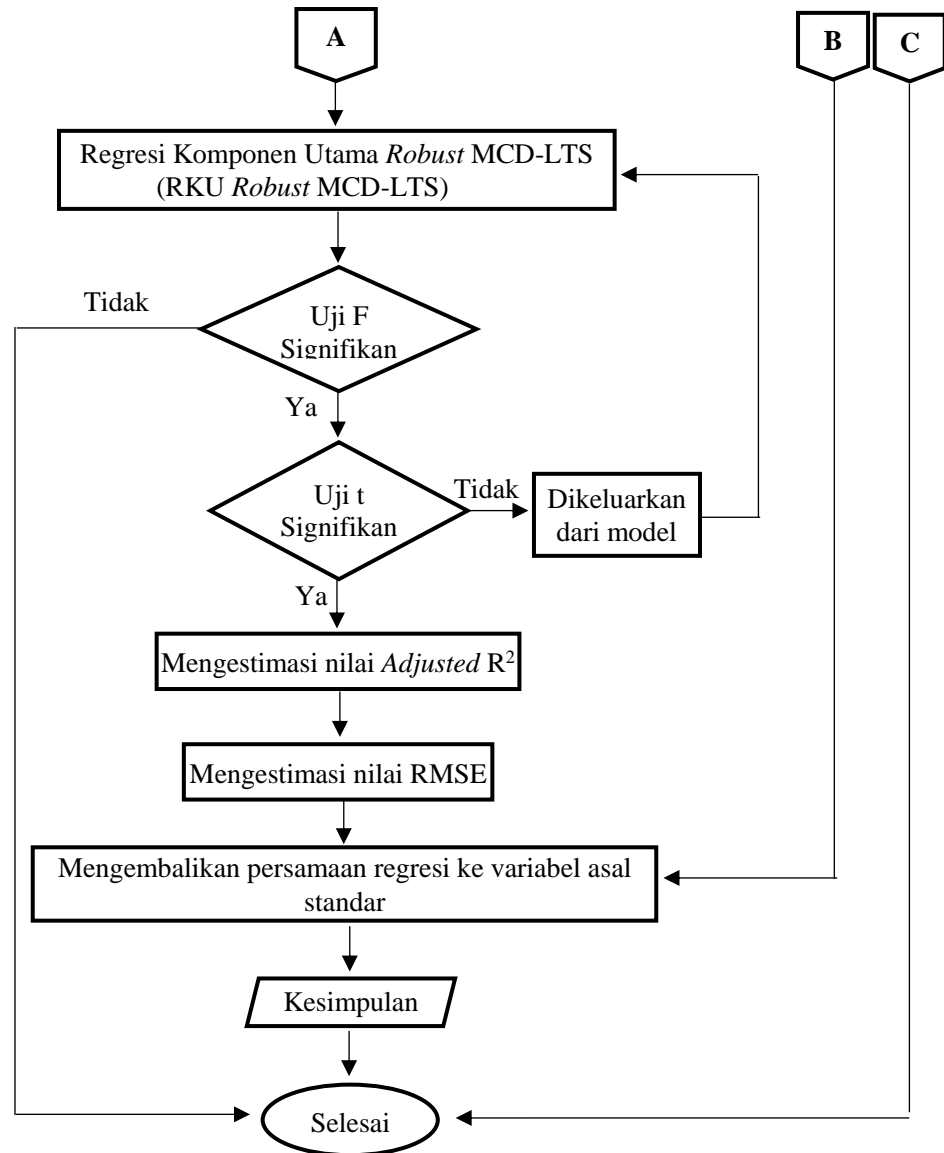


8. Meregresikan komponen-komponen utama terpilih dari AKU klasik dengan variabel terikat menggunakan metode OLS, sehingga didapatkan model RKU Klasik-OLS.
9. Melakukan pengujian pada model regresi menggunakan uji F simultan.
10. Melakukan pengujian koefisien regresi parsial menggunakan uji t parsial.
11. Mendeteksi masalah multikolinieritas dan didapatkan hasil bahwa kasus multikolinieritas telah teratasi.
12. Mengestimasi nilai *Adjusted R<sup>2</sup>*.
13. Mengestimasi nilai RSE.
14. Mendeteksi adanya *outlier* menggunakan metode *DFFITs*.
15. Jika terdapat *outlier*, maka meregresikan komponen-komponen utama yang telah terpilih dari AKU *robust* dengan variabel terikat menggunakan metode LTS. Maka, didapatkan model RKU *Robust* MCD-LTS.
16. Melakukan pengujian pada model regresi menggunakan uji F simultan.
17. Melakukan pengujian koefisien regresi parsial menggunakan uji t parsial.
18. Mengestimasi nilai *Adjusted R<sup>2</sup>*.
19. Mengestimasi nilai RMSE.
20. Memulihkan persamaan regresi dalam bentuk variabel asal standar.
21. Membuat kesimpulan.

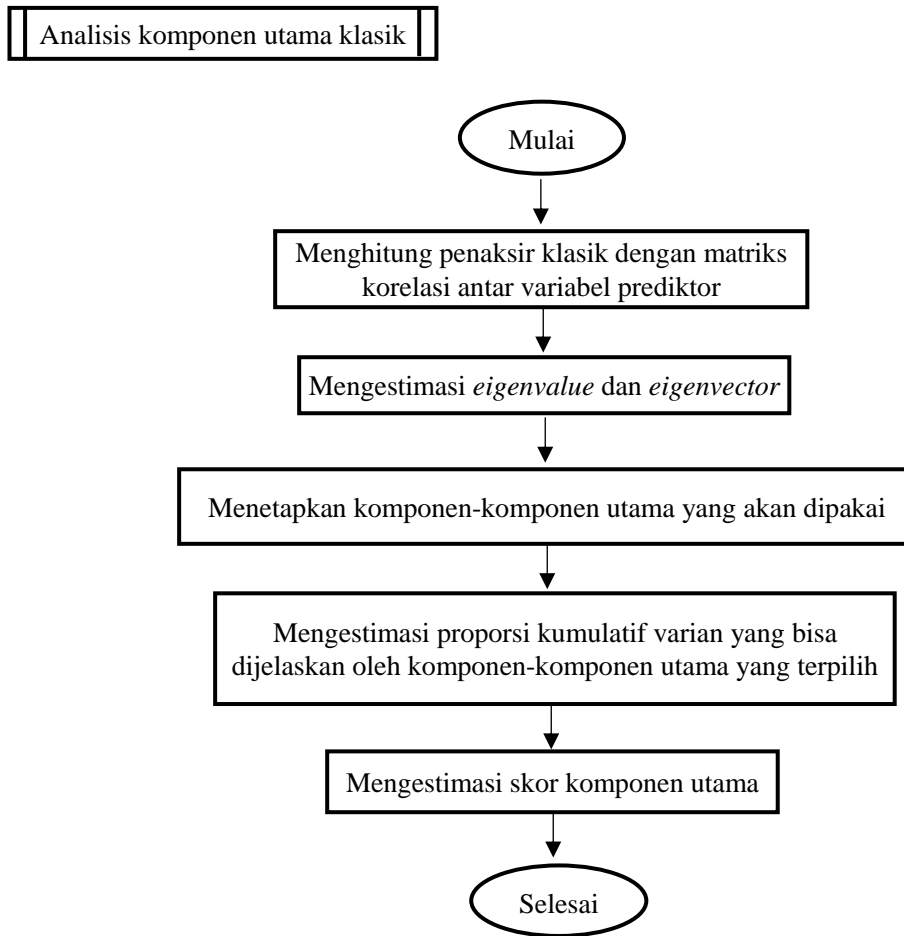
### 3.4 Diagram Alir (Flowchart)



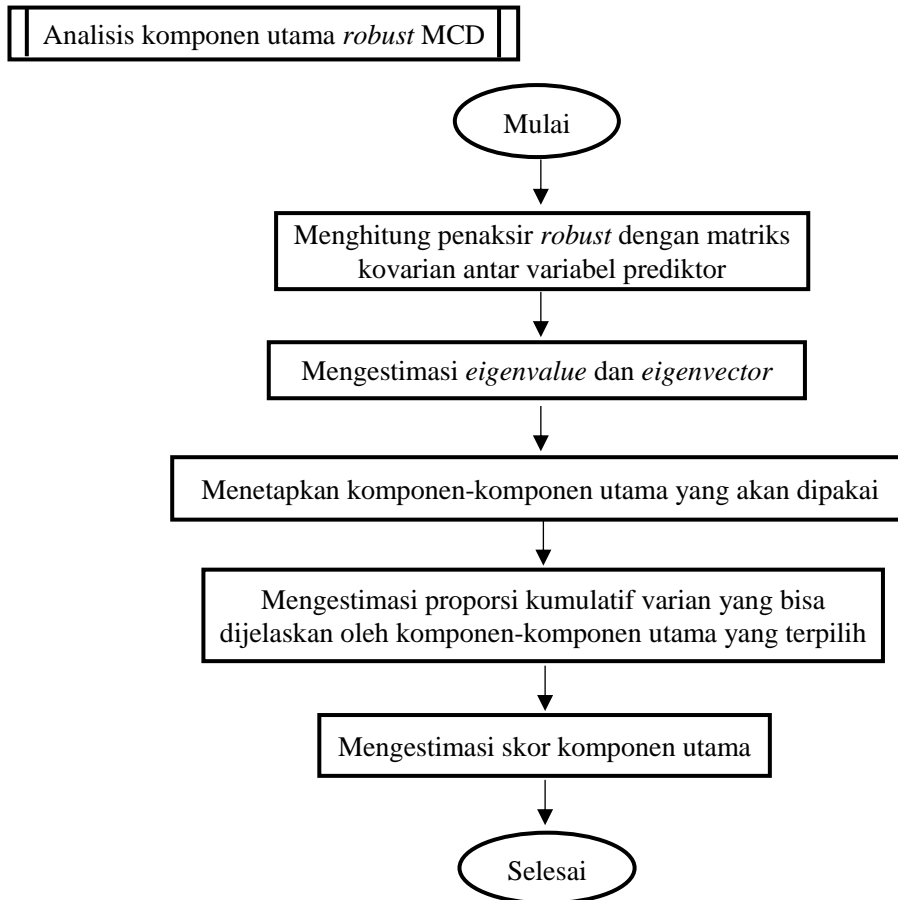
Gambar 1a. Diagram alir.



Gambar 1b. Diagram alir (Lanjutan).



Gambar 2. Diagram alir Analisis Komponen Utama Klasik.



Gambar 3. Diagram alir Analisis Komponen Utama *Robust* MCD.

## V. KESIMPULAN

Berdasarkan hasil penelitian, maka dapat disimpulkan bahwa model Regresi Komponen Utama *Robust* dengan menggunakan metode *Minimum Covariance Determinant* (MCD) yang diregresikan menggunakan metode *Least Trimmed Squares* (LTS) yaitu :

$$\hat{Q} = -0.1736 Z_1 - 0.0474 Z_2 - 0.0204 Z_3 + 0.0465 Z_4 - 0.0807 Z_5 - \\ 0.0157 Z_6 + 0.4275 Z_7$$

Dengan nilai *Adjusted R<sup>2</sup>* nya 0.5573 dan nilai RMSE 0.4339 dinilai lebih efektif dalam mengatasi masalah multikolinearitas dan *outlier* pada Data Indeks Pembangunan Manusia (IPM) provinsi Sumatera Utara.

## DAFTAR PUSTAKA

- Faizia, T., Prahutama, A., & Yasin, H. 2019. Pemodelan Indeks Pembangunan Manusia di Jawa Tengah dengan Regresi Komponen utama *Robust. JurnalGaussian*. 8(2):253-271.
- Field, A. 2013. *Discovering Statistics Using IBM SPSS Statistics*. 4<sup>th</sup> Edition. Sage Publications.
- Ghorbani, H. 2019. Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*. 1(1):583-595.
- Gujarati, D. & Porter, D.C. 2011. *Dasar-Dasar Ekonometrika*. Mangunsong, R.C. Ed. Ke-5. Salemba Empat, Jakarta.
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. 2014. *Multivariate Data Analysis*. 7<sup>th</sup> Edition. Pearson Education.
- Hubert, M., Rousseeuw, P.J., & Van Aelst, S. 2018. *Minimum Covariance Determinant and Extensions (Chapter 3)*. In S. Chretien, É. Dufour, & G. Saporta (Eds.), *Robust Statistics*. 2<sup>nd</sup> Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New Jersey.
- Johnson, R.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*. 6<sup>th</sup> Edition. Pearson Prentice Hall., New Jersey.
- Jolliffe, E.L. 2002. *Principal Component Analysis*. 2<sup>nd</sup> Edition. Springer-Verlag, Inc., New York

- Kutner M.H., Nachtsheim, C.J., Neter, J., & Li, W. 2005. *Applied Linear Statistical Models*. 5<sup>th</sup> Edition. McGraw-Hill Companies, Inc., New York.
- Larasati, S.D.A., Nisa.K., & Setiawan, E. 2020. Analisis Regresi Komponen Utama Robust dengan Metode *Minimum Covariance Determinant Least Trimmed Square* (MCD-LTS). *Jurnal Siger Matematika*. **1**(1):1-5.
- Maronna, R.A., Martin, R.D., & Yohai, V.J. 2006. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New Jersey.
- Montgomery, D.C., Peck, E.A., & Vining, G.G. 2012. *Introduction to Linear Regression Analysis*. 5<sup>th</sup> Edition. John Wiley & Sons, Inc., New Jersey.
- Nisa, K. 2006. Analisis Regresi *Robust* Menggunakan Metode *Least Trimmed Square* untuk Data mengandung Pencilan. *Jurnal Ilmiah MIPA*. **IX**(2):1-9.
- Notiragayu & Nisa, K. 2008. Analisis Regresi Komponen Utama *Robust* untuk Data Mengandung Pencilan. *Jurnal Sains MIPA*. **14**(1):45-50.
- Rousseeuw, P.J. & Driessen, K.V. 1999. A Fast Algorithm for The Minimum Covariance Determinant Estimator. *Technometrics*. **41**(3):212-223.
- Rousseeuw P.J. & Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, Inc.
- Sembiring, R.K. 2003. *Analisis Regresi*. Ed. Ke-2. Institut Teknologi Bandung, Bandung.
- Tobin, J. 1958. Estimation of Relationships for Limited Dependent Variables. *Econometrica*. **26**(1):24-36
- Widiharih, T. 2001. Penanganan Multikolinieritas (Kekolinieran Ganda) dengan Analisis Regresi Komponen Utama. *Jurnal Matematika dan Komputer*. **4**(2): 71-81.