MENGATASI MASALAH MULTIKOLINEARITAS DAN *OUTLIER* DENGAN METODE *ROBUST* PCA PADA MODEL REGRESI LINEAR BERGANDA

Skripsi

Disusun untuk melengkapi syarat-syarat

guna memperoleh gelar Sarjana Sains

YULIFIRDA ISNAINI

3125121977

PROGRAM STUDI MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI JAKARTA

2017

ABSTRACT

YULIFIRDA ISNAINI, 3125121977. Solve Multicolinearity and Outlier Problems using Robust PCA Methods on Multiple Linear Regression Models. Thesis. Faculty of Mathematics and Natural Science Jakarta State University. 2017.

The existence of multicolinearity and outlier in multiple linear regression model can disturb data analysis process, multicolinearity and outlier detection are very important to do. When the data is used in multiple linear regression models, the detection becomes difficult to do. The Robust PCA method, which is a combination of Principal Component Analysis (PCA) and Minimum Covariance Determinant (MCD) techniques, can solve multicolinearity and outlier problems in multiple linear regression models. Before doing outliers detection, it is necessary to reduce the dimensions of data, which using PCA. In this study, the outlier is detected using the MCD Method. The purpose MCD Method is to obtain a subset of all observations whose variance-covariance matrix has the smallest determinant of all possible combinations of data. Outliers detection using MCD Method is based on Robust Distance and cut-off value. An observation can be detected as an outlier when the Robust Distance is greater than the cut-off value. While to classify the outlier is done by making a diagnostic plot of Mahalanobis Distance versus Robust Distance. And then a principal component regression analysis is performed, to solve multicolinearity and to get multiple linear regression model. Therefore based on using Robust PCA Method on customer satisfaction level data, it obtains VIF value that every principal component is 1, from diagnostic plot outliers have been solved, and R^2 value has been increased becoming 66,1%.

Keywords *: Multicollinearity, Outlier, Principal Component Analysis, Minimum Covariance Determinant, Robust PCA.*

ABSTRAK

YULIFIRDA ISNAINI, 3125121977. Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Metode *Robust* PCA pada Model Regresi Linear Berganda. Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta. 2017.

Keberadaan multikolinearitas dan *outlier* pada model regresi linear berganda mengganggu proses analisis data, sehingga pendeteksian multikolinearitas dan *outlier* sangat penting untuk dilakukan. Ketika data yang digunakan merupakan data pada model regresi linear berganda, maka pendeteksian tersebut menjadi sulit untuk dilakukan. Metode *Robust* PCA merupakan gabungan teknik Analisis Komponen Utama (AKU) dan *Minimum Covariance Determinant* (MCD) yang dapat mengatasi masalah multikolinearitas dan *outlier* pada model regresi linear berganda. Sebelum melakukan pendeteksian *outlier* menggunakan metode MCD, perlu dilakukan pereduksian dimensi data dengan metode AKU. Prinsip Metode MCD adalah mendapatkan subhimpunan dari keseluruhan pengamatan yang matriks varian kovariannya memiliki determinan terkecil diantara semua kombinasi kemungkinan data. Pendeteksian *outlier* dengan Metode MCD dilakukan berdasarkan jarak *robust* dan nilai *cut-off* nya. Suatu pengamatan terdeteksi sebagai *outlier* ketika jarak *robust* lebih besar dari nilai *cut-off*. Sedangkan untuk mengklasifikasikan *outlier* dilakukan dengan cara membuat *diagnostic plot* Jarak *Mahalanobis* versus Jarak *Robust*. Selanjutnya analisis regresi komponen utama dilakukan, untuk mengatasi multikolinearitas dan untuk mendapatkan model regresi linear berganda. Penerapan Metode *Robust* PCA pada data tingkat kepuasan pelanggan menghasilkan nilai VIF pada setiap komponen utama 1, dari *diagnostic plot outlier* sudah teratasi, dan nilai R^2 meningkat menjadi 66,1%.

Kata kunci : Multikolinearitas, *Outlier*, Analisis Komponen Utama, *Minimum Covariance Determinant*, *Robust* PCA.

PERSEMBAHANKU...

"*Sesungguhnya sesudah kesulitan itu ada kemudahan, maka apabila kamu telah selesai dari suatu urusan, kerjakanlah dengan sungguh-sungguh urusan yang lain, dan hanya kepada Tuhanmulah hendaknya kamu berharap.*"

-(Q.S Al-Insyirah: 6-8)

Skripsi ini kupersembahkan untuk Ayah, Ibu, Kak Uut, dan Rafi.

"Terima kasih atas do'a, dukungan, serta kasih sayang kalian".

KATA PENGANTAR

Puji syukur kepada Allah SWT atas pengetahuan dan kemampuan sehingga penulis dapat menyelesaikan skripsi yang berjudul "Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Metode *Robust* PCA pada Model Regresi Linear Berganda" yang merupakan salah satu syarat dalam memperoleh gelar Sarjana Program Studi Matematika Universitas Negeri Jakarta.

Skripsi ini berhasil diselesaikan tidak terlepas dari adanya bantuan dari berbagai pihak. Oleh karena itu, dalam kesempatan ini penulis ingin menyampaikan terima kasih terutama kepada:

1. Ayah dan Ibu yang selalu mendukung, memberi do'a dan motivasi, serta setia membantu penulis dengan penuh cinta dan kasih sayang yang tulus.

2. Ibu Dra. Widyanti Rahayu, M.Si. selaku Dosen Pembimbing I dan Ibu Vera Maya Santi, M.Si. selaku Dosen Pembimbing II, yang telah meluangkan waktunya dalam memberikan bimbingan, saran, nasehat serta arahan sehingga skripsi ini dapat menjadi lebih baik dan terarah.

3. Ibu Dr. Lukita Ambarwati, S.Pd, M.Si., selaku Koordinator Prodi Matematika FMIPA UNJ yang telah banyak membantu penulis.

4. Ibu Ratna Widyati, S.Si, M.Kom., selaku Pembimbing Akademik atas segala bimbingan dan kerja sama Ibu selama perkuliahan, dan seluruh Bapak/Ibu dosen atas pengajarannya yang telah diberikan, serta karyawan/karyawati FMIPA UNJ yang telah memberikan informasi yang penulis butuhkan dalam menyelesaikan skripsi.

5. Kakakku Agustin Siti Nur Utami dan adikku Arrafi Naufal Fikri yang terus memberi semangat, mendoakan penulis, dan selalu menghibur ketika penulis mengalami kesulitan dalam penulisan skripsi ini.

6. Teman kesayangan Ibeth, Vinna, Anggita, Mega, Astrid, Jennyfer, Farchatun, dan Faralita sebagai dosen pembimbing ke-3 yang telah banyak membantu, memotivasi, dan menyemangati penulis, serta selalu ada dikala senang maupun susah.

7. Sahabatku Erna, Yuni, Ganis, Mawasumi yang telah memotivasi, menyemangati, dan ada dikala senang maupun susah.

8. Teman2 Matematika 2012 Heru, Dewanti, Hengki, Ziezie, Leny, Zuhai, Alphien, Chrisna, Mei, Fatmah, Mella, Yarham, Mira, Deddy, Dwi, Bety, Uyun, Jaja, Bobby, Sharah, Icha, Steven, Sidik, Habib, Mukti, Ela, Yohana, Aan, Irma, Lusia, Timah, Miqdad, Tyo untuk bantuan dan kebersamaannya.

9. Adik2 dan kakak2 tingkat di matematika terutama Eza mtk 2013, Kak Idam, dll yang memberi motivasi dan semangat untuk penulis, juga sebagai teman seperjuangan penulis.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Masukan dan kritikan akan sangat berarti. Semoga skripsi ini dapat bermanfaat bagi pembaca sekalian.

Jakarta, Agustus 2017

Yulifirda Isnaini

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Cara untuk mengetahui pengaruh beberapa variabel bebas terhadap variabel terikat dapat menggunakan beberapa metode analisis data dalam statistika, salah satunya adalah analisis regresi. Jika hubungan antara variabel terikat Y dan variabel bebas X adalah linier dan hanya terdapat satu variabel bebas X, maka model regresi yang digunakan adalah model regresi linier sederhana, sedangkan jika hubungan antara variabel terikat Y dan variabel bebas X adalah linier dan terdapat lebih dari satu variabel bebas $(X_1, X_2, \ldots, X_n)$, maka model regresi yang digunakan adalah model regresi linier berganda. Tujuan dari analisis regresi linier berganda adalah mengetahui seberapa besar pengaruh dari beberapa variabel bebas terhadap variabel terikat dan juga dapat meramal nilai variabel terikat jika seluruh variabel bebas sudah diketahui nilainya.

Model regresi linier berganda didapat dengan melakukan proses analisis data, diantaranya estimasi terhadap parameter - parameternya menggunakan metode tertentu. Salah satu metode yang dapat digunakan untuk mengestimasi parameter model regresi linier berganda adalah dengan metode kuadrat terkecil (MKT). Metode kuadrat terkecil harus memenuhi beberapa asumsi klasik, diantaranya variabel berdistribusi normal, homoskedastisitas sama untuk setiap observasi, tidak ada autokorelasi antar kesalahan pengganggu, tidak ada multiko-

linieritas diantara variabel bebas, serta tidak adanya pencilan (*outlier*).

Asumsi yang harus dipenuhi untuk melakukan pengujian hipotesis terhadap parameter pada analisis regresi linier berganda, diantaranya tidak terjadi multikolinearitas diantara variabel. Multikolinearitas terjadi karena adanya hubungan di antara variabel-variabel bebasnya. Multikolinearitas yang terdapat pada variabel-variabel bebas mengakibatkan model regresi yang diperoleh jauh dari akurat.

Pada data multikolinearitas kadang terdapat adanya *outlier*. *Outlier* didefinisikan sebagai sebagian dari data pengamatan yang memiliki pola yang berbeda dari sebagian besar data pengamatan (Hadi, 2009). *Outlier* pada data dapat menyebabkan ketidakhomogenan matriks varian kovarian. Penelitian yang telah dilakukan untuk menangani masalah *outlier* adalah Perbandingan Regresi *Robust* Penduga MM dengan Metode *Random Sample Consensus* dalam Menangani Pencilan (Irfagutami, 2014), Analisis Regresi pada Data *Outlier* dengan menggunakan *Least Trimmed Square* (LTS) dan MM-Estimasi (Nurcahyadi, 2010), dan Analisis Regresi *Robust* dengan Menggunakan Metode Penduga-M (Ridwan, 2008).

Principal Component Analysis (PCA) yang berdasarkan matriks varian kovarian sangat sensitif terhadap adanya pencilan pada data pengamatan, sehingga untuk mengatasi masalah pencilan diperlukan suatu metode penduga yang tegar terhadap pencilan. *Robust* PCA adalah suatu metode yang kuat (*robust*) untuk PCA terhadap keberadaan pencilan pada data yang mengandung multikolinearitas. Untuk mendapatkan komponen utama yang *robust* diperlukan *estimator covariance robust*, yaitu *robust Minimum Covariance Determinant* (MCD). Sunaryo (2011) pernah meneliti *Robust* PCA dalam jurnal Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Pendekatan ROBPCA (Studi Kasus: Angka Kematian Bayi di Jawa Timur).

Kelebihan Metode *Robust* PCA dibandingkan dengan metode lain dalam mengatasi *outlier*, yaitu dalam Metode *Robust* PCA dapat mengatasi dua masalah statistik sekaligus yaitu multikolinearitas dan *outlier*. Berdasarkan hal di atas, maka penulis akan mengkaji Metode *Robust* PCA untuk mengatasi multikolinearitas dan *outlier* pada model regresi linear berganda.

1.2 Perumusan Masalah

Berdasarkan latar belakang di atas, perumusan masalah yang akan dikaji pada skripsi ini adalah bagaimana cara mengatasi multikolinearitas dan *outlier* dengan Metode *Robust* PCA pada model regresi linear berganda?

1.3 Pembatasan Masalah

Masalah yang dibahas pada skripsi ini dibatasi pada penyelesaian masalah multikolinearitas dan *outlier* dengan menggunakan Metode *Robust* PCA dengan asumsi:

1. Pembahasan *estimator covariance robust* pada Metode *Robust* PCA difokuskan pada *Minimum Covariance Determinant* (MCD).

2. Nilai $30 < n \leq 600$, dimana $n =$ jumlah observasi.

1.4 Tujuan Penulisan

Tujuan yang ingin dicapai dalam skripsi ini adalah:

1. Mengkaji Metode *Robust* PCA sebagai alternatif lain dalam mengatasi masalah multikolinearitas dan *outlier* pada model regresi linear berganda,

2. Menerapkan model terbaik dari Metode *Robust* PCA pada data yang mengandung multikolinearitas dan *outlier* pada model regresi linear berganda.

1.5 Manfaat Penulisan

Manfaat yang diharapkan dari skripsi ini adalah:

1. Diharapkan bahwa Metode *Robust* PCA dapat menjadi suatu alternatif metode yang cukup baik dan efektif dalam mengatasi multikolinearitas dan *outlier* pada model regresi linear berganda,

2. Memberi wawasan kepada penulis dan pembaca tentang Metode *Robust* PCA,

3. Skripsi ini dapat dijadikan acuan untuk penulisan selanjutnya yang berkaitan dengan penggunaan Metode *Robust* PCA.

1.6 Metode Penelitian

Metode penelitian yang digunakan dalam pembuatan skripsi ini merupakan kajian teori, yaitu dengan mempelajari referensi yang berhubungan dengan Metode *Robust* PCA. Pembahasan yang diberikan merupakan hasil dari mempelajari buku-buku, jurnal-jurnal dengan jurnal utama Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Pendekatan ROBPCA (Studi Kasus: Angka Kematian Bayi di Jawa Timur (Sunaryo, 2011), diktat, dan situs matematika.

BAB II

LANDASAN TEORI

Pada bab ini akan dijelaskan landasan teori apa saja yang mendasari Metode *Robust* PCA, yaitu analisis regresi linear berganda, pembahasan tentang multikolinearitas dan *outlier*, pembentukan komponen utama dengan matriks kovarians, Jarak *Mahalanobis*, dan pengantar materi regresi *robust* yang fokus pada Metode *Robust* PCA.

2.1 Analisis Regresi Linier Berganda

2.1.1 Model Regresi Linier Berganda

Analisis regresi linier berganda adalah analisis regresi linier dengan menggunakan lebih dari satu variabel bebas dan mempunyai hubungan linier dengan variabel terikat. Model regresi linier berganda adalah:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Dalam notasi matriks menjadi:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

dengan:

$\mathbf{Y}$ = vektor berukuran $n \times 1$ dimana elemen-elemennya adalah nilai-nilai variabel bebas

$\mathbf{X}$ = matriks rancangan yang berukuran $n \times p$

$\boldsymbol{\beta}$ = vektor yang berukuran $p \times 1$ yang elemen-elemennya adalah parameter (koefisien) regresi

$\boldsymbol{\varepsilon}$ = vektor galat yang berukuran $n \times 1$

Dengan asumsi:

1. Variansi dari variabel gangguan (*error*) ε_i adalah konstan (homoskedastisitas):

$$\mathrm{Var}(\varepsilon_i|X_i) = \mathrm{E}[\varepsilon_i - \mathrm{E}(\varepsilon_i|X_i)]^2 = \mathrm{E}(\varepsilon_i^2|X_i) = \sigma^2.$$

2. Tidak ada serial korelasi antara variabel gangguan (*error*) ε_i atau variabel gangguan (*error*) ε_i tidak saling berhubungan dengan ε_i yang lain:

$$\begin{aligned} \mathrm{Cov}(\varepsilon_i, \varepsilon_j|X_i, X_j) &= \mathrm{E}[\varepsilon_i - \mathrm{E}(\varepsilon_i|X_i)][\varepsilon_j - \mathrm{E}(\varepsilon_j|X_j)] \\ &= \mathrm{E}(\varepsilon_i|X_i)\mathrm{E}(\varepsilon_j|X_j) = 0 \end{aligned}$$

3. Variabel gangguan (*error*) ε_i berdistribusi normal:

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2).$$

2.1.2 Pendugaan Parameter $\boldsymbol{\beta}$

Pendugaan parameter regresi $\boldsymbol{\beta}$ menggunakan metode kuadrat terkecil berdasarkan model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ adalah dengan meminimumkan jumlah kuadrat galat (JKG) dengan rumus:

$$\text{JKG} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

Jika $\mathbf{X}$ adalah matriks rancangan yang berukuran $n \times (p+1)$ yang bersifat full rank dengan $p + 1 \leq n$, maka penduga kuadrat terkecil untuk $\boldsymbol{\beta}$ adalah

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Penduga $\mathbf{b}$ tersebut ditentukan dengan meminimumkan jumlah kuadrat galat, dapat ditunjukan sebagai berikut: Berdasarkan $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, penduga bagi vektor galat $\boldsymbol{\varepsilon}$ dapat ditulis dengan $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$ sehingga jumlah kuadrat galat dapat dinyatakan sebagai berikut:

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} &= (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\
&= (\mathbf{Y}' - \mathbf{b}'\mathbf{X}')(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}
\end{aligned}$$

Karena $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ matriks berukuran 1×1, $\mathbf{b}'\mathbf{X}'\mathbf{Y} = (\mathbf{b}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\mathbf{b}$, maka

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \\
&= \mathbf{Y}'\mathbf{Y} - 2(\mathbf{X}'\mathbf{Y})'\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \qquad (2.1)
\end{aligned}$$

Untuk memperoleh $\mathbf{b}$ yang menyebabkan $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ minimum, persamaan (2.1) diturunkan terhadap $\mathbf{b}$, kemudian disamakan dengan nol, maka hasilnya adalah:

$$\begin{aligned}\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \mathbf{b}} &= -2\mathbf{X}'\mathbf{Y} + (\mathbf{X}'\mathbf{X})\mathbf{b} + (\mathbf{X}'\mathbf{X})'\mathbf{b} \\ &= -2\mathbf{X}'\mathbf{Y} + 2(\mathbf{X}'\mathbf{X})\mathbf{b}\end{aligned}$$

Karena $\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \mathbf{b}} = 0$, maka akan diperoleh persamaan

$$-2\mathbf{X}'\mathbf{Y} + 2(\mathbf{X}'\mathbf{X})\mathbf{b} = 0$$

atau

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y} \tag{2.2}$$

Persamaan (2.2) disebut persamaan normal. Dari persamaan normal tersebut, maka solusi untuk $\mathbf{b}$ adalah:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.3}$$

Sifat penduga $\mathbf{b}$ adalah:

1. Takbias

$$\begin{aligned}\mathrm{E}(\mathbf{b}) &= \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned}$$

Jadi $\mathbf{b}$ merupakan penduga takbias dari $\boldsymbol{\beta}$

2. Variansi minimum

$$\begin{aligned}\Sigma_b &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma_Y}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

$\Sigma_b = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ merupakan variansi terkecil dari semua penaksir linear takbias dijamin oleh teorema Gauss-Markov.

Teorema 2.1.1. Teorema Gauss-Markov
Penaksir kuadrat terkecil $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ merupakan variansi terkecil dalam himpunan semua penaksir linear takbias (Sembiring, 2003).

Bukti: Misalkan $\mathbf{b}_*$ penaksir linear lain dari β yang juga takbias. Karena $\mathbf{b}_*$ penaksir linear dapat dimisalkan bentuknya sebagai

$$\mathbf{b}_* = [(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' + U]\mathbf{Y},$$

untuk suatu U yang merupakan fungsi dari $\mathbf{X}$.
Jadi

$$\begin{aligned}\mathrm{E}(\mathbf{b}_*) &= [(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' + U]E(\mathbf{Y}) \\ &= [(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' + U](\mathbf{X}\beta) \\ &= \beta + U\mathbf{X}\beta = (I + U\mathbf{X})\beta\end{aligned}$$

Agar $\mathbf{b}_*$ penaksir takbias dari β maka haruslah $U\mathbf{X} = 0$.

$$\begin{aligned}
\Sigma_{b_*} &= [(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' + U]\mathbf{\Sigma_Y}[U' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'U' + UU' + (\mathbf{X}'\mathbf{X})^{-1} + U\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} + UU']
\end{aligned}$$

karena $U\mathbf{X} = \mathbf{X}'U' = 0$.

Matriks UU' adalah definit tak negatif, semua unsur diagonalnya berbentuk kuadrat. Jadi, terbukti bahwa variansi dari setiap unsur dari vektor $\mathbf{b}_*$ selalu lebih besar, atau paling kecil sama, dengan variansi unsur $\mathbf{b}$ yang sesuai.

2.1.3 Uji Asumsi Klasik Analisis Regresi Linear Berganda

Uji asumsi klasik pada regresi linear berganda adalah untuk menguji kelayakan model regresi dan kelayakan variabel bebasnya. Tujuannya adalah agar dapat menghasilkan nilai parameter yang baik sehingga hasil penelitian dapat diandalkan. Terdapat empat uji asumsi klasik yang harus dilakukan terhadap suatu model regresi tersebut, yaitu:

1. Uji Normalitas

   Uji normalitas data bertujuan untuk mendeteksi distribusi data dalam suatu variabel. Data yang baik dan layak untuk membuktikan model-model penelitian tersebut adalah data yang memiliki distribusi normal. Pada uji normalitas digunakan Uji Kolmogorov-Smirnov.

   Langkah-langkah uji normalitas, sebagai berikut:

   - Menentukan hipotesis:

     H_0: Berdistribusi normal

H_1: Berdistribusi tidak normal

- Menentukan derajat kepercayan α
- Menentukan nilai signifikansi:

  Jika nilai signifikansi $> \alpha$, maka terima H_0

  Jika nilai signifikansi $\leq \alpha$, maka tolak H_0
- Kesimpulan

2. Uji Autokorelasi

   Uji autokorelasi digunakan untuk mengetahui ada atau tidaknya penyimpangan korelasi yang terjadi antara residual pada satu pengamatan dengan pengamatan lain pada model regresi. Uji autokorelasi yang digunakan dalam model regresi linier berganda adalah metode Durbin-Watson.

   Langkah-langkah uji autokorelasi, sebagai berikut:

   - Menentukan hipotesis:

     H_0: Tidak ada autokorelasi

     H_1: Ada autokorelasi
   - Menentukan derajat kepercayan α
   - Menentukan nilai durbin, dL dan dU:

     Jika H_0 adalah hipotesis tidak adanya autokorelasi positif, maka keputusan dapat diartikan jika:

     $DW < dL$ = tolak H_0 maka ada autokorelasi positif

     $DW > dU$ = terima H_0 maka tidak ada autokorelasi positif

     $dL \leq DW \leq dU$ maka tidak menghasilkan kesimpulan yang pasti.

     Jika H_0 adalah hipotesis tidak adanya autokorelasi negatif, maka keputusan dapat diartikan jika:

$DW > 4 - dL$ = tolak H_0 maka ada autokorelasi negatif

$DW < 4 - dU$ = terima H_0 maka tidak ada autokorelasi negatif

$4 - dU \leq DW \leq 4 - dL$ maka tidak menghasilkan kesimpulan yang pasti.

- Kesimpulan

3. Uji Multikolinearitas

   Tujuan uji multikolinieritas adalah untuk melihat ada atau tidaknya hubungan linier atau korelasi yang tinggi antar variabel bebas dalam model regresi.

   Langkah-langkah uji multikolinearitas, sebagai berikut:

   - Menentukan hipotesis:

     H_0: Tidak ada multikolinieritas

     H_1: Ada multikolinieritas

   - Menentukan derajat kepercayaan α
   - Menentukan nilai VIF:

     Jika VIF ≥ 10, maka tolak H_0

     Jika VIF < 10, maka terima H_0

   - Kesimpulan

4. Uji Heterokedastisitas

   Uji heteroskedastisitas bertujuan menguji apakah dalam model regresi terjadi ketidaksamaan varian dari residual satu pengamatan ke pengamatan yang lain. Jika varian tetap, maka disebut homokedastisitas dan jika berbeda maka terjadi problem heterokedastisitas. Model regresi yang baik yaitu homokesdatisitas atau tidak terjadi heterokedastisitas. Untuk mendeteksi

adanya heterokedastisitas dapat melalui beberapa metode yaitu uji *Glesjer*, *scatter plot* atau uji korelasi *Rank Spearman*, yaitu dengan mengkorelasikan antara variabel bebas dengan residual.

Langkah-langkah uji heterokedastisitas, sebagai berikut:

- Menentukan hipotesis:

  H_0: Tidak ada heterokedastisitas

  H_1: Ada heterokedastisitas

- Menentukan derajat kepercayan α
- Menentukan nilai signifikansi residual:

  Jika signifikansi $\leq \alpha$, maka tolak H_0

  Jika signifikansi $> \alpha$, maka terima H_0

- Kesimpulan

2.1.4 Koefisien Determinasi

Koefisien determinasi (R^2) adalah ukuran yang menunjukkan persentase keragaman data pada variabel terikat (Y) yang diterangkan oleh model. Koefisien determinasi digunakan untuk menentukan seberapa baik model dugaan yang dihasilkan dalam *fitting* data.

Nilai koefisien determinasi (R^2) dapat diperoleh dengan rumus:

$$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{\text{JKR}}{\text{JKT}} \text{ dimana } 0 \leq R^2 \leq 1 \tag{2.4}$$

Nilai R^2 mendekati 0 (nol) menunjukkan bahwa data sangat tidak cocok dengan model regresi yang ada. Sebaliknya, jika nilai R^2 mendekati 1 (satu) menunjukkan bahwa data cocok dengan model regresi yang ada. Maka dapat

disimpulkan bahwa nilai R^2 yang didapat sesuai dengan yang dijelaskan masing-masing faktor yang tinggal di dalam regresi. Hal tersebut mengakibatkan bahwa yang dijelaskan hanyalah disebabkan faktor yang mempengaruhinya saja. Besarnya variansi yang dijelaskan penduga sering dinyatakan dalam persen. Persentase variansi penduga tersebut adalah $R^2 \times 100\%$.

2.1.5 Uji Hipotesis dan Selang Kepercayaan

Rumusan hipotesis untuk menguji parameter regresi secara simultan adalah sebagai berikut:

$H_0 : \boldsymbol{\beta} = \mathbf{0}$ artinya keseluruhan koefisien regresi tidak signifikan atau tidak ada variabel bebas yang berpengaruh nyata terhadap Y

$H_1 : \boldsymbol{\beta} \neq \mathbf{0}$ artinya paling sedikit terdapat satu variabel bebas yang berpengaruh nyata terhadap Y

dengan $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$

Statistik uji yang digunakan untuk menguji parameter regresi secara simultan adalah:

$$\begin{aligned} F_{\text{hitung}} &= \frac{\text{JKR}/p}{\text{JKG}/(n-p-1)} \\ &= \frac{(\hat{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{Y} - n\bar{Y}^2)/p}{(\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{Y})/(n-p-1)} \end{aligned} \tag{2.5}$$

Jika $F_{\text{hitung}} > F_{(p,n-p-1);\alpha/2}$, maka H_0 ditolak yang berarti paling sedikit terdapat satu variabel bebas yang berpengaruh nyata terhadap Y.

Rumusan hipotesis untuk menguji parameter regresi secara parsial adalah sebagai berikut:

$H_0 : \beta_j = 0$ artinya koefisien regresi ke-j tidak signifikan atau variabel bebas ke-j tidak berpengaruh nyata terhadap Y

$H_1 : \beta_j \neq 0$ artinya koefisien regresi ke-j signifikan atau variabel bebas ke-j berpengaruh nyata terhadap Y

Statistik uji yang digunakan untuk menguji parameter regresi secara parsial adalah sebagai berikut:

$$t_{\text{hitung}}(\hat{\beta}_j) = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \tag{2.6}$$

$\text{Var}(\hat{\beta}_j)$ diperoleh dari penguraian matriks $\text{Var}(\hat{\boldsymbol{\beta}})$, yaitu:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\
&= \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \\
&= \hat{\sigma}^2 \begin{pmatrix} c_{00} & c_{01} & \dots & c_{0p} \\ c_{10} & c_{11} & \dots & c_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p0} & c_{p1} & \dots & c_{pp} \end{pmatrix}
\end{aligned}$$

dengan c_{ij} adalah elemen dari $(\mathbf{X}'\mathbf{X})^{-1}$ pada baris ke-i, dan kolom ke-j, dimana $i, j = 0, 1, \dots, p$ dan $\hat{\sigma}^2 = \text{JKG}/(n - p - 1)$

Dengan demikian $\sqrt{\text{Var}(\hat{\beta}_j)} = \sigma\sqrt{c_{jj}}$, dan persamaan (2.6) dapat ditulis kembali sebagai berikut:

$$t_{\text{hitung}}(\hat{\beta}_j) = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \tag{2.7}$$

dengan $j = 1, 2, \ldots, p$.

Jika $\left|t_{\text{hitung}}(\hat{\beta}_j)\right| > t_{(n-p-1);\alpha/2}$, maka H_0 ditolak yang artinya variabel bebas ke-j berpengaruh nyata terhadap Y.

Selang kepercayaan untuk β_j dengan tingkat kepercayaan $100(1-\alpha)\%$ adalah

$$\hat{\beta}_j \pm t_{(n-p-1);\alpha/2}\hat{\sigma}\sqrt{c_{jj}} \tag{2.8}$$

2.2 Multikolinearitas

Multikolinearitas adalah adanya korelasi di antara variabel-variabel bebas dari model regresi. Multikolinearitas dapat memberi dampak untuk model regresi, yaitu (Neter, 1990):

1. Multikolinearitas antara variabel-variabel bebas dalam model regresi linier mengakibatkan variansi penduga kuadrat terkecil menjadi besar sehingga menghasilkan galat baku yang lebih besar. Hal ini mengakibatkan selang kepercayaan untuk parameter model regresi menjadi lebih besar.

2. Satu atau lebih variabel bebas menjelaskan variabel terikat benar-benar sama dengan yang dijelaskan oleh variabel bebas lain.

3. Pengujian hipotesis parameter berdasarkan metode kuadrat terkecil memberikan hasil yang tidak valid.

Pada analisis regresi, adanya multikolinearitas dapat dikatakan jika terdapat beberapa kondisi sebagai berikut:

1. Nilai korelasi antar variabel bebas (r_{XY}) melebihi 0,5 (Gujarati, 1995).

2. Terjadi perbedaan yang sangat besar (ketidakstabilan) pada penduga koefisien regresi bila suatu variabel bebas ditambahkan atau dibuang dari model.

3. Tidak ada satupun koefisien regresi yang bersifat signifikan jika dilakukan uji hipotesis secara parsial terhadap masing-masing koefisien, meskipun uji hipotesis terhadap seluruh koefisien regresi secara simultan bersifat signifikan.

4. Terdapat satu atau lebih nilai eigen yang mendekati nol.

5. Nilai VIF lebih besar dari 10 (O'Brien, 2007) *Variance Inflation Factor* (VIF) atau faktor inflasi ragam dapat menginterpretasikan akibat dari korelasi antar variabel bebas ke-j pada varians penduga koefisien regresi. VIF untuk kofisien regresi-j didefinisikan sebagai berikut:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2.9}$$

Nilai $1 - R_j^2$ menunjukkan nilai toleransi yang mewakili varians dari variabel bebas ke-j yang tidak dihubungkan dengan variabel bebas lain pada model, sehingga nilai toleransi berbanding terbalik dengan nilai VIF.
R_j^2 adalah koefisien determinasi antara X_j dengan variabel bebas lainnya pada persamaan/model dugaan regresi. Formula untuk R_j^2, yaitu:

$$R_j^2 = \frac{JKR_j}{JKT_j} = \frac{\hat{\beta}'\mathbf{X}'\mathbf{X}_j - [(\Sigma X_j)^2/n]}{\mathbf{X}_j'\mathbf{X}_j - [(\Sigma X_j)^2/n]} \tag{2.10}$$

dengan X_j adalah salah satu variabel bebas ke-j; dimana $j = 1, 2, \ldots, p$.
Jika $R_j^2 = 0$ atau $VIF = 1$, mengindikasikan bahwa variabel bebas ke-j orthogonal dengan variabel bebas lainnya.

2.3 *Outlier*

Outlier didefinisikan sebagai sebagian dari data pengamatan yang memiliki pola yang berbeda dari sebagian besar data pengamatan (Hadi et al, 2009). *Outlier* dapat menyebabkan ketidakhomogenan matriks varian kovarian pada data. Selain itu Hadi et al (2009) menyebutkan *outlier* memberi efek *swamping* (kesalahan mengidentifikasi data *non outliers* sebagai *outliers*) dan *masking* (mengaburkan data). Jumlah *outlier* pada data yang dapat diterima, ditentukan oleh selang kepercayaan. Apabila ada data di luar selang kepercayaan, maka dapat dihitung persentase dari *outlier* data tersebut. Pada analisis regresi, adanya *outlier* dapat menyebabkan berbagai penyimpangan, antara lain:

1. Residual yang besar dari model yang terbentuk,
2. Varian data menjadi lebih besar,
3. Rentang yang lebar pada *confidence region*.

Dalam menghadapi masalah identifikasi *outlier* terdapat dua pendekatan (Hadi et al, 2009), yaitu: estimasi *robust* dan metode yang khusus mengidentifikasi *outlier*. Metode *robust* dirancang secara khusus untuk mengatasi *outlier*, di mana kemudian saat penghitungan, estimasi parameter *robust* dapat digunakan untuk mengidentifikasi *outliers*. Di samping itu, prosedur identifikasi *outliers* dapat pula digunakan untuk mendapatkan estimator *robust*.

Berbagai cara mendeteksi keberadaan *outlier* pada data pengamatan, antara lain:

1. Membandingkan nilai F_i dengan F_{tabel}. Suatu data pengamatan dikatakan

outlier jika nilai

$$F_i > F_{\alpha;p,n-p-1} \quad \text{atau} \quad \frac{(n-p-1)nd_i^2}{p(n-1)^2npd_i^2} > F_{\alpha;p,n-p-1}$$

2. Menggunakan *Leverage* yang berkaitan dengan Jarak *Mahalanobis*. Nilai *Leverage* untuk sampel ke-i didefinisikan sebagai: $h_i = \frac{1}{n} + \mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i$ Di mana $\mathbf{X}$ adalah matriks data. Jika nilai *Leverage* $> (2p-1)/n$ maka dianggap sebagai *outlier*.

3. Pemeriksaan terhadap *outlier* dapat dilakukan dengan uji Jarak *Mahalanobis*. Jarak *Mahalanobis* dievaluasi dengan menggunakan $C = \sqrt{\chi^2_{p;1-\alpha}}$, C dinyatakan sebagai nilai *cut-off*. Data tidak memiliki *outlier* apabila Jarak *Mahalanobis* tidak lebih besar dari nilai *cut-off*.

2.4 Komponen Utama yang Dibentuk Berdasarkan Matriks Kovariansi

Teorema 2.4.1. Misal Σ merupakan matriks kovariansi dari vektor acak $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ dengan pasangan nilai eigen dan vektor eigen yang saling ortonormal adalah $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$, dimana $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Maka komponen utama ke-i didefinisikan sebagai berikut:

$$W_i = \mathbf{e_i'X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p \;;\; i = 1, 2, \ldots, p \tag{2.11}$$

dengan syarat

$$\text{Var}(W_i) = \mathbf{e_i'\Sigma e_i} = \lambda_i, \;\; i = 1, 2, \ldots, p$$

$$\text{Cov}(W_i, W_k) = \mathbf{e_i'\Sigma e_k} = 0, \;\; i \neq k$$

Jika terdapat λ_i yang sama,maka vektor eigen $\mathbf{e_i}$ menjadi tidak unik (yang tentunya juga mengakibatkan W_i tidak unik) (Johnson, 2007).

Bukti: Berdasarkan maksimisasi bentuk kuadrat, maka matriks Σ akan memaksimumkan

$$\max_{\mathbf{a}\neq 0} \frac{\mathbf{a'\Sigma a}}{\mathbf{a'a}} = \lambda_1 \quad \text{(akan dicapai pada saat } \mathbf{a} = \mathbf{e_1})$$

Karena $\mathbf{e_1'e_1}$, maka:

$$\begin{aligned}
\max_{\mathbf{a}\neq 0} \frac{\mathbf{a'\Sigma a}}{\mathbf{a'a}} &= \frac{\mathbf{e_1'\Sigma e_1}}{\mathbf{e_1'e_1}} = \mathbf{e_1'\Sigma e_1} \\
&= \mathbf{e_1'\Sigma^{1/2}\Sigma^{1/2}e_1} = \mathbf{e_1'P\Lambda^{1/2}P'P\Lambda^{1/2}P'e_1} \\
&= (\mathbf{P'e_1})'\mathbf{\Lambda P'e_1} = \lambda_1 = \mathrm{Var}(W_1)
\end{aligned}$$

dengan $\mathbf{P} = \begin{pmatrix} \mathbf{e_1} & \mathbf{e_2} & \dots & \mathbf{e_p} \end{pmatrix}$, dimana $\mathbf{PP'} = \mathbf{P'P} = \mathbf{I}$, dan Λ adalah matriks diagonal yang memiliki elemen diagonal utama $\lambda_1, \lambda_2, \dots, \lambda_p$. Dari definisi matriks kovariansi maka berlaku:

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{P\Lambda P'} \\
\mathbf{\Sigma P} &= \mathbf{P\Lambda P'P} \\
\mathbf{\Sigma}\begin{bmatrix} \mathbf{e_1} & \mathbf{e_2} & \dots & \mathbf{e_p} \end{bmatrix} &= \begin{bmatrix} \mathbf{e_1} & \mathbf{e_2} & \dots & \mathbf{e_p} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \\
\begin{bmatrix} \mathbf{\Sigma e_1} & \mathbf{\Sigma e_2} & \dots & \mathbf{\Sigma e_p} \end{bmatrix} &= \begin{bmatrix} \mathbf{e_1}\lambda_1 & \mathbf{e_2}\lambda_2 & \dots & \mathbf{e_p}\lambda_p \end{bmatrix} \qquad (2.12)
\end{aligned}$$

Dari (2.12), yaitu $\mathbf{\Sigma e_k} = \mathbf{e_k}\lambda_k$ dan dengan cara yang serupa akan diperoleh

$$\begin{aligned}\max_{\mathbf{a}\perp\mathbf{e_1},\mathbf{e_2},\ldots,\mathbf{e_k}} \frac{\mathbf{a'\Sigma a}}{\mathbf{a'a}} &= \frac{\mathbf{e'_{k+1}\Sigma e_{k+1}}}{\mathbf{e'_{k+1}e_{k+1}}} = \mathbf{e'_{k+1}\Sigma e_{k+1}} \\ &= \mathbf{e'_{k+1}e_{k+1}}\lambda_{k+1} = \lambda_{k+1} \\ &= \mathrm{Var}(W_{k+1})\end{aligned}$$

dengan $\mathbf{e'_{k+1}e_{k+1}} = 1$ dan $\mathbf{e'_{k+1}e_i} = 0$, dimana $i = 1, 2, \ldots, k$; $k = 1, 2, \ldots, p-1$.

Akan ditunjukkan $\mathbf{e_i}$ tegak lurus dengan $\mathbf{e_k}$ (dimana $\mathbf{e'_ie_k} = 0, i \neq k$) yang mengakibatkan $\mathrm{Cov}(W_i, W_k) = 0$. Vektor-vektor eigen dari $\mathbf{\Sigma}$ akan saling ortogonal jika $\lambda_1, \lambda_2, \ldots, \lambda_p$ berbeda nilainya. Namun, jika terdapat nilai eigen yang sama, maka vektor eigen yang bersesuaian dengan nilai eigen yang relatif besar yang akan dipilih untuk mencapai kondisi ortogonal. Oleh sebab itu, jika terdapat dua vektor eigen $\mathbf{e_i}$ dan $\mathbf{e_k}$, maka $\mathbf{e'_ie_k} = 0, i \neq k$, sehingga:

$$\mathrm{Cov}(W_i, W_k) = \mathbf{e'_i\Sigma e_k} = \mathbf{e'_ie_k}\lambda_k = 0$$

dengan $i, k = 1, 2, \ldots, p$, dan $i \neq k$.

Teorema 2.4.2. Misal $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ mempunyai matriks kovariansi $\mathbf{\Sigma}$, dengan pasangan nilai eigen-vektor eigen $(\lambda_1, \mathbf{e_1}), (\lambda_2, \mathbf{e_2}), \ldots, (\lambda_p, \mathbf{e_p})$; $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Jika $W_1 = \mathbf{e'_1X}, W_2 = \mathbf{e'_2X}, \ldots, W_p = \mathbf{e'_pX}$ adalah komponen-komponen utama maka total variansi komponen utama didefinisikan sebagai berikut (Johnson, 2007)

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(W_i)$$

Bukti: Dari definisi *trace*, $\sigma_{11}+\sigma_{22}+\ldots+\sigma_{pp} = \mathrm{tr}(\Sigma)$. Dari definisi matriks kova-

riansi, maka dapat ditulis $\mathbf{\Sigma} = \mathbf{P\Lambda P'}$, dengan $\mathbf{\Lambda}$ adalah matriks diagonal yang memiliki elemen nilai eigen pada diagonal utamanya, dan $\mathbf{P} = \begin{bmatrix} \mathbf{e_1} & \mathbf{e_2} & \dots & \mathbf{e_p} \end{bmatrix}$, sedemikian sehingga $\mathbf{PP'} = \mathbf{P'P} = \mathbf{I}$. Dengan demikian trace $\mathbf{\Sigma}$ adalah:

$$\text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{P\Lambda P'}) = \text{tr}(\mathbf{\Lambda P'P}) = \text{tr}(\mathbf{\Lambda})$$

Sementara itu total variansi komponen utama adalah:

$$\sum_{i=1}^{p} \text{Var}(X_i) = \text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{\Lambda}) = \sum_{i=1}^{p} \text{Var}(W_i)$$

Berdasarkan teorema 2.4.2 telah diperoleh

$$\begin{aligned} \text{Total variansi kumulatif} &= \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \ldots + \lambda_p \end{aligned} \tag{2.13}$$

Dari (2.13) maka proporsi total variansi yang dijelaskan komponen utama ke-k, yaitu:

$$\begin{pmatrix} \text{Proporsi total variansi} \\ \text{populasi yang dijelaskan oleh} \\ \text{komponen utama ke-k} \end{pmatrix} = \frac{\lambda_k}{\text{tr}(\Sigma)} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \;\; ; \;\; k = 1, 2, \ldots, p \tag{2.14}$$

Proporsi total variansi populasi yang dijelaskan oleh komponen utama ke-k adalah suatu ukuran yang menunjukkan berapa besar keragaman data asal yang dapat dijelaskan oleh komponen utama ke-k (biasanya dinyatakan dalam persen). Jika proporsi kumulatif keragaman data asal yang dijelaskan oleh k komponen utama sudah mencapai minimal 80%, maka komponen utama yang berjumlah k dapat mengganti p variabel asal dimana informasi keragaman data

asal yang tidak dijelaskan oleh k komponen utama tidak terlalu besar (Johnson, 2007).

Teorema 2.4.3. Jika $W_1 = \mathbf{e_1'X}, W_2 = \mathbf{e_2'X}, \ldots, W_p = \mathbf{e_p'X}$ merupakan komponen utama yang diperoleh dari matriks kovariansi ($\mathbf{\Sigma}$), maka korelasi antara komponen utama W_i dengan variabel X_k

$$\rho_{W_i.X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

dengan $(\lambda_1, \mathbf{e_1}), (\lambda_2, \mathbf{e_2}), \ldots, (\lambda_p, \mathbf{e_p})$, dimana $i, k = 1, 2, \ldots, p$ adalah pasangan nilai eigen dan vektor eigen dari $\mathbf{\Sigma}$ (Johnson, 2007).

Bukti: Jika $\mathbf{a_k'} = [0, \ldots, 0, 1, 0, \ldots, 0]$, maka $X_k = \mathbf{a_k'X}$, dan kovariansi W_i dan X_k adalah sebagai berikut:

$$\begin{aligned} \mathrm{Cov}(X_k, W_i) &= \mathrm{Cov}(\mathbf{a_k'X}, \mathbf{e_i'X}) = \mathrm{E}(\mathbf{a_k'XX'e_i}) \\ &= \mathbf{a_k'}\mathrm{E}(XX')\mathbf{e_i} = \mathbf{a_k'\Sigma e_i} \\ &= \mathbf{a_k'}\lambda_k\mathbf{e_i} = \lambda_k e_{ik} \end{aligned}$$

Sementara itu, karena nilai $\mathrm{Var}(W_i) = \lambda_i$, dan $\mathrm{Var}(X_k) = \sigma_{kk}$, maka korelasi antara W_i dengan X_k adalah

$$\rho_{W_i.X_k} = \frac{\mathrm{Cov}(W_i, X_k)}{\sqrt{\mathrm{Var}(W_i)}\sqrt{\mathrm{Var}(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

Korelasi antara variabel bebas dan komponen utama sering membantu dalam melakukan interpretasi data, tetapi ukuran korelasi ini hanya menunjukkan kontribusi satu variabel bebas X terhadap komponen utama W tanpa memperhatikan adanya variabel bebas lain. Johnson (2007) menyarankan, selain mengguna-

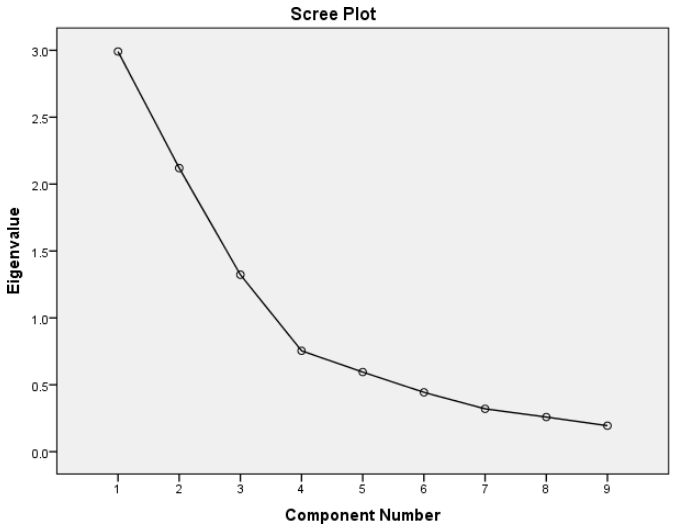
kan korelasi antara variabel bebas dengan komponen utama, sebaiknya digunakan juga koefisien e_{ik} dalam melakukan interpretasi komponen utama, karena koefisien e_{ik} menunjukkan kontribusi satu variabel X terhadap komponen utama dimana keberadaan variabel lain juga diperhatikan.

2.4.1 Kriteria Pemilihan Komponen Utama

Salah satu tujuan dari analisis komponen utama ialah mereduksi dimensi data asal yang awalnya memiliki p variabel bebas menjadi w komponen utama (dimana $w < p$). Kriteria pemilihan w adalah:

1. Proporsi kumulatif keragaman data asal yang dijelaskan oleh w komponen utama minimal 80%, dan proporsi total variansi populasi bernilai cukup besar (Johnson, 2007).

2. Dengan menggunakan *scree plot* yaitu antara *component number* (i) dengan *eigen value* (λ_i), pemilihan nilai w berdasarkan *scree plot* ditentuan dengan melihat letak terjadinya belokan pada grafik atau saat nilai eigennya kecil dengan menghapus komponen utama yang menghasilkan beberapa nilai eigen kecil yang membentuk pola garis lurus. Sebagai contoh dapat dilihat pada gambar (2.1), dimana pada gambar tersebut dari 9 komponen utama dapat dipilih 4 komponen utama saja karena nilai eigen yang kecil pada $i = 5, 6, 7, 8, 9$ berbentuk garis lurus setelah belokan pada $i = 4$ (Rencher, 1998).



Gambar 2.1: Contoh Scree Plot

2.5 Regresi Komponen Utama dari Matriks Varian Kovarian

Komponen-komponen utama berdasarkan matriks varian kovarian sesuai persamaan (2.11) adalah sebagai berikut:

$$\begin{aligned} W_1 &= \mathbf{e_1'X} = e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_p \\ W_2 &= \mathbf{e_2'X} = e_{21}X_1 + e_{22}X_2 + \ldots + e_{2p}X_p \\ &\vdots \\ W_p &= \mathbf{e_p'X} = e_{p1}X_1 + e_{p2}X_2 + \ldots + e_{pp}X_p \end{aligned} \tag{2.15}$$

dimana W_1 merupakan suatu komponen pertama yang memenuhi maksimum nilai $\mathbf{e_1'\Sigma e_1} = \lambda_1$. W_2 adalah komponen kedua yang memenuhi sisa keragaman selain komponen pertama dengan memaksimumkan nilai $\mathbf{e_2'\Sigma e_2} = \lambda_2$. W_p adalah komponen ke-p yang memenuhi sisa keragaman selain $W_1, W_2, \ldots, W_{p-1}$ dengan

memaksimumkan nilai $\mathbf{e_p'\Sigma e_p} = \lambda_p$. Urutan $W_1, W_2, \ldots, W_p$ harus memenuhi persyaratan $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Persamaan (2.15) dapat ditulis dalam notasi matriks, yaitu:

$$\mathbf{W} = \mathbf{XP} \tag{2.16}$$

dengan

$\mathbf{W}$ adalah suatu matriks berukuran $n \times p$ yang elemennya terdapat nilai (*score*) p komponen utama.

$\mathbf{X}$ adalah rancangan (*design matrix*) yang berukuran $n \times p$.

$\mathbf{P}$ adalah matriks berukuran $p \times p$ yang elemen-elemennya merupakan vektor eigen, dimana masing-masing vektor eigen $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_p}$ berukuran $p \times 1$ yang memenuhi $\mathbf{e_i'e_i'} = 1$ dan $\mathbf{e_i'e_j'} = 1; i \neq j$.

Persamaan (2.16) merupakan transformasi ortogonal yang variabel-variabel $X_1, X_2, \ldots, X_p$ diubah menjadi komponen utama $W_1, W_2, \ldots, W_p$ dengan kombinasi linier yang melibatkan matriks ortogonal $\mathbf{P}$, dimana $\mathbf{P'P} = \mathbf{PP'} = \mathbf{I}$.

Pembentukan regresi komponen utama berdasarkan matriks varian kovarian dimulai dari model regresi linier berganda dimana variabel bebasnya sudah dinilaitengahkan (*centered*). Misal $\mathbf{X_t}$ adalah matriks yang elemen-elemennya dikurang dengan rataannya (*centered*) dengan syarat rataan nol dan variansi σ^2, sehingga model regresi linier berganda yang sudah dinilaitengahkan variabel bebas adalah:

$$Y = \beta_0 + \beta_1(X_1 - \bar{X}_1) + \beta_2(X_2 - \bar{X}_2) + \ldots + \beta_p(X_p - \bar{X}_p) + \varepsilon \tag{2.17}$$

Nilai eigen yang bersesuaian vektor eigen didapat dari $|\mathbf{X_t'X_t} - \lambda\mathbf{I}| = 0$,

sedangkan hasil subtitusi nilai eigen pada persamaan $\mathbf{X_t'X_te_j} = \lambda\mathbf{e_j}$ akan didapat vektor eigen ke-j ($\mathbf{e_j}$), dimana $\mathbf{X_t'X_t} = (n-1)\Sigma; j = 1, 2, \ldots, p$.

Matriks $\mathbf{P}$ merupakan matriks ortogonal yang memenuhi persamaan berikut ini $\mathbf{P'P} = \mathbf{PP'} = \mathbf{I}$. Karena $\mathbf{W} = \mathbf{X_tP}$, sehingga persamaan regresi linier berganda pada (2.17), menjadi:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X_t}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{Y} &= \mathbf{X_tPP'}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{Y} &= \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \end{aligned} \tag{2.18}$$

dengan $\mathbf{W}$ merupakan matriks berukuran $n \times w$ yang elemennya adalah komponen utama, dan $\alpha' = (\alpha_0, \alpha_1, \ldots, \alpha_w)$ merupakan vektor koefisien komponen utama berukuran $w \times 1$. Persamaan (2.18) dapat ditulis sebagai berikut:

$$\mathbf{Y} = \alpha_0\mathbf{1} + \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{2.19}$$

dengan:

$\mathbf{1}$ adalah vektor yang setiap elemennya 1 berukuran $n \times 1$

$\mathbf{W'} = (W_1, W_2, \ldots W_w)$

$\boldsymbol{\alpha'} = (\alpha_1, \alpha_2, \ldots, \alpha_w)$.

Contoh 2.5.1. Berikut ini adalah cara untuk mereduksi dimensi data dan sekaligus menyelesaikan masalah multikolinearitas pada data dengan Metode *Principal Component Analysis* (PCA). Misal terdapat data sebagai berikut:

Tabel 2.1: Contoh Data

Y	X1	X2	X3	X4	X5	X6
574	11,20	76,00	90,00	5,59	45,00	25,33
71	3,20	64,00	65,00	0,74	21,67	21,33
115	5,40	58,00	70,00	2,64	35,00	19,33
295	5,80	72,00	93,00	3,30	46,50	24,00
116	5,00	59,00	73,00	3,50	36,50	14,75
58	8,70	45,00	23,00	2,52	11,50	15,00
184	5,30	57,00	99,00	2,60	49,50	19,00
118	2,60	74,00	86,00	2,05	43,00	24,67

Dari data di atas didapat persamaan regresi linear bergandanya adalah:

$$Y = -713 + 47,6X_1 + 6,18X_2 + 5,15X_3 + 6,1X_4 - 4,64X_5 - 0,12X_6$$

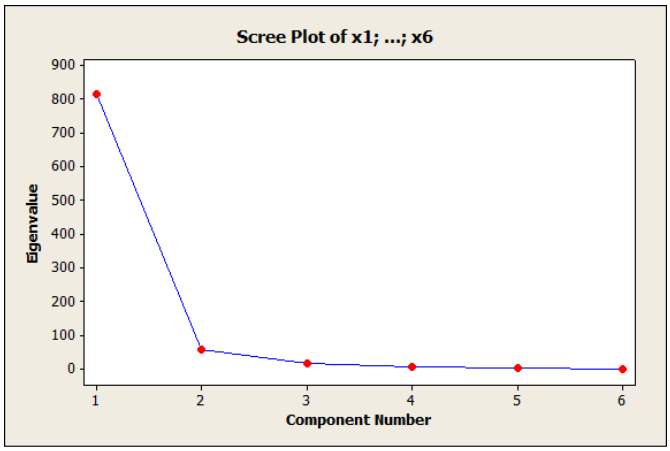
Adanya gejala multikolinearitas dapat dilihat melalui nilai *Variance Inflation Factor* (VIF) di setiap koefisien regresi linear berganda berikut:

Tabel 2.2: Uji multikolinearitas koefisien regresi linear berganda

Predictor	Coef	SE Coef	T	P	VIF
Constant	-713,32	76,31	-9,35	0,07	
X1	47,62	10,53	4,52	0,14	110,70
X2	6,18	3,27	1,89	0,31	148,50
X3	5,15	1,03	5,02	0,13	77,30
X4	6,14	26,99	0,23	0,86	176,10
X5	-4,64	2,57	-1,80	0,32	146,00
X6	-0,12	6,75	-0,02	0,99	98,70

Nilai VIF jauh lebih besar dari 10 sehingga menandakan adanya multikolinearitas. Kemudian dianalisis dengan analisis komponen yang didasarkan pada matriks varian kovarian dan diperoleh hasil sebagai berikut:

Eigenvalue	815,970	57,970	14,630	4,640	2,540	0,010
Proportion	0,911	0,065	0,016	0,005	0,003	0,000
Cumulative	0,911	0,976	0,992	0,997	1,000	1,000



Gambar 2.2: Scree Plot

Berdasarkan kriteria pemilihan komponen utama, komponen utama ke-1 (W_1) dan komponen utama ke-2 (W_2) telah menjelaskan 97,6% keragaman dari Y, dan dari *scree plot* yang ada dapat dipilih 2 komponen utama. Karena nilai eigen pada komponen utama ke 3,4,5,6 sangat kecil dan membentuk garis lurus pada *scree plot*. Sehingga didapat komponen utama terpilih, sebagai berikut:

Tabel 2.3: Komponen Utama terpilih

W1	W2
-119,8930	46,3561
-84,6300	45,3111
-92,9170	34,8342
-121,8690	41,2405
-95,9770	33,1084
-38,6630	37,3833
-123,5150	24,3980
-115,0660	45,3967

Sehingga didapat persamaan regresi komponen utamanya adalah sebagai berikut:

$$Y = -422 - 3,41W_1 + 7,15W_2$$

2.6 Jarak *Mahalanobis*

Jarak *Mahalanobis* adalah jarak antara masing-masing vektor data dengan titik pusat data atau vektor rata-rata. Jarak Mahalanobis didefinisikan sebagai berikut:

$$\lambda^2 = (X - \mu)'\Sigma^{-1}(X - \mu) \tag{2.20}$$

dengan:

λ^2 adalah Jarak *Mahalanobis*

X adalah vektor data

μ adalah vektor rata-rata

Σ adalah matriks varian kovarian

Σ^{-1} adalah invers dari matriks Σ

Suatu formula dikatakan jarak dimana $\forall x, y, k \in \Re^p$, jika memenuhi:

1. $d(x, y) \geq 0$ dan $d(x, y) = 0$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, k) + d(k, y)$

Akan ditunjukkan bahwa rumus Jarak *Mahalanobis* merupakan jarak:

1. Sifat nomor 1 dari definisi jarak dipenuhi oleh rumusan Jarak *Mahalanobis*, karena matriks kovariansi adalah matriks semi definit positif maka invers dari matriks kovariansi merupakan matriks diagonal yang positif, sehingga Jarak *Mahalanobis* pasti non negative, untuk lebih lengkapnya dapat dilihat dalam Jurnal Pengendalian Kualitas Produksi Mebel di PT. MAJAWANA dengan Diagram Kontrol D^2 (*Mahalanobis Distance*) (Primananda, 2010)

2. Sifat nomor 2 dipenuhi oleh rumusan Jarak *Mahalanobis* berikut penjelasannya:

$$\begin{aligned} d(X,\mu) &= \sqrt{(X-\mu)' * \Sigma^{-1} * (X-\mu)} \\ &= \sqrt{-(\mu-X)' * \Sigma^{-1} * (-(\mu-X))} \\ &= \sqrt{(\mu-X)' * \Sigma^{-1} * (\mu-X)} \\ &= d(\mu,X) \end{aligned}$$

3. Sifat nomor 3 (ketaksamaan segitiga) dipenuhi oleh rumusan Jarak *Mahalanobis*. Tanda ketaksamaan bias menjadi sama dengan jika vektor x, k dan y collinear (segaris).

Jadi, terbukti bahwa Jarak *Mahalanobis* adalah rumusan jarak, sehingga $\lambda^2 \geq 0$.

2.7 Regresi *Robust*

Regresi *robust* adalah metode yang penting untuk menganalisa data yang berpengaruh dalam pembentukan model regresi. Regresi *robust* ini bertujuan menghasilkan model yang menyediakan hasil yang *robust* (stabil), selain itu juga digunakan untuk mendeteksi adanya pencilan (*outlier*).

Pada regesi *robust* terdapat beberapa *estimator covariance robust* diantaranya *Minimum Covariance Determinant* (MCD) yang digunakan untuk mengatasi kelemahan estimator dalam mengestimasi matriks kovariansi. MCD estimator adalah estimator matriks varian kovarian yang menggunakan sebagian data yang menghasilkan determinan matriks varian kovarian terkecil. Data-data yang normal diberi pembobot sama dengan nol sedang data-data *outlier* diberi pembobot sama dengan satu. Pemanfaatan *estimator covariance robust* dalam dunia

analisis multivariat adalah untuk me-*robust*-kan metode multivariat itu sendiri. Mengestimasi matriks varian kovarian dapat ditentukan dengan menghitung matriks varian kovarian sampel terlebih dahulu:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

Dengan setiap elemen yang diperoleh dari rumus berikut:

$$s_{mj} = s_{jm} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{im} - \bar{x}_m)(x_{ij} - \bar{x}_j) \tag{2.21}$$

dengan,

s_{mj} adalah entri baris ke-m dan kolom ke-j dari matriks $\mathbf{S}$,

x_{im} adalah entri pada baris ke-i dan kolom ke-m dari matriks $\mathbf{X}$ dengan $i = 1, 2, \dots, n$ dan $m = 1, 2, \dots, p$

$\bar{x}_m$ adalah rata-rata pada kolom ke-m dari matriks $\mathbf{X}$ dengan $m = 1, 2, \dots, p$,

x_{ij} adalah entri pada baris ke-i dan kolom ke-j dari matriks $\mathbf{X}$ dengan $i = 1, 2, \dots, n$ dan $j = 1, 2, \dots, p$,

$\bar{x}_j$ adalah rata-rata pada kolom ke-j dari matriks $\mathbf{X}$ dengan $j = 1, 2, \dots, p$.

Berdasarkan matriks tersebut akan dapat melakukan Analisis Komponen Utama atau disebut dengan *Principal Component Analysis* (PCA), menghitung Jarak *Mahalanobis*, didapat pula penduga matriks varian kovarian dari $\mathbf{S}$ yaitu dinotasikan dengan $\mathbf{S}_{MCD}$ yang akan dibahas pada bab selanjutnya, kemudian dari matriks $\mathbf{S}_{MCD}$ dapat dihitung Jarak *Robust*, dan membuat *Diagnostic Plot*. Sehingga dapat terdeteksi adanya *outlier* pada data dan dapat mengklasifikasikan

jenis *outlier*-nya.

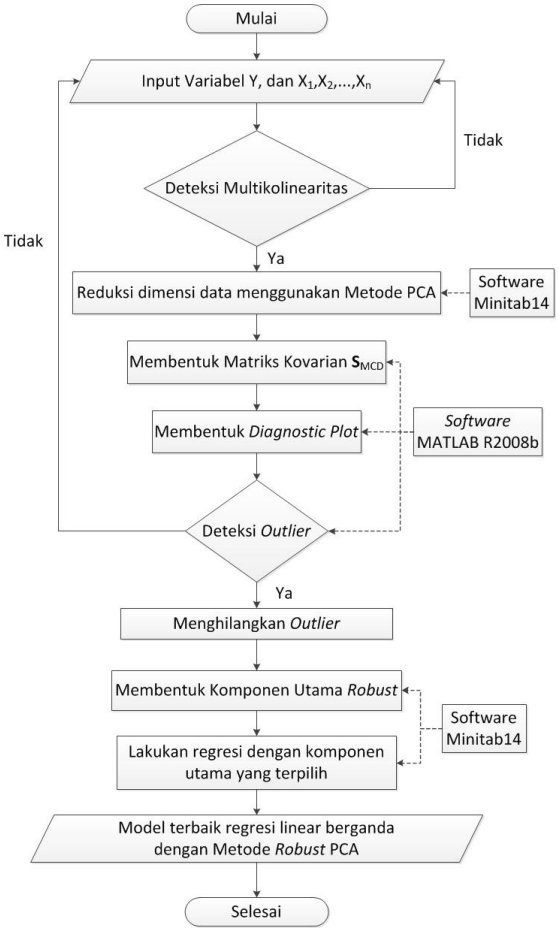
2.7.1 Metode *Robust* PCA

Metode *Robust* PCA adalah suatu metode PCA yang *robust* terhadap keberadaan *outlier* pada data. Agar mendapatkan komponen utama yang *robust* diperlukan penggabungkan konsep *Principal Component Analysis* (PCA) dengan penduga *robust Minimum Covariance Determinant* (MCD). Teknik PCA digunakan untuk reduksi dimensi awal kemudian estimator MCD diaplikasikan menghasilkan estimasi yang lebih akurat. Secara sederhana metode *Robust* PCA dapat dideskripsikan sebagai berikut, jika diasumsikan data asal berupa suatu matriks $\mathbf{X}$ berukuran $n \times p$ dimana n jumlah pengamatan dan p jumlah variabel asal maka metode *Robust* PCA dilakukan dalam 3 langkah utama. Pertama, dilakukan reduksi pada dimensi data. Kemudian dibentuk matriks kovarian $\mathbf{S}_{MCD}$ yang digunakan untuk memilih jumlah komponen k menghasilkan subruang berdimensi k yang cocok dengan data. Titik-titik data kemudian diproyeksikan ke subruang ini kemudian lokasi dan matriks sebarannya diestimasi secara *robust* dan dihitung nilai eigen $l_1, l_2, \ldots l_k$. Maka didapat vektor eigen yang bersesuaian adalah sejumlah k komponen utama yang *robust*.

BAB III

PEMBAHASAN

Pada bab ini akan dijelaskan bagaimana cara mengatasi multikolinearitas dan *outlier* pada model regresi linear berganda dengan Metode *Robust* PCA. Berikut diagram alir yang menjelaskan tentang Metode *Robust* PCA.



Gambar 3.1: Diagram Alir Metode *Robust* PCA

Dimulai dari reduksi dimensi data, mencari *estimator covariance robust* dengan MCD, mendeteksi *outlier*, membentuk komponen utama yang *robust*, dan melakukan regresi dengan komponen utama robust yang terpilih, sehingga didapat model terbaik regresi linear berganda yang sudah tidak mengandung *outlier*. Selain itu, akan dibuat contoh aplikasi dengan menggunakan Metode *Robust* PCA.

3.1 Reduksi Dimensi dengan PCA

Untuk mereduksi dimensi data dengan metode PCA, pertama-tama akan dianalisis dengan teknik analisis komponen utama. Misal dari vektor acak $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ didapat $\mathbf{\Sigma}$ yang merupakan matriks kovariansi, selanjutnya diperoleh pasangan nilai eigen dan vektor eigen sebagai berikut $(\lambda_1, e_1), (\lambda_2, e_2), \ldots,$ (λ_p, e_p), dimana $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Maka komponen utamanya diperoleh sebagai berikut:

$$KU1 = W_1 = \begin{pmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{n1} \end{pmatrix} = e_{11} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + e_{21} \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \ldots + e_{p1} \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$$

$$KU2 = W_2 = \begin{pmatrix} w_{12} \\ w_{22} \\ \vdots \\ w_{n2} \end{pmatrix} = e_{12} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + e_{22} \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \ldots + e_{p2} \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$$

$$\vdots$$

$$KUp = W_p = \begin{pmatrix} w_{1p} \\ w_{2p} \\ \vdots \\ w_{np} \end{pmatrix} = e_{1p} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + e_{2p} \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \ldots + e_{pp} \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$$

Persamaan di atas dapat ditulis sebagai berikut:

$$\begin{aligned} W_1 &= e_{11}X_1 + e_{21}X_2 + \ldots + e_{p1}X_p \\ W_2 &= e_{12}X_1 + e_{22}X_2 + \ldots + e_{p2}X_p \\ &\vdots \\ W_p &= e_{1p}X_1 + e_{2p}X_2 + \ldots + e_{pp}X_p \end{aligned} \tag{3.1}$$

dalam bentuk matriks dapat dinyatakan dengan:

$$\mathbf{W} = \mathbf{XP} \tag{3.2}$$

$$\begin{bmatrix} W_1 & W_2 & \ldots & W_p \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & \ldots & X_p \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} & \ldots & e_{1p} \\ e_{21} & e_{22} & \ldots & e_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{p1} & e_{p2} & \ldots & e_{pp} \end{bmatrix}$$

$$= \begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1p} \\ w_{21} & w_{22} & \ldots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \ldots & w_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \vdots \\ \mathbf{L}_n \end{bmatrix}$$

dengan **W** adalah vektor komponen utama, **P** adalah vektor eigen dari matriks varians-kovarians sampel, **X** adalah variabel asli.

Setelah didapat komponen-komponen utamanya, berdasarkan kriteria pemilihan komponen utama, yaitu proporsi kumulatif keragaman data asal yang dijelaskan oleh w komponen utama minimal 80% dan dilihat dari *scree plot* maka akan didapat dimensi data baru yang telah direduksi.

3.2 Jarak *Mahalanobis* Komponen Utama

Identifikasi *outlier* pada data multivariat, biasanya didasarkan pada Jarak *Mahalanobis*. Jarak *Mahalanobis* pada komponen utama adalah jarak antara masing-masing vektor komponen utama dengan vektor rata-rata setiap komponen utama. Jarak *Mahalanobis* komponen utama didefinisikan sebagai berikut:

$$(MD_i)^2 = (\mathbf{L}_i - \bar{\mathbf{W}})'\mathbf{S}^{-1}(\mathbf{L}_i - \bar{\mathbf{W}}), i = 1, 2, \ldots, n$$

dengan:

MD_i adalah Jarak *Mahalanobis* pada pengamatan ke-i

$\mathbf{L}_i$ adalah vektor komponen utama pada pengamatan ke-i

$$\mathbf{L}_i = \begin{pmatrix} w_{i1} & w_{i2} & \ldots & w_{ip} \end{pmatrix}$$

$\bar{\mathbf{W}}$ adalah vektor rata-rata dari $\mathbf{W}_j$

$$\bar{\mathbf{W}} = \begin{pmatrix} \bar{W}_1 & \bar{W}_2 & \ldots & \bar{W}_p \end{pmatrix}$$

S adalah matriks varian-kovarian sampel

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

$\mathbf{S}^{-1}$ adalah invers dari matriks **S**

Contoh 3.2.1. Diketahui dari contoh (2.5.1) $\mathbf{W} = \begin{bmatrix} -119,8930 & 46,3561 \\ -84,6300 & 45,3111 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -38,6630 & 37,3833 \\ -123,5150 & 24,3980 \\ -115,0660 & 45,3967 \end{bmatrix}$

$$\begin{aligned} (MD_1)^2 &= \left[\begin{pmatrix} -119,89 \\ 46,36 \end{pmatrix} - \begin{pmatrix} -99,07 \\ 38,50 \end{pmatrix} \right]' \begin{bmatrix} 0,00123 & 0 \\ 0 & 0,01725 \end{bmatrix} \\ &\quad \left[\begin{pmatrix} -119,89 \\ 46,36 \end{pmatrix} - \begin{pmatrix} -99,07 \\ 38,50 \end{pmatrix} \right] \\ &= \begin{bmatrix} -20,8268 & 7,8526 \end{bmatrix} \begin{bmatrix} 0,00123 & 0 \\ 0 & 0,01725 \end{bmatrix} \begin{bmatrix} -20,8268 \\ 7,8526 \end{bmatrix} \\ &= \begin{bmatrix} -0,0255 & 0,1355 \end{bmatrix} \begin{bmatrix} -20,8268 \\ 7,8526 \end{bmatrix} \\ MD_1 &= \sqrt{1,5952} = 1,2630 \end{aligned}$$

3.3 Penduga MCD

Identifikasi *outlier* dengan Jarak *Mahalanobis* tidak maksimal jika data mengandung lebih dari satu pengamatan *outlier*. Hal tersebut yang mengakibatkan adanya pengaruh *masking* dan *swamping*. *Masking* terjadi saat pengamatan *outlier* tidak terdeteksi karena adanya pengamatan *outlier* lain yang berdekatan. *Swamping* terjadi saat pengamatan yang bukan merupakan *outlier* teridentifikasi sebagai *outlier*.

Masking dan *swamping*, keduanya dapat diatasi dengan menggunakan penaksir *robust* untuk vektor rata-rata dan matriks kovariansi, sehingga dihasilkan jarak *robust*. Salah satu penaksir *robust* yang memiliki kemampuan mengukur jarak sekaligus mendeteksi titik *leverage* adalah MCD.

Metode MCD diperkenalkan oleh Rousseeuw dan Van Driessen pada tahun 1985. Tujuan dari metode ini untuk mendapatkan h dari keseluruhan pengamatan n, yang matriks varian kovariansnya memiliki determinan terkecil diantara semua kombinasi kemungkinan data, dengan

$$h = \frac{n + w + 1}{2} \tag{3.3}$$

Berdasarkan pengamatan 3.3, maka terdapat kombinasi pengamatan matriks himpunan bagian data dari matriks pengamatan $\mathbf{X}$ sebanyak r, dimana:

$$r = C_h^n$$

Dari setiap kombinasi kemungkinan data yang ada, dicari nilai determinan terkecil dari setiap matriks varian-kovariansnya. Sehingga diperoleh matriks himpunan bagian data $\mathbf{H}$ yang memiliki nilai detreminan matriks varian-kovarians

terkecil, sebagai berikut:

$$\begin{bmatrix} w_{11} & w_{12} & \dots & w_{1w} \\ w_{21} & w_{22} & \dots & w_{2w} \\ \vdots & \vdots & \ddots & \vdots \\ w_{h1} & w_{h2} & \dots & w_{hw} \end{bmatrix}$$

dari matriks himpunan sebanyak r, akan diperoleh matriks himpunan yang memiliki determinan matriks varian-kovarian terkecil dan nilai rata-ratanya, atau disebut dengan penduga MCD yaitu $\bar{\mathbf{M}}_{MCD}$ dan $\mathbf{S}_{MCD}$, dengan rumus:

$$\bar{\mathbf{M}}_{MCD} = \frac{1}{h}(\mathbf{H}'\mathbf{V}^*)$$

$$\mathbf{S}_{MCD} = \frac{1}{h-1}(\mathbf{H} - \mathbf{V}^*(\bar{\mathbf{M}}_{MCD})')'(\mathbf{H} - \mathbf{V}^*(\bar{\mathbf{M}}_{MCD})')$$

dengan $\mathbf{V}^*$ adalah matriks satuan berukuran $h \times 1$.

$$\mathbf{V}^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Contoh 3.3.1. Berdasarkan dari contoh (2.5.1), didapat variabel tereduksi yang menggunakan Metode PCA adalah sebagai berikut:

Tabel 3.1: Komponen Utama terpilih

W1	W2
-119,8930	46,3561
-84,6300	45,3111
-92,9170	34,8342
-121,8690	41,2405
-95,9770	33,1084
-38,6630	37,3833
-123,5150	24,3980
-115,0660	45,3967

diperoleh nilai $h = 5$, sehingga $r = 56$ kemungkinan. Beberapa kemungkinannya adalah

$$\mathbf{H_1} = \begin{bmatrix} -119,8930 & 46,3561 \\ -84,6300 & 45,3111 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \end{bmatrix}; \mathbf{H_2} = \begin{bmatrix} -84,6300 & 45,3111 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -38,6630 & 37,3833 \end{bmatrix};$$

$$\mathbf{H_3} = \begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix}$$

Dari berbagai kemungkinan tersebut, kemudian dicari determinan terkecil dari masing-masing matriks kovariannya. Sehingga didapat matriks $\mathbf{H}$ yang determinan matriks varian-kovariansnya minimum, adalah:

$$\mathbf{H} = \begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix}$$

sehingga cara untuk menghitung $\bar{\mathbf{M}}_{MCD}$ adalah sebagai berikut:

$$\begin{aligned}
\bar{\mathbf{M}}_{MCD} &= \frac{1}{5}\left(\begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix}\right)' \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\
&= \frac{1}{5}\begin{bmatrix} -119,8930 & -92,9170 & -121,8690 & -95,9770 & -115,0660 \\ 46,3561 & 34,8342 & 41,2405 & 33,1084 & 45,3967 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\
&= \frac{1}{5}\begin{bmatrix} -545,7220 \\ 200,9359 \end{bmatrix} \\
&= \begin{bmatrix} -109,1444 \\ 40,1872 \end{bmatrix} \\
\mathbf{S}_{MCD} &= \frac{1}{4}\left(\begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} -109,1444 & 40,1872 \end{bmatrix}\right)'
\end{aligned}$$

$$
\left(\begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -109,1444 & 40,1872 \end{bmatrix} \right)
$$

$$
= \frac{1}{4} \left(\begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix} - \begin{bmatrix} -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \end{bmatrix} \right)'
$$

$$
\left(\begin{bmatrix} -119,8930 & 46,3561 \\ -92,9170 & 34,8342 \\ -121,8690 & 41,2405 \\ -95,9770 & 33,1084 \\ -115,0660 & 45,3967 \end{bmatrix} - \begin{bmatrix} -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \\ -109,1444 & 40,1872 \end{bmatrix} \right)
$$

$$
= \frac{1}{4} \begin{pmatrix} -10,7486 & 16,2274 & -12,7246 & 13,1674 & -5,9216 \\ 6,1689 & -5,3530 & 1,0533 & -7,0788 & 5,2095 \end{pmatrix} \begin{pmatrix} -10,7486 & 6,1689 \\ 16,2274 & -5,3530 \\ -12,7246 & 1,0533 \\ 13,1674 & -7,0788 \\ -5,9216 & 5,2095 \end{pmatrix}
$$

$$
= \frac{1}{4} \begin{pmatrix} 749,2221 & -290,6331 \\ -290,6331 & 145,0677 \end{pmatrix}
$$

$$
= \begin{pmatrix} 187,3055 & -72,6583 \\ -72,6583 & 36,2669 \end{pmatrix}
$$

3.4 Pendeteksian *Outlier* dengan MCD

$\mathbf{S}_{MCD}$ adalah penduga matriks varian-kovarian untuk MCD. Metode MCD memiliki kemampuan mengukur jarak *robust* yang dapat digunakan untuk mendeteksi *outlier leverage*.

Jarak *robust* adalah suatu pendekatan untuk mendeteksi *outlier* pada data regresi linear berganda, yaitu dengan menggunakan penduga dari $\bar{\mathbf{M}}_{MCD}$ dan $\mathbf{S}_{MCD}$ pada metode *robust*. Oleh karena itu metode ini mampu meminimumkan pengaruh dari adanya efek *masking* dan *swamping* dalam pendeteksian *outlier*. Ada beberapa penyebab munculnya *oulier*, salah satunya disebabkan oleh variabel bebas, yang dinamakan *oulier leverage*. *Outlier leverage* dideteksi dengan menggunakan jarak *robust* (RD_i) untuk setiap pengamatan ke-i. Jarak *robust* didefinisikan oleh persamaan berikut:

$$(RD_i)^2 = (\mathbf{L}_i - \bar{\mathbf{M}}_{MCD})'\mathbf{S}_{MCD}^{-1}(\mathbf{L}_i - \bar{\mathbf{M}}_{MCD}), i = 1, 2, \ldots, n \tag{3.4}$$

dengan:

RD_i adalah jarak *robust* untuk setiap pengamatan ke-i

$\mathbf{L}_i$ adalah vektor komponen utama pada pengamatan ke-i

$$\mathbf{L}_i = \begin{pmatrix} w_{i1} & w_{i2} & \ldots & w_{ip} \end{pmatrix}$$

$\bar{\mathbf{M}}_{MCD}$ adalah vektor rata-rata dari $\mathbf{W}_j$ dengan metode MCD

$$\bar{\mathbf{M}} = \begin{pmatrix} \bar{W}_1 & \bar{W}_2 & \ldots & \bar{W}_w \end{pmatrix}$$

$\mathbf{S}_{MCD}$ adalah matriks varian kovarian sampel dengan metode MCD

$$\mathbf{S}_{MCD} = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1w} \\ s_{21} & s_{22} & \ldots & s_{2w} \\ \vdots & \vdots & \ddots & \vdots \\ s_{w1} & s_{w2} & \ldots & s_{ww} \end{bmatrix}$$

$\mathbf{S}_{MCD}^{-1}$ adalah invers dari matriks $\mathbf{S}_{MCD}$

Contoh 3.4.1. Dari contoh sebelumnya, setelah diperoleh

$$\mathbf{S}_{MCD} = \begin{bmatrix} 187,3055 & -72,6583 \\ -72,6583 & 36,2669 \end{bmatrix}$$

maka didapat $\mathbf{S}_{MCD}^{-1} = \begin{bmatrix} 0,0240 & 0,0480 \\ 0,0480 & 0,1237 \end{bmatrix}$, sehingga untuk menghitung jarak *robust* berdasarkan rumus (3.4) diperoleh perhitungan sebagai berikut:

$$\begin{aligned}
(RD_1)^2 &= \left[\begin{pmatrix} -119,8930 \\ 46,3561 \end{pmatrix} - \begin{pmatrix} -109,1444 \\ 40,1872 \end{pmatrix} \right]' \begin{bmatrix} 0,0240 & 0,0480 \\ 0,0480 & 0,1237 \end{bmatrix} \\
&\quad \left[\begin{pmatrix} -119,8930 \\ 46,3561 \end{pmatrix} - \begin{pmatrix} -109,1444 \\ 40,1872 \end{pmatrix} \right] \\
&= \begin{bmatrix} -10,7486 & 6,1689 \end{bmatrix} \begin{bmatrix} 0,0240 & 0,0480 \\ 0,0480 & 0,1237 \end{bmatrix} \begin{bmatrix} -10,7486 \\ 6,1689 \end{bmatrix} \\
&= \begin{bmatrix} 0,0386 & 0,2474 \end{bmatrix} \begin{bmatrix} -10,7486 \\ 6,1689 \end{bmatrix} \\
RD_1 &= \sqrt{1,1115} = 1,0543
\end{aligned}$$

sehingga didapat jarak *robust* untuk setiap pengamatan adalah

Tabel 3.2: Jarak *robust* untuk setiap pengamatan

Pengamatan	*RDi*
1	1,0543
2	5,4502
3	1,2311
4	1,6522
5	1,1859
6	10,0507
7	7,5879
8	1,1121

Pendeteksian *outlier leverage* menggunakan jarak *robust* (RD_i) untuk setiap pengamatan ke-i dapat dituliskan sebagai berikut:

$$leverage = \begin{cases} \text{jika } RD_i \leq C, \text{ maka pengamatan bukan } outier \text{ (diberi kode 0)} \\ \text{jika } RD_i > C, \text{ maka pengamatan merupakan } outier \text{ (diberi kode 1)} \end{cases}$$

dengan $C = \sqrt{\chi^2_{p;1-\alpha}}$, dimana $\alpha = 0,975$ dan C dinyatakan sebagai nilai *cut-off*. Nilai *cut-off* adalah suatu nilai yang digunakan untuk menentukan apakah suatu pengamatan dideteksi sebagai *outlier* atau bukan. Notasi $\chi^2_{p;1-\alpha}$ adalah nilai χ^2 yang membuat luas diujung kanan distribusinya sama dengan $1 - \alpha$ dan RD_i adalah jarak *robust*untuk setiap pengamatan ke-i.

Outlier dapat dideteksi dan diklasifikasikan dengan *diagnostic plot* atau peta *outlier* yang berguna untuk membedakan data pengamatan. Pada *diagnostic plot*, data pengamatan dibedakan menjadi empat tipe, yaitu *bad leverage*, *outlier* orthogonal, pengamatan biasa, dan *good leverage*.

Bad leverage (terdapat pada kuadran 1) adalh suatu titik yang memiliki nilai jarak *robust* dan nilai jarak mahalanobis lebih besar dari nilai *cut-off*. Titik ini adalah jenis *outlier* yang sangat berpengaruh, tetapi tidak tidak cocok untuk model regresi. Dengan adanya titik ini dapat merubah garis regresi sehingga dapat mempengaruhi hasil secara keseluruhan. Karena akibat yang dapat ditimbulkan oleh titik ini, maka yang sebaiknya dilakukan adalah menghapus pengamatan yang tergolong ke dalam jenis *bad leverage* tersebut.

Outlier orthogonal (terdapat pda kuadran 2), adalah titik yang memiliki nilai jarak *robust* lebih besar dari nilai *cut-off* dan nilai jarak mahalanobis lebih kecil atau sama dengan nilai *cut-off*, pengamatan ini dapat dihapus.

Pengamatan biasa (terdapat pada kuadran 3), adalah titik yang memiliki

nilai jarak *robust* dan nilai jarak mahalanobis lebih kecil atau sama dengan nilai *cut-off*. Pengamatan ini bukan merupakan *outlier*.

Good leverage (terdapat pada kuadran 4), adalah titik yang memiliki nilai jarak *robust* lebih kecil atau sama dengan nilai *cut-off* dan nilai jarak mahalanobis lebih besar dari nilai *cut-off*. Titik ini berada jauh dari pengamatan biasa, namun tetap mengikuti garis regresi. Pengamatan jenis ini tetap dapat diikutsertakan dalam analisis data.

3.5 Aplikasi Metode *Robust* PCA

Metode *robust* PCA ini akan diaplikasikan pada data tingkat kepuasan pelanggan yang diperoleh dari HATCO DATA SET (Hair, et al., 1998) dikutip dalam Jurnal Korelasi Kanonikal (Siregar, 2008) dan makalah Pencilan (*Outlier*) (Soemartini, 2007), yang dilampirkan pada Lampiran A. Untuk mendeteksi apakah data pada Lampiran A mengandung multikolinearitas dan *outlier* atau tidak, maka langkah-langkah yang dilakukan adalah mendeteksi multikolinearitas, melakukan analisis komponen utama, menghitung jarak mahalanobis, menghitung jarak *robust*, dan membuat *diagnostic plot* untuk mengetahui jenis-jenis *outlier*. AKU menghasilkan *score* komponen utama yang dijadikan variabel baru selanjutnya yang digunakan untuk mendeteksi *outlier* dengan metode MCD.

1. Analisis Regresi Linear Berganda

   Pada data aplikasi didapat model regresi linear berganda sebagai berikut:

$$\begin{aligned} Y &= -452 + 109X_1 + 123X_2 + 13,2X_3 + 16,4X_4 - 234X_5 + 18,1X_6 + \\ &\quad 0,3X_7 + 0,39X_8 + 4,0X_9 + 2,78X_{10} + 22,2X_{11} + 6,01X_{12} \end{aligned}$$

Adanya gejala multikolinearitas, dapat dilihat dari nilai VIF sebagai berikut:

Tabel 3.3: Uji multikolinearitas koefisien regresi linear berganda

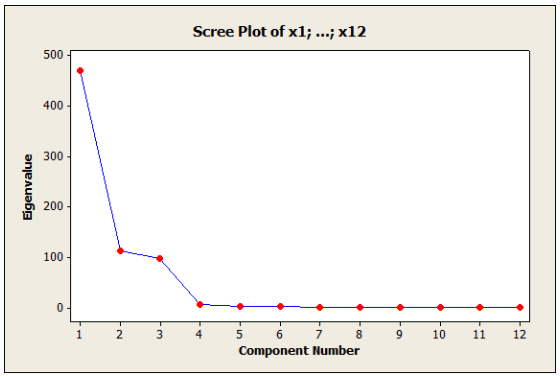
Predictor	Coef	SE Coef	T	P	VIF
Constant	-452,20	176,70	-2,56	0,01	
X1	108,61	62,08	1,75	0,09	46,80
X2	122,91	64,54	1,90	0,06	43,40
X3	13,18	9,65	1,37	0,18	1,80
X4	16,40	10,42	1,57	0,12	1,20
X5	-234,40	120,40	-1,95	0,06	55,90
X6	18,11	11,22	1,61	0,11	1,30
X7	0,31	10,40	0,03	0,98	2,00
X8	0,39	1,66	0,23	0,82	2,00
X9	4,02	19,73	0,20	0,84	7,10
X10	2,78	0,71	3,91	0,00	1,60
X11	22,19	36,86	0,60	0,55	11,20
X12	6,01	5,15	1,17	0,25	21,40

Dari tabel di atas terdapat beberapa indikasi adanya multikolinearitas, yaitu terdapat 5 variabel (X1, X2, X5, X11, X12) yang nilai VIFnya lebih besar dari 10.

2. Reduksi Dimensi Data

   Selanjutnya dilakukan reduksi dimensi dengan analisis komponen utama sebagai tahap awal Metode *Robust* PCA. Berikut ini merupakan analisis komponen utama yang didasarkan pada matriks varian kovarian:

Eigenvalue	469,470	112,490	97,020	5,180	1,780	1,610
Proportion	0,680	0,163	0,140	0,007	0,003	0,002
Cumulative	0,680	0,842	0,983	0,990	0,993	0,995
Eigenvalue	1,030	0,820	0,720	0,610	0,100	0,010
Proportion	0,001	0,001	0,001	0,001	0,000	0,000
Cumulative	0,997	0,998	0,999	1,000	1,000	1,000



Gambar 3.2: Scree Plot

Berdasarkan kriteria pemilihan komponen utama, komponen utama ke-1 (W_1) sampai komponen utama ke-3 (W_3) telah menjelaskan 98,3% keragaman dari Y, dan dari *scree plot* yang ada dapat dipilih 3 komponen utama. Karena jumlah proporsi kumulatif 3 komponen utama sudah lebih dari 80%, jadi cukup 3 komponen utama saja yang dipilih. Sehingga, dari analisis tersebut hasil variabel tereduksi atau *score* komponen utama untuk setiap pengamatan adalah sebagai berikut. Lebih lengkap pada Lampiran B.

Tabel 3.4: *Score* komponen utama untuk setiap pengamatan

Pengamatan	W1	W2	W3
1	121,4590	22,9131	14,4579
2	77,0200	6,4680	34,1029
3	71,7730	8,7992	39,5021
4	45,8560	13,8135	25,0029
5	90,6290	15,8577	46,3819
6	87,3070	14,4708	33,1895
7	68,9050	11,2603	37,5094
8	86,1850	6,3815	33,9886
9	99,9310	14,3399	50,5299
10	100,0750	5,8741	42,9440
⋮	⋮	⋮	⋮
54	118,6580	18,3366	70,7113

dengan nilai VIF masing-masing komponen utama adalah:

Tabel 3.5: Uji multikolinearitas koefisien regresi linear berganda

Predictor	Coef	SE Coef	T	P	VIF
Constant	-233,76	66,51	-3,51	0,00	
W1	4,15	0,55	7,49	0,00	1,00
W2	7,18	1,13	6,34	0,00	1,00
W3	-0,62	1,22	-0,51	0,61	1,00

Dari variabel data yang sudah direduksi, dapat dilihat nilai VIF di atas, dari masing-masing komponen utama bernilai 1, hal ini menandakan multikolinearitas sudah teratasi.

3. Menghitung Jarak *Mahalanobis*

   Sebelum menghitung Jarak *Mahalanobis*, terlebih dahulu menghitung varian kovarian dari *score* komponen utama yang diperoleh dari AKU. Berikut ini merupakan hasil perhitungan Jarak *Mahalanobis* untuk setiap pengamatan. Lebih lengkap pada Lampiran C.

Tabel 3.6: Jarak *Mahalanobis*

Pengamatan	*Mdi*
1	2,853313
2	0,848861
3	0,890131
4	2,040753
5	1,167281
6	0,278948
7	0,787635
8	0,809424
9	1,712038
10	1,364767
⋮	⋮
54	3,948446

4. Penduga MCD

   Kemudian, dari data tereduksi didapat nilai $h = 29$ dan didapat banyaknya kemungkinan C_{29}^{54}, sehingga akan didapat himpunan kemungkinan data yang matriks varian kovariannya memiliki determinan terkecil diantara semua kombinasi kemungkinan data. Dengan bantuan *Software* MATLAB 7.7.0 (R2008b) melalui program pada Lampiran D, didapat nilai rata-rata dan matriks varian kovarian dari metode MCD pada tabel berikut.

Tabel 3.7: Nilai rata-rata dan matriks varian kovarian dengan Metode MCD

	Matriks Varian Kovarian			Nilai Rata-Rata
	W1	W2	W3	
W1	346,3510	-19,3531	-15,7072	83,5816
W2	-19,3531	45,6373	16,2576	12,4230
W3	-15,7072	16,2576	68,9533	35,5834

5. Pendeteksian *Outlier*

   Setelah didapat nilai rata-rata dan matriks varian kovarian dengan metode MCD, selanjutnya adalah menghitung Jarak *Robust*. Berikut ini adalah hasil perhitungan Jarak *Robust* untuk setiap pengamatan. Lebih lengkap pada Lampiran E.

Tabel 3.8: Jarak *Robust*

Pengamatan	*RDi*
1	4,084021
2	1,011843
3	1,090890
4	2,534878
5	1,413292
6	0,545576
7	0,877645
8	0,897073
9	2,095328
10	1,761085
⋮	⋮
54	4,828416

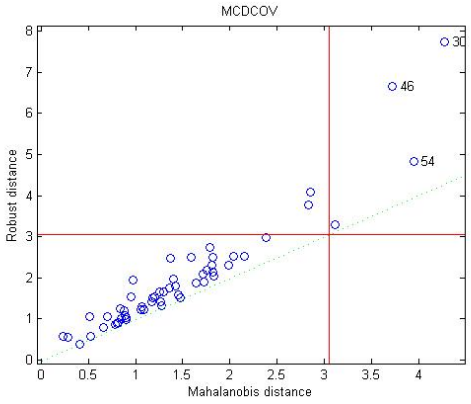
Nilai *cut-off* pada Jarak *Robust* adalah $\sqrt{\chi^2_{3;\alpha}} = \sqrt{\chi^2_{3;0,025}} = 3,057515915$. Nilai Jarak Robust yang lebih besar dari nilai *cut-off* dideteksi sebagai *outlier*. Sedangkan nilai jarak *robust* yang lebih kecil dari nilai *cut-off* bukan dideteksi sebagai *outlier*. Asumsikan pengamatan yang terdeteksi *outlier* diberi nilai 1 dan yang bukan *outlier* diberi nilai 0. Hasil pendeteksian *outlier* untuk setiap pengamatan pada jarak *robust* dapat dilihat pada tabel berikut. Lebih lengkap pada Lampiran F.

Tabel 3.9: Pendeteksian *Outlier* untuk setiap pengamatan jarak *robust*

Pengamatan	Pendeteksian *Outlier*
1	1
2	0
3	0
4	0
5	0
⋮	⋮
54	1

6. Membentuk *Diagnostic Plot*

   Berdasarkan tabel pendeteksian *outlier*, terdapat 6 pengamatan yang terdeteksi sebagai *outlier*. Dapat dilihat pada *diagnostic plot* sebagai berikut:



Gambar 3.3: *Diagnostic Plot*

Dari *Diagnostic Plot* di atas, untuk mengklasifikasi *outlier*, dapat dibuat *plot diagnostic* antara Jarak *Mahalanobis* dan Jarak *Robust*. Pada 2 *outlier* Jarak *Mahalanobis*-nya lebih kecil dari nilai *cut-off* dan Jarak *Robust*-nya lebih besar dari nilai *cut-off* maka 2 *outlier* tersebut termasuk kedalam klasifikasi *Outlier* ortogonal, dan terdapat 4 *outlier* yang Jarak *Mahalanobis* dan Jarak *Robust*-nya lebih besar dari nilai *cut-off* maka 4 *outlier* tersebut termasuk kedalam klasifikasi *Bad Leverage*, sehingga semua *outlier* tersebut harus dihilangkan dari data pengamatan.

7. Membentuk model komponen utama pada *outlier* yang dihilangkan
   Setelah semua *outlier* dihilangkan, didapat model regresi komponen utama sebagai berikut:

$$Y = -253 - 4,04W_1 - 2,58W_2 - 7,48W_3$$

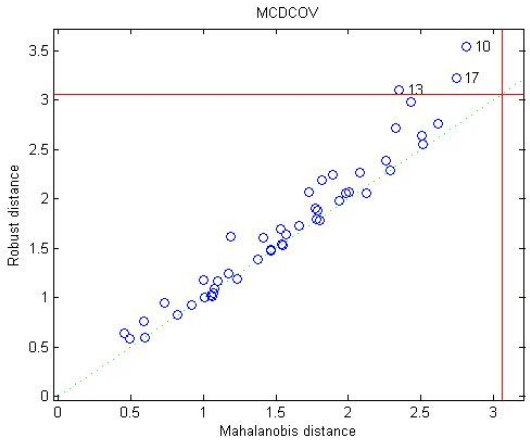
dengan nilai VIF masing-masing komponen utama, adalah sebagai berikut:

Tabel 3.10: Uji multikolinearitas koefisien regresi linear berganda

Predictor	Coef	SE Coef	T	P	VIF
Constant	-253,16	84,31	-3,00	0,00	
W1	-4,04	0,68	-5,96	0,00	1,00
W2	-2,58	1,60	-1,61	0,11	1,00
W3	-7,48	2,08	-3,59	0,00	1,00

dari nilai VIF di atas, masing-masing variabel komponen utama nilai VIF nya sudah 1, sehingga data sudah tidak mengandung multikolinearitas, dan R^2 sebesar 64%, yang artinya keragaman data pada variabel terikat yang dapat dijelaskan oleh variabel bebas adalah 64%.

Kemudian dilakukan cara yang sama untuk mendeteksi apakah masih terdapat *outlier* atau tidak, dan didapat *diagnostic plot* sebagai berikut:



Gambar 3.4: *Diagnostic Plot*

dari *diagnostic plot* tersebut dapat dilihat masih ada *outlier* pada data, namun jumlah *outlier* lebih sedikit dibandingkan dengan data awal. Terdapat 3 *outlier* yang termasuk kedalam klasifikasi *Outlier* ortogonal dan harus dihilangkan.

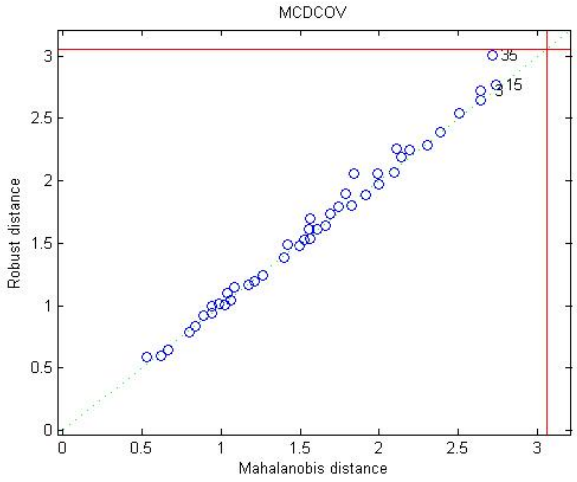
Karena *outlier* masih ada, namun tidak sebanyak *outlier* sebelum dilakukan Metode *Robust* PCA, maka dilakukan metode yang sama untuk menghilangkannya. Kemudian setelah *outlier* dihilangkan, didapat hasil sebagai berikut:

$$Y = -266 + 4,37W_1 + 3,69W_2 - 6,78W_3$$

Tabel 3.11: Uji multikolinearitas koefisien regresi linear berganda

Predictor	Coef	SE Coef	T	P	VIF
Constant	-266,42	92,16	-2,89	0,01	
W1	4,37	0,75	5,84	0,00	1,00
W2	3,69	1,68	2,20	0,03	1,00
W3	-6,78	2,53	-2,68	0,01	1,00

dari persamaan regresi komponen utama di atas dapat dilihat semua nilai VIF adalah 1, serta nilai R^2 juga sudah meningkat menjadi 66,1%, yang artinya keragaman data pada variabel terikat yang dapat dijelaskan oleh variabel bebas adalah 66,1%.



Gambar 3.5: *Diagnostic Plot*

dari *diagnostic plot* tersebut dapat dilihat, ternyata sudah tidak ada *outlier* pada data. Sehingga dari analisis di atas, multikolinearitas dan *outlier* telah teratasi.

Dari persamaan regresi komponen utama di atas, dilakukan transformasi balik ke regresi linear berganda, sehingga didapat persamaan regresi linear berganda adalah sebagai berikut:

$$\begin{aligned} Y &= -266 + 0,16X_1 - 0,28X_2 + 0,50X_3 - 0,04X_4 - 0,02X_5 - 0,15X_6 - \\ &\quad 0,52X_7 + 1,32X_8 + 1,32X_9 + 3,44X_{10} + 0,90X_{11} + 7,87X_{12} \end{aligned}$$

BAB IV

PENUTUP

4.1 Kesimpulan

Berdasarkan pembahasan mengenai Metode *Robust* PCA, dimana metode yang kuat (*robust*) untuk PCA, dapat mengatasi masalah multikolinearitas dan *outlier* pada model regresi linear berganda. Dengan melihat hasil pada nilai VIF, *diagnostic plot*, dan perubahan nilai R^2.

Dapat dilihat melalui Aplikasi Metode *Robust* PCA pada sub bab 3.5. dengan menggunakan *estimator covariance robust* yaitu *Minimum Covariance Determinant* (MCD) pada Metode *Robust* PCA, yaitu:

1. Pada data tingkat kepuasan pelanggan, didapat model regresi linear berganda sebagai berikut:

$$\begin{aligned} Y &= -452 + 109X_1 + 123X_2 + 13,2X_3 + 16,4X_4 - 234X_5 + 18,1X_6 + \\ &\quad 0,3X_7 + 0,39X_8 + 4,0X_9 + 2,78X_{10} + 22,2X_{11} + 6,01X_{12} \end{aligned}$$

2. Dari model tersebut terdapat 5 variabel yang mengandung multikolinearitas.

3. Dilakukan langkah awal Metode *Robust* PCA dengan mereduksi dimensi data melalui Metode PCA, dan didapat penduga matriks varian kovarian dan rata-rata dengan penduga MCD sebagai berikut:

Tabel 4.1: Nilai rata-rata dan matriks varian kovarian dengan Metode MCD

	Matriks Varian Kovarian			Nilai Rata-Rata
	W1	W2	W3	
W1	346,3510	-19,3531	-15,7072	83,5816
W2	-19,3531	45,6373	16,2576	12,4230
W3	-15,7072	16,2576	68,9533	35,5834

4. Dari penduga matriks varian kovarian dan rata-rata dengan MCD, dapat dideteksi adanya *outlier* pada data sebanyak 6 *outlier*. Dilakukan uji ulang, total menjadi 9 *outlier*

5. Setelah multikolineritas dan *outlier*-nya teratasi menggunakan Metode *Robust* PCA, didapat model regresi linear berganda terbaik, sebagai berikut:

$$\begin{aligned} Y &= -266 + 0,16X_1 - 0,28X_2 + 0,50X_3 - 0,04X_4 - 0,02X_5 - 0,15X_6 - \\ &\quad 0,52X_7 + 1,32X_8 + 1,32X_9 + 3,44X_{10} + 0,90X_{11} + 7,87X_{12} \end{aligned}$$

dengan:

(a) Nilai VIF pada masing-masing komponen utama adalah 1, yang berarti data sudah tidak mengandung multikolinearitas.

(b) Dari hasil *diagnostic plot* akhir, dimana Jarak *Mahalanobis* dan Jarak *Robust* sudah tidak ada yang melebihi nilai *Cut-Off*, sehingga *outlier* sudah teratasi.

(c) Nilai R^2 pada model juga sudah meningkat menjadi 66,1%, yang artinya keragaman data pada variabel terikat yang dijelaskan oleh variabel bebas adalah 66,1% saat multikolinearitas dan *outlier* telah teratasi, sedangkan sisanya sebesar 33,9% dijelaskan oleh variabel lainnya yang tidak dimasukkan ke dalam model regresi dalam pada penelitian ini.

4.2 Saran

1. Untuk penelitian selanjutnya, jika data amatan $n > 600$ dapat menggunakan Metode FastMCD.

2. Bagi peneliti lain yang ingin meneliti tentang *outlier*, dapat menggunakan *estimator covariance robust* yang lain seperti metode *Minimum Volume Ellipsoid* (MVE) dan metode *Welsch* untuk kemudian dibandingkan tingkat efisiensinya dengan metode MCD.

DAFTAR PUSTAKA

Anton, Howard dan Rorres, Chris. 2004. *Aljabar Linear Elementer Versi Aplikasi*. Edisi Kedelapan/jilid 1. Diterjemahkan oleh Refina Indriasari dan Irzam Harmein. Jakarta: Erlangga.

Draper, N.R. and H. Smith. 1992. *Analisis Regresi Terapan*. Edisi Kedua. Diterjemahkan oleh Bambang Sumantri. Jurusan Statistika FMIPA IPB. Jakarta: PT Gramedia Pustaka Utama Jakarta.

Gujarati N, Damorar. 1995. *Ekonometrika Dasar*. Jakarta: Erlangga.

Hadi, A.S. Imon, A.H.M.R, & Werner, M. 2009. *Detection of Outliers. WIREs Computational Statistics, 1, 57-70.*

Irfagutami, Ni Putu N. Srinadi, I Gusti A M. dan Sumarjaya, I Wayan. 2014. Perbandingan Regresi *Robust* Penduga MM dengan Metode *Random Sample Consensus* dalam Menangani Pencilan. *e-Jurnal Matematika. Vol.3 No.2. 45-52* . Universitas Udayana, Bukit Jimbaran-Bali.

Johnson, R. A. Dan Wichern, D. W., 2007, *Applied Multivariate Statistical Analysis, 6th edition*. New Jersey: Printice Hall.

Neter, J., Wasserman, W. & Kutner, M. H. 1990. *Applied Linear Regression Models*. Homewood, IL: Irwin.

Nurcahyadi, Heru. 2010. *Analisis Regresi pada Data Outlier dengan menggunakan Least Trimmed Square (LTS) dan MM-Estimasi*. Skripsi. Universitas Islam Negeri Syarif Hidayatullah.

O'Brien, R M. 2007. *A Caution Regarding Rules of Thumb for Variance Inflation Factor*. Departement of Sociology of Oregon, Eugene, USA.

Prasetyo, Haris B. 2008. *Analisis Regresi Komponen Utama untuk Mengatasi Ma-*

salah Multikolinearitas Dalam Analisis Regresi Linear Berganda. Skripsi. Universitas Negeri Jakarta.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. Universitas Michigan, Wiley.

Primananda, Taufiq. Mulyono, Slamet. dan Prastyo, Dedy Dwi. 2010. Pengendalian Kualitas Produksi Mebel di PT. MAJAWANA dengan Diagram Kontrol D^2 (*Mahalanobis Distance*). Jurusan Statistika FMIPA ITS, Surabaya.

Ridwan, Ponco. 2008. *Analisis Regresi Robust dengan Menggunakan Metode Penduga-M*. Skripsi. Universitas Negeri Jakarta.

Sembiring, R K. 2003. *Analisis Regresi*. Edisi Kedua. Bandung: Penerbit ITB.

Sifriyani. 2011. Metode *Minimum Covariance Determinan* pada Analisis Regresi Linier Berganda dengan Kasus Pencilan. *Jurnal Eksponensial. Vol.2 No.2*. Prodi Statistika FMIPA Universitas Mulawarman. ISSN 2085-7829

Siregar, Suzzana L. 2008. *Korelasi Kanonikal Komputasi dengan Menggunakan SPSS dan Interpretasi Hasil*. Fakultas Ekonomi Universitas Gunadarma. Jakarta.

Soemartini. 2007. *Pencilan (Outlier)*. Jurusan Statistika. FMIPA UNPAD. Jatinangor.

Sunaryo, S. 2011. Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Pendekatan ROBPCA (Studi Kasus: Angka Kematian Bayi di Jawa Timur). *Jurnal Matematika, Saint dan Teknologi. Vol.12 No.1. pp. 1-10*. Jurusan Statistika, ITS.

Verboven, S. dan Hubert, M. 2016. *LIBRA: a MATLAB Library for Robust Analysis*. [ONLINE]. Tersedia: https://wis.kuleuven.be/stat/robust/LIBRA. [20 Mei 2017]

LAMPIRAN-LAMPIRAN

Lampiran A

Keterangan:
Y: Tingkat kepuasan pelanggan
X1: Kecepatan pengantaran
X2: Kecepatan pelayanan
X3: Fleksibilitas harga dari supplier
X4: Citra produsen
X5: Layanan keseluruhan
X6: Citra tenaga penjual
X7: Kualitas produk
X8: Tingkat keuntungan
X9: Fleksibilitas harga
X10: Tingkat harga
X11: Kualitas bahan baku
X12: Tingkat penggunaan

Tabel 4.2: Data Pengamatan

Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
509	4,10	0,60	15,00	4,70	2,40	2,30	5,20	32,00	7,80	115,00	4,30	33,54
80	1,80	3,00	6,30	6,60	2,50	4,00	8,40	43,00	5,80	72,00	1,42	8,24
101	3,40	5,20	5,70	6,00	4,30	2,70	8,20	48,00	5,10	66,00	1,70	8,67
101	2,70	1,00	7,10	5,90	1,80	2,30	7,80	32,00	6,50	41,00	2,01	13,07
204	6,00	0,90	9,60	7,80	3,40	4,60	4,50	58,00	7,40	83,00	2,16	15,98
200	1,90	3,30	7,90	4,80	2,60	1,90	9,70	45,00	6,70	81,00	2,59	17,35
80	4,60	2,40	9,50	6,60	3,50	4,50	7,60	46,00	5,70	63,00	1,91	10,89
127	1,30	4,20	6,20	5,10	2,80	2,20	6,90	44,00	3,70	81,00	2,57	9,51
202	5,50	1,60	9,40	4,70	3,50	3,00	7,60	63,00	6,00	92,00	2,50	15,00
203	4,00	3,50	6,50	6,00	3,70	3,20	8,70	54,00	3,70	94,00	2,40	8,88
329	2,40	1,60	8,80	4,80	2,00	2,80	5,80	32,00	6,30	83,00	4,13	26,02
65	3,90	2,20	9,10	4,60	3,00	2,50	8,30	47,00	6,70	43,00	1,86	12,46
217	2,80	1,40	8,10	10,00	2,10	1,40	6,60	39,00	7,40	68,00	2,40	17,76
168	3,70	1,50	8,60	5,70	2,70	3,70	6,70	38,00	7,70	67,00	3,40	26,18
330	4,70	1,30	9,90	6,70	3,00	2,60	6,80	54,00	5,80	88,00	3,95	22,91
215	3,40	2,00	9,70	4,70	2,70	1,70	4,80	49,00	7,30	74,00	3,56	25,99
172	3,20	4,10	5,70	5,10	3,60	2,90	6,20	38,00	5,60	87,00	3,02	16,91
87	4,90	1,80	7,70	4,30	3,40	1,50	5,90	40,00	6,00	28,00	2,98	17,88
34	5,30	1,40	9,70	6,10	3,30	3,90	6,80	54,00	3,70	41,00	1,55	5,74
109	4,70	1,30	9,90	6,70	3,00	2,60	6,80	55,00	5,20	76,00	2,85	14,82

Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
70	3,30	0,90	8,60	4,00	2,10	1,80	6,30	41,00	6,70	68,00	2,10	14,07
136	3,40	0,40	8,30	2,50	1,20	1,70	5,20	35,00	3,40	53,00	1,12	3,81
830	3,00	4,00	9,10	7,10	3,50	3,40	8,40	55,00	5,80	114,00	3,95	22,91
220	2,40	1,50	6,70	4,80	1,90	2,50	7,20	36,00	5,80	86,00	3,40	19,72
276	5,10	1,40	8,70	4,80	3,30	2,60	3,80	49,00	6,30	100,00	2,95	18,59
144	4,60	2,10	7,90	5,80	3,40	2,80	4,70	49,00	5,80	73,00	3,50	20,30
181	2,40	1,50	6,60	4,80	1,90	2,50	7,20	36,00	5,20	86,00	2,56	13,31
178	5,20	1,30	9,70	6,10	3,20	3,90	6,70	54,00	5,80	59,00	2,58	14,96
72	3,50	2,80	9,90	3,50	3,10	1,70	5,40	49,00	5,20	56,00	2,71	14,09
574	4,10	3,70	5,90	5,50	3,90	3,00	8,40	46,00	11,20	90,00	5,59	62,61
71	3,00	3,20	6,00	5,30	3,10	3,00	8,00	43,00	3,20	65,00	0,74	2,37
115	2,80	3,80	8,90	6,90	3,30	3,20	8,20	53,00	5,40	70,00	2,64	14,26
295	5,20	2,00	9,30	5,90	3,70	2,40	4,60	60,00	5,80	93,00	3,30	19,14
116	3,40	3,70	6,40	5,70	3,50	3,40	8,40	47,00	5,00	73,00	3,50	17,50
58	2,40	1,00	7,70	3,40	1,70	1,10	6,20	35,00	8,70	23,00	2,52	21,92
184	1,80	3,30	7,50	4,50	2,50	2,40	7,60	39,00	5,30	99,00	2,60	13,78
118	3,60	4,00	5,80	5,80	3,70	2,50	9,30	44,00	2,60	86,00	2,05	5,33
148	4,00	0,90	9,10	5,40	2,40	2,60	7,30	46,00	5,40	88,00	1,18	6,37
151	0,00	2,10	6,90	5,40	1,10	2,60	8,90	29,00	4,80	76,00	2,45	11,76
120	2,40	2,00	6,40	4,50	2,10	2,20	8,80	28,00	4,30	120,00	2,85	12,26
95	1,90	3,40	7,60	4,60	2,60	2,50	7,70	40,00	5,20	72,00	1,84	9,57
191	5,90	0,90	9,60	7,80	3,40	4,60	4,50	58,00	3,40	93,00	1,48	5,03
123	4,90	2,30	9,30	4,50	3,60	1,30	6,20	53,00	6,50	84,00	3,00	19,50
311	5,00	1,30	8,60	4,70	3,10	2,50	3,70	48,00	4,50	106,00	3,05	13,73
75	2,00	2,60	6,50	3,70	2,40	1,70	8,50	38,00	3,60	99,00	1,30	4,68
483	5,00	2,50	9,40	4,60	3,70	1,40	6,30	54,00	8,80	88,00	6,40	56,32
153	3,10	1,90	10,00	4,50	2,60	3,20	3,80	55,00	6,70	77,00	2,85	19,10
158	3,40	3,30	5,60	5,60	3,60	2,30	9,10	43,00	5,10	77,00	2,86	14,59
313	5,80	0,20	8,80	4,50	3,00	2,40	6,70	57,00	8,80	72,00	3,20	28,16
398	5,40	2,10	8,00	3,00	3,80	9,30	5,20	53,00	4,80	101,00	4,10	19,68
128	3,70	0,70	8,20	6,00	2,10	2,50	5,20	41,00	6,40	40,00	1,21	7,74
124	2,60	4,80	8,20	5,00	3,60	2,50	9,00	53,00	6,60	46,00	1,95	12,87
198	4,50	4,10	6,30	5,90	4,30	3,40	8,80	50,00	6,40	85,00	2,33	14,91
310	2,80	2,40	6,70	4,90	2,50	2,60	9,20	86,00	3,80	108,00	4,55	17,29

Lampiran B

Tabel 4.3: *Score* komponen utama untuk setiap pengamatan

Pengamatan	W1	W2	W3
1	121,459	22,913	14,458
2	77,020	6,468	34,103
3	71,773	8,799	39,502
4	45,856	13,814	25,003
5	90,629	15,858	46,382
6	87,307	14,471	33,190
7	68,905	11,260	37,509
8	86,185	6,382	33,989
9	99,931	14,340	50,530
10	100,075	5,874	42,944
11	88,956	20,045	18,632
12	49,495	16,082	39,622
13	73,861	15,805	28,555
14	73,914	23,964	25,882
15	96,024	20,777	40,436
16	82,032	24,996	35,917
17	92,386	11,674	26,163
18	34,654	22,188	32,993
19	47,492	11,080	48,327
20	83,247	14,917	44,207
21	73,598	12,649	31,204
22	56,786	3,456	28,942
23	121,748	16,874	39,118
24	91,515	14,144	23,559
25	106,659	13,865	35,465
26	80,329	19,493	37,409
27	90,660	7,881	24,967
28	66,423	17,488	44,705
29	62,759	16,047	40,102
30	102,275	57,330	23,643
31	69,361	1,618	36,093
32	77,043	14,734	42,804
33	101,130	17,634	46,668
34	79,744	16,108	36,006

Pengamatan	W1	W2	W3
35	29,620	25,986	27,352
36	103,877	6,978	26,659
37	90,575	1,474	34,688
38	92,857	3,045	36,170
39	79,800	6,315	19,237
40	123,080	-0,021	14,547
41	76,852	7,192	30,876
42	99,050	3,332	48,027
43	91,486	17,967	40,461
44	111,824	7,892	35,072
45	102,533	-2,263	27,655
46	100,431	53,257	33,053
47	84,774	19,048	43,012
48	82,806	11,849	32,361
49	81,255	29,153	43,447
50	108,361	15,348	39,238
51	45,202	10,760	35,082
52	53,206	17,105	45,035
53	91,536	12,535	38,505
54	118,658	18,337	70,711

Lampiran C

Tabel 4.4: Jarak *Mahalanobis*

Pengamatan	*Mdi*	Pengamatan	Mdi
1	2,853313	28	1,264447
2	0,848861	29	1,085447
3	0,890131	30	4,274939
4	2,040753	31	1,405258
5	1,167281	32	0,814217
6	0,278948	33	1,427199
7	0,787635	34	0,231659
8	0,809424	35	2,830638
9	1,712038	36	1,479113
10	1,364767	37	1,293040
11	1,788303	38	1,184486
12	1,640686	39	1,829809
13	0,833580	40	3,116922
14	1,373810	41	0,903014
15	0,954762	42	1,819639
16	0,970934	43	0,698892
17	1,058159	44	1,450910
18	2,381501	45	1,985630
19	2,153861	46	3,719835
20	0,897067	47	0,876482
21	0,659654	48	0,412138
22	1,761000	49	1,590377
23	1,808489	50	1,205828
24	1,254084	51	1,815682
25	1,063580	52	1,729216
26	0,516634	53	0,525118
27	1,280223	54	3,948446

Lampiran D

```
function [rew,raw,hsetsfull] = DetMCD(x,varargin)

if rem(nargin-1,2)~=0
    error('The number of input arguments should be odd!');
end
% Assigning some input parameters
data = x;
rew.plane=[];
raw.cor=[];
rew.cor=[];
if size(data,1)==1
    data=data';
end

% Observations with missing or infinite values are ommitted.
ok=all(isfinite(data),2);
data=data(ok,:);
xx=data;
[n,p]=size(data);

% Some checks are now performed.
if n==0
    error('All observations have missing or infinite values.')
end
if n < p
    error('Need at least (number of variables) observations.')
end

%internal variables
hmin=quanf(0.5,n,p);
%Assiging default values
h=quanf(0.75,n,p);
default=struct('alpha',0.75,'h',h,'plots',1,'scale_est',1,'cor',0,'hsetsful
l',NaN,'classic',0);
list=fieldnames(default);
options=default;
IN=length(list);
i=1;
counter=1;

%Reading optional inputarguments
if nargin>2
    %
    % placing inputfields in array of strings
    %
    for j=1:nargin-1
        if rem(j,2)~=0
            chklist{i}=varargin{j};
            i=i+1;
        end
    end
    dummy=sum(strcmp(chklist,'h')+2*strcmp(chklist,'alpha'));
    switch dummy
        case 0 % default values should be taken
            alfa=options.alpha;
            h=options.h;
```

```
case 3
            error('Both input arguments alpha and h are provided. Only one
is required.')
    end
    %
    % Checking which default parameters have to be changed
    % and keep them in the structure 'options'.
    %
    while counter<=IN
        index=strmatch(list(counter,:),chklist,'exact');
        if ~isempty(index) % in case of similarity
            for j=1:nargin-2 % searching the index of the accompanying
field
                if rem(j,2)~=0 % fieldnames are placed on odd index
                    if strcmp(chklist{index},varargin{j})
                        I=j;
                    end
                end
            end
            options=setfield(options,chklist{index},varargin{I+1});
            index=[];
        end
        counter=counter+1;
    end
    if dummy==1% checking inputvariable h
        % hmin is the minimum number of observations whose covariance
determinant
        % will be minimized.

        if options.h < hmin
            disp(['Warning: The MCD must cover at least ' int2str(hmin) '
observations.'])
            disp(['The value of h is set equal to ' int2str(hmin)])
            options.h = hmin;
        elseif options.h > n
            error('h is greater than the number of non-missings and non-
infinites.')
        elseif options.h < p
            error(['h should be larger than the dimension ' int2str(p)
'.'])
        end

        options.alpha=options.h/n;
    elseif dummy==2
        if options.alpha < 0.5
            options.alpha=0.5;
            mess=sprintf(['Attention (detmcd.m): Alpha should be larger
than 0.5. \n',...
                'It is set to 0.5.']);
            disp(mess)
        end
        if options.alpha > 1
            options.alpha=0.75;
            mess=sprintf(['Attention (detmcd.m): Alpha should be smaller
than 1.\n',...
                'It is set to 0.75.']);
            disp(mess)
        end
```

```
options.h=quanf(options.alpha,n,p);
    end
end

h=options.h;  %number of regular data points on which estimates are based.
h=[alpha*n]
plots=options.plots; %relevant plots are plotted
alfa=options.alpha; %percentage of regular observations
scale_est=options.scale_est;
hsetsfull=options.hsetsfull;
cor=options.cor;

%-------------------------------------------------------------------------
%MAIN part


switch scale_est
    case 1
        if n>=1000
            scales='W_scale';
        else
            scales='qn';
        end
    case 2
        scales='qn';
    case 3
        scales='W_scale';
end

med=median(data);
sca=feval(scales,data);
ii=find((sca < eps),1);
if ~isempty(ii)
    error(['DetMCD.m: Variable ', int2str(ii), ' has zero scale. MCD can
not be computed.']);
end
data=(data-repmat(med,n,1))./repmat(sca,n,1);
cutoff.rd=sqrt(chi2inv(0.975,p)); %cutoff value for the robust distance
cutoff.md=cutoff.rd; %cutoff value for the Mahalanobis distance
clmean=mean(data);
clcov=cov(data);

if p==1

[rew.center,rewsca,weights,raw.center,raw.cov,raw.rd,Hopt]=unimcd(data,h);
    rew.Hsubsets.Hopt = Hopt';
    raw.cov=raw.cov*sca^2;
    raw.objective=raw.cov;
    raw.center=raw.center*sca+med;
    raw.cutoff=cutoff.rd;
    raw.wt=weights;
    rew.cov=rewsca^2;
    mah=(data-rew.center).^2/rew.cov;
    rew.rd=sqrt(mah');
    rew.flag=(rew.rd<=cutoff.rd);
    rew.cutoff=cutoff.rd;
```

```
rew.center=rew.center*sca+med;
    rew.cov=rew.cov*sca^2;
    rew.mahalanobis=abs(data'-clmean)/sqrt(clcov);

    %classical analysis?
    if options.classic==1
        classic.cov=clcov;
        classic.center=clmean;
        classic.md=rew.mahalanobis;
        classic.flag = (classic.md <= cutoff.md);
        classic.class='COV';
    else
        classic=0;
    end
    %assigning the output
    rewo=rew;rawo=raw;

rew=struct('center',{rewo.center},'cov',{rewo.cov},'cor',{rewo.cor},'h',{h}
,'Hsubsets',{rewo.Hsubsets},...
        'alpha',{alfa},'rd',{rewo.rd},'cutoff',{cutoff},'flag',{rewo.flag},
'plane',{rewo.plane},...
        'class',{'MCDCOV'},'md',{rewo.mahalanobis},'classic',{classic});

raw=struct('center',{rawo.center},'cov',{rawo.cov},'cor',{rawo.cor},'object
ive',{rawo.objective},...
        'rd',{rawo.rd},'wt',{rawo.wt});
    if plots
        makeplot(rew);
    end
    return
end

if isnan(hsetsfull)
    hsetsfull=NaN(6,n);
    %Determining initial shape estimates

    %1) Hyperbolic tangent of standardized data
    y1=tanh(data);
    R1=corr(y1);
    [P,L]=eig(R1);
    ind=initset(data,scales,P,n,p);
    hsetsfull(1,:)=ind;

    %2) Spearmann correlation matrix
    y2=data;
    for j=1:p
        y2(:,j)=tiedrank(data(:,j));
    end
    R2=corr(y2);
    [P,L]=eig(R2);
    ind=initset(data,scales,P,n,p);
    hsetsfull(2,:)=ind;

    %3) Tukey normal scores
    y3=norminv((y2-1/3)/(n+1/3));
    R3=corr(y3);
```

```
[P,L]=eig(R3);
    ind=initset(data,scales,P,n,p);
    hsetsfull(3,:)=ind;

    %4) Spatial sign covariance matrix
    znorm=sqrt(sum(data.^2,2));
    ii=znorm>eps;
    zznorm=data;
    zznorm(ii,:)=data(ii,:)./repmat(znorm(ii),1,p);
    SCM=(zznorm'*zznorm)./(n-1);
    [P,L]=eig(SCM);
    ind=initset(data,scales,P,n,p);
    hsetsfull(4,:)=ind;

    %5) BACON
    [not,ind5]=sort(znorm);
    half=ceil(n/2);
    Hinit=ind5(1:half);
    covx=cov(data(Hinit,:));
    [P,L]=eig(covx);
    ind=initset(data,scales,P,n,p);
    hsetsfull(5,:)=ind;

    %6) Raw OGK estimate for scatter
    P=ogkscatter(data,scales);
    ind=initset(data,scales,P,n,p);
    hsetsfull(6,:)=ind;

Isets=hsetsfull(:,1:half);
nIsets=size(Isets,1);

for k=1:nIsets
    xk=data(Isets(k,:),:);
    [P,T,L,r,centerX,meanvct] = classSVD(xk);
    if r < p
        error('DetMCD.m: More than half of the observations lie on a
hyperplane.')
    end
    score=(data - repmat(meanvct,n,1))*P;
    [dis,sortdist]=sort(mahalanobis(score,zeros(size(score,2),1),'cov',L));
    hsetsfull(k,:)=sortdist;
end

end

% construction of h-subsets
nIsets=size(hsetsfull,1);
Hsets=hsetsfull(:,1:h);
%-------------------------------------------------------------------------
%Applying C-steps as in mcdcov.m

% Some initializations.
raw.wt=NaN(1,length(ok));
raw.rd=NaN(1,length(ok));
rew.rd=NaN(1,length(ok));
```

```
rew.mahalanobis=NaN(1,length(ok));
rew.flag=NaN(1,length(ok));

%nsamp=size(Hsets,1);
csteps=100;
prevdet=0;
bestobj=inf;
cutoff.rd=sqrt(chi2inv(0.975,p)); %cutoff value for the robust distance
cutoff.md=cutoff.rd; %cutoff value for the Mahalanobis distance

for i=1:nIsets
    for j=1:csteps
        if j==1
            obs_in_set=Hsets(i,:);
        else
            score=(data - repmat(meanvct,n,1))*P;
            mah=mahalanobis(score,zeros(size(score,2),1),'cov',L);
            [dis2,sortdist]=sort(mah);
            obs_in_set=sortdist(1:h);
        end
        [P,T,L,r,centerX,meanvct] = classSVD(data(obs_in_set,:));
        obj=prod(L);

        if r < p
            error('DetMCD.m: More than h of the observations lie on a
hyperplane.');
        end
        if j >= 2 && obj == prevdet
            break;
        end
        prevdet=obj;

    end
    if obj < bestobj
        % bestset           : the best subset for the whole data.
        % bestobj           : objective value for this set.
        % initmean, initcov : resp. the mean and covariance matrix
        % of this set.
        bestset=obs_in_set;
        bestobj=obj;
        initmean=meanvct;
        initcov=P*diag(L)*P';
        raw.initcov=initcov;
        rew.Hsubsets.Hopt=bestset;
        rew.Hsubsets.i=i; %to determine which subset gives best results.
    end
    rew.Hsubsets.csteps(i)=j; %how many csteps necessary to converge.
end

[P,T,L,r,centerX,meanvct] = classSVD(data(bestset,:));
mah=mahalanobis((data - repmat(meanvct,n,1))*P,zeros(size(P,2),1),'cov',L);
sortmah=sort(mah);

factor = sortmah(h)/chi2inv(h/n,p);
raw.cov=factor*initcov;
% We express the results in the original units.
```

```
raw.cov=raw.cov.*repmat(sca,p,1).*repmat(sca',1,p);
raw.center=initmean.*sca+med;
raw.objective=bestobj*prod(sca)^2;
mah=mah/factor;
raw.rd=sqrt(mah);
weights=raw.rd<=cutoff.rd;
raw.wt=weights;
[rew.center,rew.cov]=weightmecov(data,weights);
trcov=rew.cov.*repmat(sca,p,1).*repmat(sca',1,p);
trcenter=rew.center.*sca+med;

mah=mahalanobis(data,rew.center,'cov',rew.cov);
rew.rd=sqrt(mah);
rew.flag=(rew.rd <= cutoff.rd);

rew.mahalanobis=sqrt(mahalanobis(data,clmean,'cov',clcov));
rawo=raw;
reso=rew;

if options.classic==1
    classic.center=clmean.*sca+med;
    classic.cov=clcov.*repmat(sca,p,1).*repmat(sca',1,p);
    classic.md=rew.mahalanobis;
    classic.flag = (classic.md <= cutoff.md);
    if cor==1
        diagcl=sqrt(diag(clcov));
        classic.cor=clcov./(diagcl*diagcl');
    end
    classic.class='COV';

else
    classic=0;
end

if cor==1
    diagraw=sqrt(diag(raw.cov));
    raw.cor=raw.cov./(diagraw*diagraw');
    diagrew=sqrt(diag(rew.cov));
    rew.cor=rew.cov./(diagrew*diagrew');
end

rew=struct('center',{trcenter},'cov',{trcov},'cor',{rew.cor},'h',{h},'Hsubs
ets',{reso.Hsubsets},'alpha',{alfa},...

'rd',{reso.rd},'cutoff',{cutoff},'flag',{reso.flag},'plane',{reso.plane},..
.

'class',{'MCDCOV'},'md',{reso.mahalanobis},'classic',{classic},'X',{xx});
raw=struct('center',{rawo.center},'cov',{rawo.cov},'cor',{raw.cor},'objecti
ve',{rawo.objective},...
    'rd',{rawo.rd},'cutoff',{cutoff},'wt',{rawo.wt});

if size(data,2)~=2
    rew=rmfield(rew,'X');
end

if plots
```

```
makeplot(rew);
end
%-------------------------------------------------------------------------
%-------------------------------------------------------------------------
%Some auxiliary functions:
%-------------------------------------------------------------------------
function quan=quanf(alfa,n,rk)
quan=floor(2*floor((n+rk+1)/2)-n+2*(n-floor((n+rk+1)/2))*alfa);
%-------------------------------------------------------------------------
function [scale]=W_scale(x)
c=4.5;
[n,p]=size(x);
Wc=inline('(1-(x./c).^2).^2.*(abs(x)<c)');
sigma0=mad(x,1);
w=Wc(c,(x-repmat(median(x),n,1))./repmat(sigma0,n,1));
loc=diag(x'*w)'./sum(w);

c=3;
rc=inline('min(x.^2,c^2)');
sigma0=mad(x,1);
b=c*norminv(3/4);
nes=n*(2*((1-b^2)*normcdf(b)-b*normpdf(b)+b^2)-1);
scale=sigma0.^2./nes.*sum(rc(c,(x-repmat(loc,n,1))./repmat(sigma0,n,1)));
scale=sqrt(scale);
%-------------------------------------------------------------------------
function [P,L]=ogkscatter(x,scales)

[n,p]=size(x);
U=eye(p);
for i=1:p
    sYi=x(:,i);
    for j=1:(i-1)
        sYj=x(:,j);
        sY=sYi+sYj;
        dY=sYi-sYj;
        U(i,j)= 0.25*(feval(scales,sY)^2-feval(scales,dY)^2);
    end
end
U=tril(U,-1)+U';
[P,L]=eig(U);

%-------------------------------------------------------------------------
function [ind]=initset(data,scales,P,n,p)

lambda=feval(scales,data*P);
sqrtcov=P*diag(lambda)*P';
sqrtinvcov=P*diag(1./lambda)*P';
estloc=median(data*sqrtinvcov)*sqrtcov;
centeredx=(data-repmat(estloc,n,1))*P;
[not,ind]=sort(mahalanobis(centeredx,zeros(p,1),'cov',diag(lambda).^2));
```

Lampiran E

Tabel 4.5: Jarak *Robust*

Pengamatan	*RDi*	Pengamatan	*RDi*
1	4,084021	28	1,410912
2	1,011843	29	1,227069
3	1,090890	30	7,742372
4	2,534878	31	1,960046
5	1,413292	32	0,910836
6	0,545576	33	1,804967
7	0,877645	34	0,571923
8	0,897073	35	3,763253
9	2,095328	36	1,510293
10	1,761085	37	1,671041
11	2,738411	38	1,507579
12	1,865579	39	2,050252
13	1,260169	40	3,283107
14	2,464236	41	0,993412
15	1,538916	42	2,503933
16	1,940688	43	1,069619
17	1,221934	44	1,596700
18	2,975471	45	2,300710
19	2,530810	46	6,644685
20	1,045535	47	1,197619
21	0,802486	48	0,396941
22	2,194804	49	2,503058
23	2,306724	50	1,536776
24	1,653700	51	2,142874
25	1,305787	52	1,911106
26	1,050321	53	0,584236
27	1,333786	54	4,828416

Lampiran F

Tabel 4.6: Pendeteksian *Outlier* untuk setiap pengamatan jarak *robust*

Pengamatan	Pendeteksian *Outlier*	Pengamatan	Pendeteksian *Outlier*
1	1	28	0
2	0	29	0
3	0	30	1
4	0	31	0
5	0	32	0
6	0	33	0
7	0	34	0
8	0	35	1
9	0	36	0
10	0	37	0
11	0	38	0
12	0	39	0
13	0	40	1
14	0	41	0
15	0	42	0
16	0	43	0
17	0	44	0
18	0	45	0
19	0	46	1
20	0	47	0
21	0	48	0
22	0	49	0
23	0	50	0
24	0	51	0
25	0	52	0
26	0	53	0
27	0	54	1

SURAT PERNYATAAN KEASLIAN SKRIPSI

Dengan ini saya yang bertanda tangan di bawah ini, mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta:

Nama	:	Yulifirda Isnaini
No. Registrasi	:	3125121977
Program Studi	:	Matematika

Menyatakan bahwa skripsi ini yang saya buat dengan judul "**Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Metode *Robust* PCA pada Model Regresi Linear Berganda**" adalah :

1. Dibuat dan diselesaikan oleh saya sendiri.
2. Bukan merupakan duplikat skripsi yang pernah dibuat oleh orang lain atau jiplakan karya tulis orang lain.

Pernyataan ini dibuat dengan sesungguhnya dan saya bersedia menanggung segala akibat yang timbul jika pernyataan saya tidak benar.

Jakarta, Agustus 2017

Yang membuat pernyataan

Yulifirda Isnaini

DAFTAR RIWAYAT HIDUP

YULIFIRDA ISNAINI. Lahir di Jakarta, 4 Juli 1994. Anak kedua dari pasangan Bapak Makmun dan Ibu Nurchayati. Saat ini bertempat tinggal di Jalan Kramat IV RT 003/ RW 03 No 76, Jakarta Timur 17810.

No. Ponsel : 085691929899

Email : yulifirda.isnaini@gmail.com

Riwayat Pendidikan: Penulis mengawali pendidikan di TK Fathul Jabbar selama 1 tahun, dan kemudian melanjutkan pendidikan di SD Negeri 04 PG Lubang Buaya pada tahun 2000 - 2006. Setelah itu, penulis melanjutkan ke SMP Negeri 81 Jakarta pada tahun 2006 - 2009. Kemudian kembali melanjutkan ke SMA Negeri 113 Jakarta pada tahun 2009 - 2012. Di Tahun yang sama penulis melanjutkan ke Universitas Negeri Jakarta (UNJ), jurusan Matematika, melalui jalur SNMPTN. Di pertengahan tahun 2017 penulis telah memperoleh gelar Sarjana Sains untuk Jurusan Matematika, Program Studi Matematika, FMIPA, UNJ.

Riwayat Organisasi: Selama di bangku perkuliahan, penulis aktif di berbagai kegiatan yang diselenggarakan oleh BEM Matematika. Dalam tahun kedua dan ketiga, penulis menjadi panitia divisi Infokom Pelangi XX dan Pelangi XXI.

Riwayat Pekerjaan: Penulis pernah melakukan praktik kerja lapangan di Badan Kepegawaian Negara divisi keuangan pada tahun 2015.