Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)
Volume 01, No. 1 (2012), hal 31 – 40.

MENDETEKSI *OUTLIER* DENGAN METODE *MINIMUM COVARIANCE DETERMINANT*

Mukti Kurniadi, Marisi Aritonang, Muhlasah Novitasari Mara

INTISARI

Keberadaan outlier pada data dapat mengganggu proses analisis data, sehingga pendeteksian outlier merupakan hal yang sangat penting untuk dilakukan. Ketika data yang digunakan adalah data multivariat, maka pendeteksian tersebut menjadi sulit untuk dilakukan. Pada umumnya, sebelum melakukan pendeteksian outlier pada data multivariat, terlebih dahulu perlu dilakukan peyederhanaan atau pereduksian dimensi data. Salah satu metode yang dapat digunakan adalah Analisis Komponen Utama (AKU). Dalam penelitian ini, outlier dideteksi menggunakan metode Minimum Covariance Determinant (MCD). Prinsip metode MCD adalah mendapatkan subhimpunan dari keseluruhan pengamatan yang matriks varians-kovariansnya memiliki determinan terkecil diantara semua kombinasi kemungkinan data. Pendeteksian outlier dengan metode MCD dilakukan berdasarkan jarak robust dan nilai cut-offnya. Suatu pengamatan terdeteksi sebagai outlier ketika jarak robust lebih besar dari nilai cut-off. Sedangkan untuk mengklasifikasikan outlier tersebut dilakukan dengan cara membuat plot jarak mahalanobis versus jarak robust yang disebut dengan diagnostic plot. Pada penelitian ini dilakukan deteksi outlier untuk data tahun 1994 tentang gaji pegawai pada perguruan tinggi di Amerika. Berdasarkan analisis yang dilakukan dengan bantuan software R versi 2.13.2, dapat disimpulkan bahwa sebanyak 309 pengamatan terdeteksi sebagai outlier, yang terdiri dari 48 pengamatan termasuk jenis bad leverage dan 261 pengamatan termasuk jenis outlier orthogonal. Bagi peneliti lain yang ingin meneliti tentang pendeteksian outlier pada kasus multivariat dapat menggunakan metode-metode lain seperti metode Minimum Volume Ellipsoid (MVE) dan metode Welsch untuk kemudian dibandingkan tingkat efisiensinya dengan metode MCD.

Kata Kunci : Analisis Komponen Utama, jarak *robust*, *diagnostic plot*, nilai *cut-off*, *outlier multivariate*.

PENDAHULUAN

Salah satu metode statistik yang dapat menyederhanakan atau mereduksi dimensi data tanpa mengabaikan variabel-variabel asli adalah Analisis Komponen Utama (AKU). AKU pertama kali diperkenalkan oleh Harold Hottelling pada tahun 1933. AKU merupakan salah satu metode analisis multivariat yang sudah cukup dikenal. Metode ini mampu mereduksi dimensi data yang besar dan saling berkorelasi menjadi dimensi data yang lebih kecil dan tidak saling berkorelasi, tanpa kehilangan banyak informasi. Perhitungan dalam AKU didasarkan pada matriks varians-kovarians sampel (**S**). Matriks varians-kovarians ini sangat sensitif terhadap keberadaan *outlier*. *Outlier* yaitu pengamatan yang tidak mengikuti sebagian besar pola dan terletak jauh dari pusat data [1]. Keberadaan *outlier* pada suatu data dapat mengganggu proses analisis data, sehingga mengakibatkan varians pada data tersebut menjadi besar dan dugaan interval memiliki rentang yang [2]. Akan tetapi membuang begitu saja suatu pengamatan *outlier* bukanlah tindakan yang bijaksana, karena adakalanya pengamatan *outlier* memberikan informasi yang cukup berarti. Seriusnya permasalahan dan efek yang ditimbulkan *outlier*, maka pendeteksian *outlier* menjadi sangat penting untuk dilakukan. Pada penelitian ini untuk mendeteksi *outlier* tersebut matriks varians-kovarians diduga dengan matriks varians-kovarians yang *robust* menggunakan metode*Minimum Covariance Determinant*(MCD). Dalam artikel ini MCD diterapkan pada suatu data yang dianalisis dengan *software* R versi 2.13.2.

Penelitian tentang penanganan data *outlier* dengan mengkaji konsep ROBPCA dan membandingkan hasilnya dengan analisis komponen utama klasik (CPCA) dilakukan pada tahun

2007 oleh Suryana. Hasil penelitiannya adalah metode ROBPCA menghasilkan jumlah komponen utama yang lebih sedikit daripada CPCA untuk mereduksi data. Dengan varians yang dapat dijelaskan 80%, ROBPCAmembutuhkan tiga komponen utama sedangkan CPCA membutuhkan 11 komponen utama [3]. Kemudian pada tahun 2009, Cahyawati dkk juga melakukan penelitian untuk membandingkan efektivitas metode regresi *robust Welsch* (MRR-W) dengan metode kuadrat terkecil (MKT) dalam melakukan pendugaan parameter model regresi. Hasil penelitian mereka menunjukkan bahwa untuk berbagai ukuran sampel yang diamati, pendugaan parameter MRR-W menghasilkan model yang lebih baik dari MKT [4]. Selanjutnya penelitian untuk membandingkan efektivitas metode regresi *robust Welsch* (MRR-W) dengan metode kuadrat terkecil (MKT) dalam melakukan pendugaan parameter model regresi dilakukan pada tahun 2010 oleh Makkulau dkk. Hasil penelitian mereka adalah metode LDL dapat mendeteksi adanya *outlier* pada pengamatan produksi gula dan tetes tebu pada PGDB Jombang. Berdasarkan 122 pengamatan yang berhasil dikumpulkan dapat diidentifikasi *outlier* pada pengamatan ke-3, ke-71, dan ke-116[5]. Artikel ini mengkaji tentang pendeteksian *outlier* dengan metode MCD dan menerapkannya pada data gaji pegawai perguruan tinggi di Amerika. Dalam penggunaannya, metode MCD ini diasumsikan bahwa matriks varians-kovarians dari MCD tidak nol.

METODE *MINIMUM COVARIANCE DETERMINANT* (MCD)

Metode MCD digunakan untuk mendeteksi *outlier* pada kasus multivariat. Untuk melakukan pendeteksian tersebut, ada beberapa tahapan yang harus dilakukan terlebih dahulu, yaitu melakukan Analisis Komponen Utama (AKU), menghitung jarak Mahalanobis, menghitung jarak *Robust*, dan membuat *Diagnostic Plot*. Analisis komponen utama merupakan suatu prosedur alternatif yang pertama kali diperkenalkan oleh Harold Hottelling. Metode ini bertujuan untuk mereduksi dimensi data dan mencari variabel baru yang saling bebas. Selanjutnya variabel baru ini dinamakan *score* komponen utama. Komponen utama dapat ditentukan dengan matriks varians-kovarians sampel [6], yang dihitung dengan rumus berikut:

$$s_{pj} = s_{jp} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ip} - \bar{x}_p\right)\left(x_{ij} - \bar{x}_j\right)$$

dengan s_{pj}merupakan entri pada baris ke-p dan kolom ke-j dari matriks **S**, x_{ip}merupakan entri pada baris ke-i dan kolom ke-p dari matriks **X** dengan $i = 1, 2, \ldots,$ n dan $p = 1, 2, \ldots, m$, $\bar{x}_p$ merupakan rata-rata pada kolom ke-p dari matriks **X** dengan $p = 1, 2, \ldots, m$, x_{ij}merupakan entri pada baris ke-i dan kolom ke-j dari matriks **X** dengan $i = 1, 2, \ldots,$ n dan $j = 1, 2, \ldots, m$, dan $\bar{x}_j$ merupakan rata-rata pada kolom ke-j dari matriks **X** dengan $j = 1, 2, \ldots, m$. Sehingga dari hasil perhitungan dapat dibentuk matriks varians-kovarians sampel sebagai berikut:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix}$$

Sasaran analisis komponen utama adalah menemukan komponen utama K dimana banyaknya $K<m$, K diharapkan dapat memuat semua informasi yang terdapat pada m variabel asli sehingga data menjadi lebih sederhana. Dalam bentuk kombinasi linear, komponen utama dinyatakan sebagai berikut [6]:

$$\begin{pmatrix} k_{11} \\ k_{21} \\ \vdots \\ k_{n1} \end{pmatrix} = a_{11}\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{1n} \end{pmatrix} + a_{21}\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} + \cdots + a_{m1}\begin{pmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{pmatrix}$$

$$\begin{pmatrix} k_{12} \\ k_{22} \\ \vdots \\ k_{n2} \end{pmatrix} = a_{11}\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{1n} \end{pmatrix} + a_{21}\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} + \cdots + a_{m1}\begin{pmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{pmatrix}$$

$$\vdots$$

$$\begin{pmatrix} k_{1m} \\ k_{2m} \\ \vdots \\ k_{nm} \end{pmatrix} = a_{11}\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{1n} \end{pmatrix} + a_{21}\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} + \cdots + a_{m1}\begin{pmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{pmatrix}$$

Persamaan diatas juga dapat dinyatakan sebagai berikut:

$$\begin{aligned} \mathbf{K}_1 &= a_{11}\mathbf{X}_1 + a_{21}\mathbf{X}_2 + \ldots + a_{m1}\mathbf{X}_m \\ \mathbf{K}_2 &= a_{12}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \ldots + a_{m2}\mathbf{X}_m \\ &\vdots \\ \mathbf{K}_m &= a_{1m}\mathbf{X}_1 + a_{2m}\mathbf{X}_2 + \ldots + a_{mm}\mathbf{X}_m \end{aligned} \tag{1}$$

atau dalam bentuk matriks dinyatakan dengan:

$$\mathbf{K} = \mathbf{a}_j^T\mathbf{X}$$

$$\begin{pmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \\ \vdots \\ \mathbf{K}_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{mm} \end{pmatrix}\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}$$

dengan **K** merupakanvektor komponen utama, $\mathbf{a}_j$ merupakan vektor eigen dari matriks varians-kovarians sampel, a_j^T merupakan *transpose* dari a_j dengan j=1,2, …,m, $\mathbf{X}_j$ merupakan variabel asli dengan j = 1, 2, …, m.

Langkah yang dilakukan dalam analisis komponen utama adalah menentukan *score* komponen utama. Cara memperoleh *score* komponen utama adalah menyelesaikan Persamaan (1) dengan mensubstitusikan nilai-nilai pada setiap variabel asli. Sehingga diperoleh matriks *score* komponen utama yang dinotasikan dengan **M**, yaitu:

$$\mathbf{M} = \begin{pmatrix} \mathbf{K}_1^T \\ \mathbf{K}_2^T \\ \vdots \\ \mathbf{K}_m^T \end{pmatrix} = \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{n1} \\ k_{12} & k_{22} & \cdots & k_{n2} \\ \vdots & \vdots & & \vdots \\ k_{1m} & k_{2m} & \cdots & k_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 & \cdots & \mathbf{L}_n \end{pmatrix}$$

Setelah melakukan AKU, langkah selanjutnya adalah menghitung jarak Mahalanobis. Jarak mahalanobis merupakan jarak antara masing-masing vektor data dengan titik pusat data atau vektor rata-rata. Jarak mahalanobis didefinisikan pada Persamaan berikut [8]:

$$MD_i = \sqrt{\left(\mathbf{L}_i - \overline{\mathbf{M}}\right)^T \mathbf{S}^{-1}\left(\mathbf{L}_i - \overline{\mathbf{M}}\right)},\ i = 1, 2, ..., n$$

dengan MD_i merupakan jarak mahalanobis pada pengamatan ke-i, $\mathbf{L}_i$ merupakan vektor komponen utama pada pengamatan ke-i, $\overline{\mathbf{M}}$ merupakan vektor rata-rata dari $\mathbf{K}_j$,$\mathbf{S}$ merupakan matriks varians-kovarians sampel, $\mathbf{S}^{-1}$ merupakan invers dari matriks $\mathbf{S}$ dengan, $\mathbf{L}_i = \begin{pmatrix} k_{i1} & k_{i2} & \cdots & k_{im} \end{pmatrix}$, $\overline{\mathbf{M}} = \begin{pmatrix} \overline{\mathbf{K}}_1 & \overline{\mathbf{K}}_2 & \cdots & \overline{\mathbf{K}}_m \end{pmatrix}$, dan $\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}$.

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

$$h = \frac{n + k + 1}{2}$$

dan k menyatakan banyak variabel. Jika nilai h merupakan pecahan maka nilai h dibulatkan ke bawah [9]. Misalkan terdapat sampel acak, yaitu $\mathbf{K}_1$, $\mathbf{K}_2$, …,$\mathbf{K}_m$ diambil dari distribusi yang mempunyai vektor rata-rata $\overline{\mathbf{M}}$ dan matriks varians-kovarians $\mathbf{S}$, dengan $\mathbf{M} = \begin{pmatrix} \mathbf{K}_1 & \mathbf{K}_2 & \cdots & \mathbf{K}_m \end{pmatrix}$, $\mathbf{K}_1 = \begin{pmatrix} k_{11} \\ k_{21} \\ \vdots \\ k_{n1} \end{pmatrix}$, $\mathbf{K}_2 = \begin{pmatrix} k_{12} \\ k_{22} \\ \vdots \\ k_{n2} \end{pmatrix}$ sampai $\mathbf{K}_m = \begin{pmatrix} k_{1m} \\ k_{2m} \\ \vdots \\ k_{nm} \end{pmatrix}$. Penduga MCD untuk $\overline{\mathbf{M}}$ dan $\mathbf{S}$ masing-masing adalah $\overline{\mathbf{M}}_{MCD}$ dan $\mathbf{S}_{MCD}$ dengan

$$\overline{\mathbf{M}}_{MCD} = \frac{1}{h}\sum_{i=1}^{h} \mathbf{K}_i$$

$$\mathbf{S}_{MCD} = \frac{1}{h-1}\sum_{i=1}^{h} \left(\mathbf{K}_i - \overline{\mathbf{M}}_{MCD}\right)\left(\mathbf{K}_i - \overline{\mathbf{M}}_{MCD}\right)^T$$

dan determinan matriks varians-kovarians $\mathbf{S}_{MCD}$ minimum diantara semua kemungkinan h. Jika n kecil ($n \leq 600$) maka pendugaan MCD mudah dan relatif lebih cepat untuk ditemukan. Sedangkan jika n besar ($n > 600$)maka banyak sekali kombinasi subhimpunan yang harus ditemukan untuk mendapatkan pendugaan MCD. Keterbatasan ini kemudian diatasi oleh Rousseeuw and Van Driessen tahun 1998 dengan algoritma yang dikenal dengan istilah FAST-MCD. Berdasarkan Rousseeuw and Van Driessen tahun 1998, algoritmanya adalah sebagai berikut:

1. Ambil himpunan bagian dari matriks $\mathbf{M}$ yang terdiri atas $h = \frac{n + k + 1}{2}$ buah data dan disimbolkan dengan H_{lama}.
2. Hitung vektor rata-rata $\overline{\mathbf{M}}_{i\,lama}$ dan matriks varians-kovarians $\mathbf{S}_{lama}$.

3. Kemudian hitung jarak mahalanobis $MD_{lama} = \sqrt{\left(\mathbf{L}_i - \overline{\mathbf{M}}\right)^T \mathbf{S}^{-1}\left(\mathbf{L}_i - \overline{\mathbf{M}}\right)}$ dengan $\mathbf{L}_i$ merupakan vektor komponen utama pada pengamatan ke-*i*.
4. Urutkan $\mathbf{L}_i$ berdasarkan nilai MD_{lama} dari yang terkecil ke nilai yang terbesar.
5. Definisikan himpunan bagian baru yang dinotasikan dengan $H_{baru} = \{\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_h\}$.
6. Hitung $\overline{\mathbf{M}}_{i\,baru}$, $\mathbf{S}_{baru}$, dan MD_{baru}.
7. Ulangi Langkah 1 sampai langkah 6 sampai ditemukan $\det(\mathbf{S}_{baru}) \le \det(\mathbf{S}_{lama})$.

Karena $\overline{\mathbf{M}}_{MCD}$ dan $\mathbf{S}_{MCD}$ merupakan penduga untuk MCD maka $\overline{\mathbf{M}}_{MCD} = \overline{\mathbf{M}}_{baru}$ dan $\mathbf{S}_{MCD} = \mathbf{S}_{baru}$. Metode MCD mempunyai kemampuan mengukur jarak *robust* yang dapat digunakan untuk mendeteksi *outlier leverage*.

Jarak robust merupakan suatu pendekatan untuk mendeteksi *outlier* pada data multivariat, yaitu dengan menggunakan penduga dari $\overline{\mathbf{M}}_{MCD}$ dan $\mathbf{S}_{MCD}$ pada metode *robust*. Sehingga metode ini mampu meminimumkan pengaruh dari adanya efek *masking* dan *swamping* dalam pendeteksian *outlier* [10]. Terdapat beberapa penyebab munculnya *outlier*, salah satunya *outlier* yang disebabkan oleh variabel independen, dinamakan *outlierleverage*. *Outlier leverage* dideteksi dengan menggunakan jarak *robust* (RD_i) untuk setiap pengamatan ke-*i*. Jarak *robust* didefinisikan pada Persamaan berikut:

$$RD_i = \sqrt{\left(\mathbf{L}_i - \overline{\mathbf{M}}_{MCD}\right)^T \mathbf{S}_{MCD}^{-1}\left(\mathbf{L}_i - \overline{\mathbf{M}}_{MCD}\right)},\ i = 1, 2, ..., n.$$

dengan RD_i merupakan jarak *robust* untuk setiap pengamatan ke-*i*, $\mathbf{L}_i$ merupakan vektor komponen utama pada pengamatan ke-*i*, $\overline{\mathbf{M}}_{MCD}$ merupakan vektor rata-rata dari $\mathbf{K}_j$ dengan metode MCD, $\mathbf{S}_{MCD}$ merupakan matriks varians-kovarians sampel dengan metode MCD, $\mathbf{S}_{MCD}^{-1}$ merupakan invers dari matriks $\mathbf{S}$ dengan, $\mathbf{L}_i = \begin{pmatrix} k_{i1} & k_{i2} & \cdots & k_{im} \end{pmatrix}, \overline{\mathbf{M}}_{MCD} = \begin{pmatrix} \overline{\mathbf{K}}_1 & \overline{\mathbf{K}}_2 & \cdots & \overline{\mathbf{K}}_m \end{pmatrix}$, dan

$$\mathbf{S}_{MCD} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}.$$

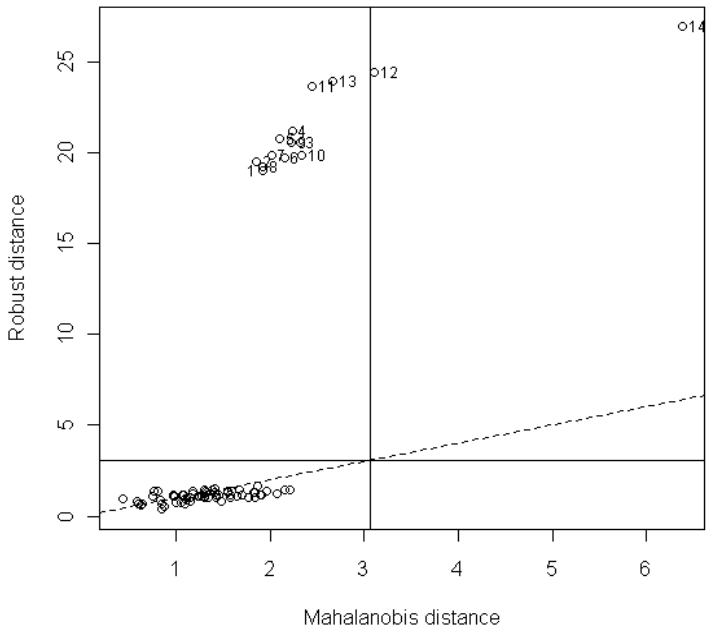
Pendeteksian *outlierleverage* menggunakan jarak *robust* (RD_i) untuk setiap pengamatan ke-*i* dapat dituliskan sebagai berikut (Chen, 2002):

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

dengan $C = \sqrt{\chi^2_{p;\alpha}}$, C dinyatakan sebagai nilai *cut-off*. Dalam hal ini, nilai *cut-off* merupakan suatu nilai yang digunakan untuk menentukan apakah suatu pengamatan dideteksi sebagai *outlier* atau bukan. Notasi $\chi^2_{p;\alpha}$ merupakan nilai χ^2 yang membuat luas di ujung kanan distribusinya sama dengan α dan *RDi* merupakan jarak *robust* untuk setiap pengamatan ke-*i*.

Outlier dapat dideteksi dan diklasifikasikan berdasarkan *diagnostic plot* atau peta *outlier* yang berguna untuk membedakan data pengamatan. Di dalam *diagnostic plot*, data pengamatan dibedakan menjadi empat tipe yaitu *bad leverage*, *outlier* orthogonal, pengamatan biasa dan *good leverage*[2]. *Bad leverage* (terletak pada kuadran 1), yaitu suatu titik yang memiliki nilai jarak *robust* dan nilai jarak mahalanobis yang lebih besar dari nilai *cut-off*. Titik ini merupakan jenis *outlier* yang sangat berpengaruh, akan tetapi tidak cocok untuk model regresi. Keberadaan titik *bad leverage* ini dapat merubah garis regresi sehingga dapat mempengaruhi hasil secara keseluruhan. Karena akibat yang

ditimbulkan oleh titik *bad leverage* ini, maka tindakan yang sebaiknya dilakukan adalah menghapus pengamatan yang tergolong jenis *bad leverage* tersebut. *Outlier* ortogonal (terletak pada kuadran 2), yaitu suatu titik yang memiliki nilai jarak *robust* lebih besar dari nilai *cut-off* dan nilai jarak mahalanobis lebih kecil atau sama dengan nilai *cut-off*. Titik ini berpengaruh pada pendugaan parameter regresi. Alternatif yang dapat dilakukan jika terdapat pengamatan jenis *outlier* ortogonal adalah tetap mengikutsertakan pengamatan tersebut tetapi pendugaan parameternya menggunakan metode *Least Median Square* (LMS). Pengamatan biasa (terletak pada kuadran 3), yaitu suatu 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* (terletak pada kuadran 4),yaitu suatu 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 terletak jauh dari pengamatan biasa, tetapi tetap mengikuti garis regresi. Pengamatan jenis *good leverage* ini tetap dapat diikutsertakan dalam analisis data. Berikut merupakan contoh gambar *diagnostic plot*:



Gambar 1. Contoh *diagnostic plot* jarak mahalanobis vs jarak *robust*

PENERAPAN METODE *MINIMUM COVARIANCE DETERMINANT* (MCD)

Dalam penelitian ini metode *Minimum Covariance Determinant* (MCD) diterapkan pada data gaji pegawai pada perguruan tinggi di Amerika pada bulan Maret-April tahun 1994. Data tersebut di*download* dari http://www.amstat.org/publications/jse/jse_data_archive.htm yang terdiri dari 1.074 pengamatan dan 13 variabel, yaitu: X_1 = rata-rata gaji professor ($), X_2 = rata-rata gaji associate profesor ($), X_3 = rata-rata gaji asisten profesor ($), X_4 = rata-rata gaji semua karyawan fakultas ($), X_5 = rata-rata kompensasi profesor ($), X_6 = rata-rata kompensasi associate profesor ($), X_7 = rata-rata kompensasi asisten professor ($), X_8 = rata-rata kompensasi semua karyawan fakultas ($), X_9 = jumlah professor, X_{10} = jumlah associate professor, X_{11} = jumlah asisten professor, X_{12} = jumlah instruktur, X_{13} = jumlah karyawan fakultas. Untuk mendeteksi apakah data gaji pegawai pada perguruan tinggi di Amerika pada bulan Maret-April tahun 1994 mengandung *outlier* atau tidak, maka langkah-langkah yang dilakukan adalah 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 yang selanjutnya

digunakan untuk mendeteksi *outlier* dengan metode MCD. Berikut ini merupakan hasil *score* komponen utama untuk setiap pengamatan.

Tabel 1. *Score* komponen utama untuk setiap pengamatan

Pengamatan	K_1	K_2
1	303,8079	-34,5539
2	-332,922	-314,785
3	153,6219	-210,428124
4	-267,852	-205,073284
5	145,0784	195,432335
6	237,0046	71,391286
7	273,9556	-1,593806
8	-832,575	440,143534
9	211,3586	-56,045491
10	108,5461	7,578159
⋮	⋮	⋮
1.074	-318,6	177,775
$\overline{\mathbf{M}}$	-1,1125E-14	-1,39E-14

Sebelum menghitung jarak mahalanobis, terlebih dahulu menghitung varians kovarians dari *score* komponen utama yang diperoleh dari AKU. Berikut ini merupakan hasil perhitungan jarak mahalanobis untuk setiap pengamatan.

Tabel 2. Jarak Mahalanobis Untuk Setiap Pengamatan

Pengamatan	Md_i
1	0,75691
2	1,70385
3	1,06956
4	1,17397
5	0,9954
6	0,66917
7	0,66625
8	2,91438
9	0,57916
10	0,26642
⋮	⋮
1.074	1,14767

Nilai rata-rata dan varians-kovarians dari metode MCD dapat dilihat pada tabel berikut:

Tabel 3. Nilai rata-rata dan varians-kovarians dengan metode MCD

	Nilai varians-kovarians		Nilai rata-rata
	K_1	K_2	
K_1	55.650	57.495	216,7
K_2	57.495	70.846	-11,98

Setelah didapat nilai rata-rata dan varians-kovarians dengan metode MCD, selanjutnya adalah menghitung jarak *robust*. Berikut ini merupakan hasil perhitungan jarak *robust* untuk setiap pengamatan.

Tabel 5. Jarak *Robust* Untuk Setiap Pengamatan

Pengamatan	RD_i
1	1,1150955
2	3,4008813
3	1,2741463
4	3,5330079
5	2,6479195
6	0,5895633
7	0,5163942
8	15,0323827
9	0,3609889
10	1,3101590
⋮	⋮
1.074	7,3046398

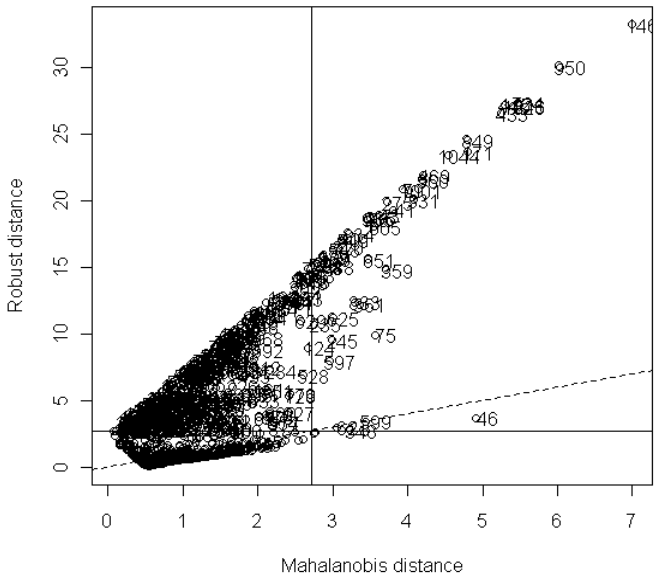
Nilai *cut-off* pada jarak *robust* adalah $\sqrt{\chi^2_{2;\alpha}} = \sqrt{\chi^2_{2;\,0,025}} = 2{,}71620$. Nilai jarak *robust* yang lebih besar dari nilai *cut-off* dideteksi sebagai *outlier*. Sedangkan nilai jarak *robust* yang lebih kecil dari nilai *cut-off* bukan dideteksi sebagai *outlier*. 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:

Tabel 6. Pendeteksian *Outlier* Untuk Setiap Pengamatan Pada Jarak *Robust*

Pengamatan	Pendeteksian *Outlier*
1	0
2	1
3	0
4	1
5	0
6	0
7	0
8	1
9	0
10	0
⋮	⋮
1.074	1

Berdasarkan Tabel 6diatas, terdapat 309 pengamatan terdeteksi sebagai *outlier*. Untuk mengklasifikasikan *outlier* tersebut dapat dilakukan dengan cara membuat plot jarak mahalanobis *versus* jarak *robust* yang disebut dengan *diagnostic plot*. Dari Gambar 2 dapat dilihat bahwa sebanyak 309 pengamatan terletak pada kuadran 1 dan kuadran 2 yang berarti bahwa pengamatan-pengamatan tersebut terdeteksi sebagai *outlier*. Pengamatan yang terletak di kuadran 1 termasuk jenis *bad leverage*, yaitu sebanyak 48 pengamatan. Hal itu berarti bahwa pengamatan-pengamatan tersebut sangat berpengaruh padahal tidak cocok untuk model regresi. Keberadaan titik *bad leverage* ini dapat merubah garis regresi sehingga dapat mempengaruhi hasil secara keseluruhan. Karena akibat yang ditimbulkan oleh titik *bad leverage* ini sangat berbahaya, maka tindakan yang sebaiknya dilakukan adalah menghapus seluruh pengamatan yang tergolong jenis *bad leverage* tersebut. Sedangkan pada kuadran 2 terdapat 261 pengamatan yang berarti termasuk tipe *outlier* ortogonal. Pengamatan-pengamatan tersebut berpengaruh pada pendugaan parameter regresi. Alternatif yang dapat dilakukan adalah tetap mengikutsertakan pengamatan tersebut tetapi pendugaan parameternya menggunakan

metode *Least Median Square* (LMS). Pengamatan yang terletak di kuadran 3 sebanyak 762 pengamatan yang termasuk tipe pengamatan biasa. Pengamatan-pengamatan tersebut bukan merupakan *outlier*. Pengamatan yang terdapat di kuadran 4 termasuk tipe *good leverage*, yaitu sebanyak 3 pengamatan. Pengamatan jenis *good leverage* ini tetap dapat diikutsertakan dalam analisis data. Gambar 2 menyajikan *diagnostic plot* untuk setiap pengamatan.



Gambar 2 *Diagnostic plot* jarak mahalanobis vs jarak *robust*

PENUTUP

Berdasarkan hasil kajian yang telah dilakukan, dapat ditarik kesimpulan bahwa 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* tersebut dapat dilakukan dengan cara membuat plot jarak mahalanobis *versus* jarak *robust* yang disebut dengan *diagnostic plot*.

Penerapan metode MCD pada data gaji pegawai Perguruan Tinggi di Amerika pada bulan Maret-April tahun 1994 menghasilkan sebanyak 309 pengamatan terdeteksi sebagai *outlier*. Dari 309 pengamatan tersebut, terdapat 48 pengamatan yang termasuk jenis *bad leverage*. Hal itu berarti bahwa pengamatan-pengamatan tersebut sangat berpengaruh padahal tidak cocok untuk model regresi. Oleh karena itu, 48 pengamatan ini tidak boleh diikutsertakan dalam analisis data atau dengan kata lain 48 pengamatan ini sebaiknya dihapus. Sedangkan sebanyak 261 pengamatan termasuk jenis *outlier* ortogonal. Alternatif yang dapat dilakukan adalah tetap mengikutsertakan pengamatan tersebut tetapi pendugaan parameternya menggunakan metode *Least Median Square* (LMS). Pengamatan yang terletak di kuadran 3 sebanyak 762 pengamatan yang termasuk tipe pengamatan biasa. Pengamatan-pengamatan tersebut bukan merupakan *outlier*. Pengamatan yang terdapat di kuadran 4 termasuk tipe *good leverage*, yaitu sebanyak 3 pengamatan.

DAFTAR PUSTAKA

[1]. Barnett V, Lewis T. *Outliers in Statistical Data*. New York: John Wiley & Sons; 1980.

[2]. Sunaryo S, Setiawan, Siagian TH. Mengatasi Masalah Multikolinearitas dan *Outlier* dengan Pendekatan ROBPCA (Studi Kasus Analisis Regresi Angka Kematian Bayi di Jawa Timur). Jurnal *Matematika,Sains dan Teknologi*. 2011; 12:1-10.

[3]. Suryana. Analisis Data Outlier pada Data Pengeluaran Rumah Tangga di Kota Kupang, Nusa Tenggara Timur Tahun 2005 dengan Metode ROBPCA, Institut Teknologi Sepuluh Nopember Fakultas Matematika dan IlmuPengetahuan Alam. Surabaya. 2007.
[4]. Cahyawati D, Tanuji H, Abdiati R. Efektivitas Metode Regresi Robust Penduga Welsch dalam Mengatasi Pencilan pada Pemodelan Regresi Linear Berganda. *Jurnal Penelitian Sains*. 2009;12:1-7.
[5]. Makkulau, Linuwih S, Puhardi, Mashuri M. Pendeteksian *Outlier* dan Penentuan Faktor-Faktor yang Mempengaruhi Produksi Gula dan Tetes Tebu dengan Metode *Likelihood Displacement Statistic-Lagrange*. *Jurnal Teknik Industri*. 2010;12:95-100.
[6]. Darmanto. *Pengantar Analisis Regresi*. Malang: Universitas Brawijaya; 2010.
[7]. Filzmoser P. Identification of Multivariate Outliers: A Performance Study. *Austrian Journal Of Statistics*. 2005;34:127–138.
[8]. Rousseuw PJ, Driessen KV. A Fast Algorithm for The MinimumCovariance Determinant Estimator. *Journal Technometrics*. 1998;41:212-223.
[9]. Rencher AC. *Methods of Multivariate Analysis*. Ed ke-2. Canada: John Wiley &Sons; 2002.

MUKTI KURNIADI : FMIPA UNTAN Pontianak, mukti_f3@yahoo.co.id
MARISI ARITONANG : FAPERTA UNTAN Pontianak, hetty_aritonang@ymail.com.
MUHLASAH NOVITASARI MARA : FMIPA UNTAN Pontianak, novee_mara@yahoo.co.id