Pendeteksian *Outlier* pada Pengamatan dalam Model Linear Multivariat dengan Metode *Likelihood Displacement Statistic-Lagrange*

Outlier Detection in Observation at Multivariate Linear Models with Likelihood Displacement Statistic-Lagrange Method

Makkulau [1)], Susanti Linuwih [2)], Purhadi [2)] & Muhammad Mashuri [2)]
[1)]*Jurusan Matematika, FMIPA Universitas Haluoleo, Kendari*
[2)]*Jurusan Statistika, FMIPA Institut Teknologi Sepuluh Nopember*

ABSTRACT

There are two different outliers, i.e outlier in observations and outlier in models. The existing outlier detection method in models is using common Likelihood method. The limitation of this method is the optimal value produced might be not the real optimal values. This research yields a method for outlier detection in multivariate linear models with Likelihood Displacement Statistic-Lagrange method (LDL method). This method uses multiplier Lagrange with constraint the confidence interval of parameter's vector. This parameter's vector is obtained from the data set which is outlier free. This parameter estimation process uses numerical method with Karush-Kuhn Tucker condition in nonlinear programming. This method compares between LDL value and the table F value that follows the distribution of F value to indentify the outlier in models.

Keywords : Distribution of F, Likelihood Displacement Statistic-Lagrange, multivariate linear models, nonlinear programming, outlier detection

PENDAHULUAN

Outlier merupakan pengamatan yang menyimpang sedemikian jauh dari pengamatan lain (Hawkins 1980), dapat mempunyai efek bagi pengambilan suatu kesimpulan atau keputusan pada penelitian. *Outlier* dibedakan atas *outlier* pada pengamatan univariat atau multivariat dan *outlier* pada model linear univariat atau multivariat.

Pendeteksian *outlier* pada model linear telah dilakukan antara lain oleh Srivastava & von Rosen (1998), Cook (2000), Adnan *et al.* (2003), dan Diaz-Garcia *et al.* (2007). Xu *et al.* (2005) mengembangkan jarak Cook's univariat untuk mendeteksi *outlier* pada model linear multivariat. Metode yang digunakan adalah Metode *Likelihood Displacement Statistic* (Metode LD), yaitu suatu metode yang menghilangkan pengamatan yang *outlier* pada model; Metode *Likelihood Ratio Statistic for a Mean Shift* (Metode LR), yaitu suatu metode dengan cara pergeseran rata-rata pada model; dan Metode *Multivariate Leverage* yang mengguna-kan elemen dari *the average diagonal* Q_{A_m} untuk mengukur keekstriman dari *m* pengukuran pada variabel independen. Makkulau *et al.* (2008) membahas prosedur Metode LD sedangkan Makkulau *et al.* (2009) membahas aplikasinya di Pabrik Gula Djombang Baru Jombang Provinsi Jawa Timur. Xu *et al.* (2005) dalam mengestimasi parameter dengan Metode LD dari model linear multivariat bersifat umum, sehingga nilai optimal yang diperoleh dari fungsi tujuan dapat saja bukan merupakan nilai yang paling optimal. Oleh karena itu digunakan pengganda *Lagrange,* sehingga nilai optimal yang diperoleh merupakan nilai yang paling optimal pada daerah kepercayaan yang telah ditentukan.

Penelitian ini mengkaji tentang pendeteksian *outlier* pada pengamatan dalam model linear multivariat sebagai pengembangan Metode LD dengan menggunakan pengganda *Lagrange* yang disebut Metode *Likelihood Displacement Statistic-Lagrange* (Metode LDL). Pengganda *Lagrange* yang digunakan adalah daerah kepercayaan dari vektor parameter dimana pengamatan yang diduga *outlier* telah dihilangkan.

Model linear multivariat

Model linear multivariat adalah model linear dengan variabel dependen lebih dari satu (Christensen 1991). Misalkan $X_1, X_2, \ldots, X_p$ adalah variabel independen dan $Y_1, Y_2, \ldots, Y_q$

adalah variabel dependen, jika dilakukan *n* pengamatan yang diambil pada setiap variabel dependen yang ditulis $y_{i1}, y_{i2}, \ldots, y_i$ dimana $i = 1, 2, \ldots, n,$ atau:

$$y_{ih} = \beta_{0h} + \beta_{1h} X_{i1} + \beta_{2h} X_{i2} + \ldots + \beta_{ph} X_{ip}$$

dimana $h = 1, 2, \ldots, q$.

Model linear multivariat yang terdiri dari *q* model linear secara simultan dapat ditulis sebagai:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (1)$$

dengan

$\mathbf{Y}_{nxq} = [Y_1, Y_2, \ldots, Y_q]$

$\mathbf{X}_{n(p+1)} = [1, X_1, X_2, \ldots, X_p]$

$\mathbf{B}_{(p+1)xq} = [\beta_1, \beta_2, \ldots, \beta_q]$

dan $\mathbf{E}_{nxq} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_q]$

Estimasi parameter dan uji hipotesis model linear multivariat

Pada model linear multivariat matriks *error* $\mathbf{E}_{nxq} = (\varepsilon_{ih})$ merupakan matriks acak, dimana $i = 1, 2, \ldots, n$ dan $h = 1, 2, \ldots, q$. Diasumsikan juga bahwa $E(\mathbf{E}) = \mathbf{0}$ dan matriks varian-kovariansi-nya adalah $\mathrm{Var}(\mathbf{E}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$, dimana $\otimes$ adalah perkalian Kronecker dan $\boldsymbol{\Sigma} = (\sigma_{hh^*})$ dengan $h = 1, 2, \ldots, q$, $h^* = 1, 2, \ldots, q$, sehingga $\mathbf{E} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$.

Dengan mengestimasi parameter $\mathbf{B}$ pada (1), maka diperoleh:

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (2)$$

Estimasi parameter $\boldsymbol{\Sigma}$ pada (1), yaitu:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$$

$\hat{\boldsymbol{\Sigma}}$ adalah estimator bias untuk $\boldsymbol{\Sigma}$. Sedangkan estimasi parameter $\boldsymbol{\Sigma}$ yang lain adalah:

$$\mathbf{S} = \frac{1}{n - \mathrm{rank}(\mathbf{X})}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$$

merupakan estimator tak bias untuk $\boldsymbol{\Sigma}$.

Vektorisasi matriks variabel dependen pada (1) ditulis Vec($\mathbf{Y}$) (Christensen 1991) adalah:

$$\mathrm{Vec}(\mathbf{Y}) = (\mathbf{I}_q \otimes \mathbf{X})\mathrm{Vec}(\mathbf{B}) + \mathrm{Vec}(\mathbf{E})$$

Dimana

$$\mathrm{Vec}(\mathbf{E}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n) \quad (3)$$

dan $\mathrm{Vec}(\mathbf{Y}) \sim N_{nq}((\mathbf{I}_q \otimes \mathbf{X})\mathrm{Vec}(\mathbf{B}), \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$

Dengan menggunakan sifat hasil kali *kronecker* diperoleh:

$$\mathrm{Vec}(\hat{\mathbf{B}}) = (\mathbf{I}_q \otimes (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathrm{Vec}(\mathbf{Y})$$

dan distribusi $\mathrm{Vec}(\hat{\mathbf{B}})$ adalah:

$$\mathrm{Vec}(\hat{\mathbf{B}}) \sim N_{q(p+1)}(\mathrm{Vec}(\mathbf{B}), \boldsymbol{\Sigma} \otimes (\mathbf{X}^T\mathbf{X})^{-1})$$

Prosedur uji hipotesis parameter pada model linear multivariat adalah:

$H_0 : \Lambda^T\mathbf{B} = 0$ terhadap $H_1 : \Lambda^T\mathbf{B} \neq 0$

dimana $\Lambda^T = \mathbf{P}^T\mathbf{X}$ dan $\mathbf{P}$ adalah matriks ortogonal (Christensen 1991). Uji ini didasarkan pada statistik uji:

$$\mathbf{H} = \mathbf{Y}^T\mathbf{M}_{\mathbf{MP}}\mathbf{Y} = (\Lambda^T\hat{\mathbf{B}})^T(\Lambda^T(\mathbf{X}^T\mathbf{X})^{-1}\Lambda)^{-1}(\Lambda^T\hat{\mathbf{B}})$$

dimana $\mathbf{M}_{\mathbf{MP}} = \mathbf{MP}(\mathbf{P}^T\mathbf{MP})^{-1}\mathbf{P}^T\mathbf{M}$.

$\mathbf{M}_{\mathbf{MP}}$ adalah matriks proyeksi pada $\mathbf{H}$.

$$\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$\mathbf{M}$ adalah matriks proyeksi pada ruang kolom dari **X.**

Hipotesis nol ditolak jika nilai maksimum dari *likelihood* di bawah H_0 lebih besar dari nilai maksimum keseluruhan.

***Outlier* pada pengamatan dalam model linear multivariat**

Outlier pada pengamatan dalam model linear multivariat dapat dibagi atas 3 kategori, yaitu *outlier* terhadap nilai *X*; *outlier* terhadap nilai *Y*; dan *outlier* terhadap nilai *X* dan *Y*.

Outlier dapat diklasifikasikan ke dalam empat kelompok berdasarkan penyebabnya, yaitu observasi umum; titik *leverage* baik; *outlier* vertikal; dan titik *leverage* jelek (Rousseeuw & Hubert 1997).

Metode pendeteksian *Outlier* pada pengamatan dalam model linear multivariat

Fungsi *likelihood* dalam model linear multivariat ditulis sebagai berikut (Christensen 1991 serta Rencher & Schaalje 2008):

$$L(\mathbf{B}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp(\Delta) \quad (4)$$

dimana $\Delta = \left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB}))\right)$

Pendeteksian *outlier* pada pengamatan dalam model linear multivariat oleh Xu *et al.* (2005) mengembangkan jarak Cook's univariat

dengan menggunakan Metode LD, Metode LR, dan Metode *Multivariate Leverage*. Pendeteksian *outlier* dengan Metode LD dilakukan dengan cara menghilangkan pengamatan yang *outlier* pada model. Misalkan ada m pengamatan dikumpulkan pada himpunan tertentu, dengan m pengamatan diduga *outlier*. Indeks A_m adalah kumpulan dari m pengamatan yang diduga *outlier*. Dengan kata lain, indeks A_m artinya ada *outlier*, sehingga:

$\mathbf{Y}_{A_m}$: himpunan dari variabel dependen dengan pengamatan yang ada *outlier*.

$\mathbf{Y}^{C}_{A_m}$: himpunan dari variabel dependen dengan pengamatan tanpa *outlier*.

Berikut adalah definisi dari fungsi *Likelihood Displacement Statistic* (LD) untuk pengamatan yang ada *outlier*.

Definisi 1 (Christensen 1991)
LD dengan pengamatan yang ada *outlier* untuk $\mathbf{B}$ dengan diberikan $\boldsymbol{\Sigma}$ adalah:

$$LD_{A_m}(\mathbf{B}|\boldsymbol{\Sigma}) = \left(2 \ln\left(\hat{\mathbf{B}}, \hat{\boldsymbol{\Sigma}}\right) - \ln\left(\hat{\mathbf{B}}_{A_m}^{C}\left(\hat{\boldsymbol{\Sigma}}^{C}\right)\right)\right)_A \quad (5)$$

dimana $\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{A_m}\right)$ adalah estimasi maksimum *likelihood* dari $\boldsymbol{\Sigma}$ ketika $\mathbf{B}$ diestimasi oleh $\hat{\mathbf{B}}^{C}_{A_m}$.

Definisi 2 (Christensen 1991)
LD dengan pengamatan yang ada *outlier* dan bersyarat adalah:

$$LD_{A_m}(\mathbf{B}_1, \boldsymbol{\Sigma} \mid \mathbf{B}_2, \boldsymbol{\Sigma})_2 = \left(2\left(\ln \hat{\mathbf{B}}, \boldsymbol{\Sigma} \right) - \ln \right) \quad (6)$$

dimana:

$$= \left(\hat{\mathbf{B}}^{C}_{1A_m}, \hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{1A_m}\right)\right)$$

$$= \left\{\left(\hat{\mathbf{B}}^{C}_{2A_m}, \hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{2A_m}\right)\right)\right\}\left(\hat{\mathbf{B}}^{C}_{1A_m}, \hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{1A_m}\right)\right)$$

$\{\ \}$ menotasikan suatu fungsi yang bentuknya $\hat{\boldsymbol{\theta}}_2\left(\hat{\boldsymbol{\theta}}^{C}_{1A_m}\right)$

Estimator dari $\mathbf{B}$ dengan pengamatan tanpa *outlier* adalah:

$$\hat{\mathbf{B}}^{C}_{A_m} = \hat{\mathbf{B}} - \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{K}^{T}_{A_m}\left(\mathbf{Q}_{A_m}\right)\hat{\mathbf{E}}^{-1}_{A_m},$$

dimana

$$\mathbf{Q}_{A_m} = \mathbf{X}_{A_m}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}_{A_m}; \quad \hat{\mathbf{E}}_{A_m} = \mathbf{Y}_{A_m} - \mathbf{X}\hat{\mathbf{B}}$$

$$\mathbf{I} + \left(\mathbf{I} - \mathbf{Q}_{A_m}\right)^{-1}\mathbf{Q}_{A_m} = \left(\mathbf{I} - \mathbf{Q}_{A_m}\right)^{-1}$$

dan estimator dari $\boldsymbol{\Sigma}$ dengan pengamatan tanpa *outlier* yaitu $\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{A_m}\right)$ adalah:

$$\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{A_m}\right) = \frac{n}{n-m}\hat{\boldsymbol{\Sigma}} - \frac{1}{n-m}\hat{\mathbf{E}}^{T}_{A_m}\left(\mathbf{I}_{A_m}\right)^{-1},$$

sehingga fungsi *likelihood* dalam model linear multivariat dengan pengamatan yang ada *outlier* adalah:

$$LD_{A_m}(\mathbf{B}|\boldsymbol{\Sigma}) = \left(2 \ln\left(\hat{\mathbf{B}}, \hat{\boldsymbol{\Sigma}}\right) - \ln\left(\hat{\mathbf{B}}^{C}_{A_m}\left(\hat{\boldsymbol{\Sigma}}^{C}\right)\right)\right)$$

Optimasi nonlinear

Masalah optimasi ditentukan oleh karakteristik fungsi tujuan dan fungsi kendala. Permasalahan optimasi disebut nonlinear jika fungsi tujuan dan fungsi kendalanya mempunyai bentuk nonlinear pada salah satu atau keduanya (Bazaara *et al.*1993).

Pada optimasi dengan kendala, persamaan yang akan dioptimasi dapat dituliskan sebagai berikut:

Maksimumkan (minimumkan):

$$g(\mathrm{Z}) = g(Z_1, Z_2, \ldots, Z_p)$$

Kendala : $k_p(\mathrm{Z}) = b_p$ dan $\mathrm{Z} \geq 0$.

Optimasi dengan kendala ini dapat diselesaikan dengan menggunakan pengganda *Lagrange* seperti berikut ini:

$$L = g(\mathrm{Z}) - \sum_{p=1}^{n}\lambda_p\left(k_p(\mathrm{Z}) - b_p\right).$$

METODE

Pembahasan dalam makalah ini dilakukan dengan pendekatan teoretik. Secara garis besar, langkah pertama yang dilakukan dalam penelitian ini adalah mencari solusi persamaan Schrodinger PDM sistem osilator harmonik secara analitik, yaitu berupa nilai eigen energi (En) dan fungsi eigen (Ψn), dengan menggunakan metode transformasi. Langkah kedua adalah melakukan aproksimasi $m(x)$=1 dan $\mu(x)$=x untuk melihat apakah hasilnya mereduksi menjadi hasil untuk sistem osilator harmonik dengan massa konstan. Langkah kedua ini sekaligus sebagai verifikasi apakah hasil yang didapat konsisten atau tidak. Pada bagian akhir akan ditinjau pula beberapa bentuk fungsi massa yang bergantung posisi dan potensial yang dibangkitkannya, serta interpretasinya secara fisis.

Pendeteksian *outlier* dengan Metode LDL berdasarkan langkah-langkah:

a. Mengumpulkan m pengamatan yang diduga *outlier*.
b. Mendeteksi *outlier* pada model dengan asumsi $\mathrm{Ved}(\mathbf{E}) \sim N_p(\boldsymbol{\Sigma}0, \mathbf{I})$, dimulai dengan membuat

$L(\mathbf{B},\boldsymbol{\Sigma})$, lalu menentukan $\hat{\mathbf{B}}$ dan $\hat{\boldsymbol{\Sigma}}$ untuk mendapatkan $L(\hat{\mathbf{B}},\hat{\boldsymbol{\Sigma}})$.

c. Memaksimumkan $L(\mathbf{B},\boldsymbol{\Sigma})$ dengan kendala jika ada *m* buah peng-amatan adalah *outlier* menggunakan pengganda *Lagrange*, lalu membuat $\ln L(\mathbf{B},\boldsymbol{\Sigma})$ dan menentukan $\mathrm{Vec}(\hat{\mathbf{B}})$ dengan program nonlinear.

HASIL DAN PEMBAHASAN

Nilai eigen energi E_n dan fungsi eigen $\Psi(x)$ sistem osilator harmonik PDM

Penelitian ini dibatasi hanya pada pendeteksian *outlier* pada pengamatan dalam model linear multivariat dengan Metode LDL. Pendeteksian *outlier* dalam model linear multivariat dimulai dengan memisalkan ada *m* pengamatan yang diduga *outlier* (A_m) dari $Y_1, Y_2, \ldots, Y_n, \ldots, Y_{n+m}$, sehingga $\mathbf{Y}_{A_m}$ adalah himpunan dari variabel dependen dengan pengamatan yang ada *outlier* dan $\mathbf{Y}^{C}_{A_m}$ adalah himpunan dari variabel dependen dengan pengamatan tanpa *outlier*. Sebelumnya, jika dipunyai variabel independen sebanyak *p* dan variabel dependen sebanyak *q*, maka model linear multivariat (1) secara simultan dapat ditulis sebagai:

$$\begin{aligned}\mathbf{Y}_{nxq} &= \left(\mathbf{J}_{nx1} \; \mathbf{X}_{1_{nx(p+1)}}\right)\mathbf{B}_{(p+1)xq} + \mathbf{E}_{nxq} \\ &= \mathbf{X}_{nx(p+1)}\mathbf{B}_{(p+1)xq} + \mathbf{E}_{nxq} \qquad (7)\end{aligned}$$

Dari (7) dapat ditulis dalam bentuk vektor:

$$\mathrm{Vec}(\mathbf{Y}) = (\mathbf{I}_q \otimes \mathbf{X})\mathrm{Vec}(\mathbf{B}) + \mathrm{Vec}(\mathbf{E})$$

Pendeteksian *outlier* pada dalam model linear multivariat dengan asumsi seperti pada (3) dimulai dengan membuat fungsi *likelihood* untuk populasi $L(\mathbf{B},\boldsymbol{\Sigma})$, lalu menentukan $\hat{\mathbf{B}}$ dan $\hat{\boldsymbol{\Sigma}}$. Estimasi parameter $\mathbf{B}$ pada (7) dengan fungsi *likelihood* seperti pada (4) dengan Metode MLE dimulai dengan meloga-ritmanaturalkan (4), sehingga:

$$\ln L(\mathbf{B},\boldsymbol{\Sigma}) = -\frac{nq}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\right) \qquad (8)$$

Kondisi optimal (maksimum atau minimum) dicapai bila memenuhi kondisi berikut ini:

Jika logaritma natural dari fungsi *likelihood* (8) diturunkan terhadap $\mathbf{B}$ dan disama-kan dengan nol, maka:

$$\frac{\partial \ln L(\mathbf{B},\boldsymbol{\Sigma})}{\partial \mathbf{B}} = -\frac{1}{2}\mathrm{tr}\left(-2\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})^T\mathbf{X}\right) = 0$$

diperoleh:

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Kemudian jika (8) diturunkan terhadap $\boldsymbol{\Sigma}$ dan disamakan dengan nol, maka:

$$\frac{\partial \ln L(\mathbf{B},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\mathrm{tr}\left(n\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})\right) = 0,$$

diperoleh:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})$$

Untuk memaksimumkan fungsi *likelihood* $L(\mathbf{B},\boldsymbol{\Sigma})$ dengan kendala $(\mathrm{Vec}(\hat{\mathbf{B}})-\mathrm{Vec}(\mathbf{B}))^T(\hat{\mathrm{Var}}(\mathrm{Vec}(\hat{\mathbf{B}}))^{-1}(\mathrm{Vec}(\hat{\mathbf{B}})-\mathrm{Vec}(\mathbf{B}))$ jika ada *m* buah variabel dependen adalah *outlier* dengan menggunakan pengganda *Lagrange*, dimulai dengan membuat $\ln L(\mathbf{B},\boldsymbol{\Sigma})$ dan menentukan $\mathrm{Vec}(\hat{\mathbf{B}})$ dengan program nonlinear. Metode ini disebut dengan Metode LDL yaitu suatu metode yang menghilangkan pengamatan yang *outlier* pada model dengan kendala yang menggunakan pengganda *Lagrange*.

Fungsi *likelihood* untuk $\mathbf{B}$ yaitu (4), sehingga fungsi *likelihood* untuk $\hat{\mathbf{B}}^{C}_{A_m}$ adalah:

$$L(\mathbf{B}^{C}_{A_m},\boldsymbol{\Sigma}^{C}_{A_m}) = (2\pi)^{-\frac{(n-m)p}{2}}\left|\boldsymbol{\Sigma}^{C}_{A_m}\right|^{-\frac{n-m}{2}}\exp(\Omega)$$

dimana

$$\Omega = -\frac{1}{2}\mathrm{tr}\left((\boldsymbol{\Sigma}^{C}_{A_m})^{-1}(\mathbf{Y}^{C}_{A_m}-\mathbf{X}^{C}_{A_m}\hat{\mathbf{B}}^{C}_{A_m})^T(\mathbf{Y}^{C}_{A_m}-\mathbf{X}^{C}_{A_m}\hat{\mathbf{B}}^{C}_{A_m})\right)$$

Berdasarkan (7), diperoleh:

$\mathbf{Y}^{C}_{A_m} = \mathbf{X}^{C}_{A_m}\mathbf{B}^{C}_{A_m}$; dan

$\mathrm{Vec}(\mathbf{Y}^{C}_{A_m}) \sim N_{pn}(\boldsymbol{\mu}, \boldsymbol{\Sigma}\otimes\mathbf{I}_{n-m})$,

dan berdasarkan (2) diperoleh:

$$\hat{\mathbf{B}}^{C}_{A_m} = \left(\left(\mathbf{X}^{C}_{A_m}\right)^T\mathbf{X}^{C}_{A_m}\right)^{-1}\left(\mathbf{X}^{C}_{A_m}\right)^T\mathbf{X}^{C}_{A_m}$$

dimana

$\left(\mathbf{X}^{C}_{A_m}\right)^T\mathbf{X}^{C}_{A_m} = \mathbf{X}^T\mathbf{X}$ - .

$\left(\mathbf{X}^{C}_{A_m}\right)^T\mathbf{Y}^{C}_{A_m} = \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T_{A_m}\mathbf{Y}_{A_m}$.

Estimasi dari $\mathbf{B}$ setelah *outlier* dikeluarkan $\left(\hat{\mathbf{B}}^{C}_{A_m}\right)$ adalah:

$$\hat{\mathbf{B}}^{C}_{A_m} = \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T\left(\mathbf{X}\hat{\mathbf{B}} + \mathbf{Y}_{A_m} - \mathbf{X}_{A_m}\hat{\mathbf{B}} \right)$$

$$= \hat{\mathbf{B}} - \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T_{A_m}\left(\mathbf{I} - \mathbf{Q}_{A_m} \right)^{-1}\left(\mathbf{Y}_{A_m} - \mathbf{X}_{A_m}\hat{\mathbf{B}} \right)$$

sehingga:

$$\hat{\mathbf{B}}^{C}_{A_m} = \hat{\mathbf{B}} - \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T_{A_m}\left(\mathbf{Q}_{A_m} - \mathbf{I} \right)^{-1}\hat{\mathbf{E}}_{A_m}$$

dan diperoleh pula:

$$\hat{\mathbf{B}}^{C}_{A_m} \sim N\left(\mathbf{B}, \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\hat{\boldsymbol{\Sigma}} + \Delta \right)$$

dimana

$$\Delta = \left(\left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T_{A_m}\left(\mathbf{Q}_{A_m} - \mathbf{I} \right)^{-1} \right) \mathrm{Var}\left(\hat{\mathbf{E}}_{A_m} \right) \left(\left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T_{A_m}\left(\mathbf{Q}_{A_m} - \mathbf{I} \right)^{-1} \right)^T$$

Permasalahan di atas bersifat umum, sehingga nilai optimal yang diperoleh bisa saja bukan nilai yang paling optimal. Oleh karena itu digunakan pengganda *Lagrange,* sehingga nilai optimal yang diperoleh diharapkan merupakan nilai yang paling optimal pada daerah kepercayaan yang telah ditentukan. Secara umum daerah kepercayaan $(1-\alpha)100\%$ untuk $\mathrm{Vec}(\mathbf{B})$ adalah:

$$\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B} \right) \right)^T \left(\widehat{\mathrm{Var}}\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) \right) \right)^{-1} \left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B} \right) \right) \leq$$

sehingga daerah kepercayaan $(1-\alpha)\ 100$ untuk model dimana *outlier*-nya telah dihilangkan $\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) \right)$ adalah:

$$\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B} \right) \right)^T \left(\widehat{\mathrm{Var}}\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) \right) \right)^{-1} \left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B} \right) \right) \leq K \quad (9)$$

untuk $F_{v_1, v_2, \alpha} = K$

dimana $v_1 = p$ dan $v_2 = n - p - 1$.

Untuk menyelesaikan permasalahan nonlinear di atas digunakan pengganda *Lagrange*. Misalkan fungsi tujuan (8) dengan kendala (9), maka fungsi *Lagrange*-nya dapat ditulis:

$$F\left(\mathbf{B}, \boldsymbol{\Sigma}, \lambda \right) = \ln L\left(\mathbf{B}, \boldsymbol{\Sigma} \right) - \left\{ F^T S F - K \right\} \quad (10)$$

dimana:

$$\mathbf{F} = \left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B}^{C}_{A_m} \right) \right)$$

dan $\mathbf{S} = \widehat{\mathrm{Var}}\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) \right)^{-1}$

Kondisi optimal dicapai bila memenuhi kondisi berikut ini:

Fungsi *Lagrange* (10) diturunkan terhadap $\mathbf{B}$ dan disamakan dengan nol, maka:

$$\frac{\partial F\left(\mathbf{B}, \boldsymbol{\Sigma}, \lambda \right)}{\partial \mathbf{B}} = \frac{\partial \ln L\left(\mathbf{B}, \boldsymbol{\Sigma} \right)}{\partial \mathbf{B}} = \mathbf{0}$$

Berdasarkan (8), maka diperoleh:

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Dengan menggunakan sifat hasil kali *kronecker* diperoleh:

$$\mathrm{Vec}\left(\hat{\mathbf{B}} \right) = \left(\mathbf{I}_q \otimes \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T \right) \mathrm{Vec}\left(\mathbf{Y} \right).$$

dengan $\mathrm{E}\left(\mathrm{Vec}\left(\hat{\mathbf{B}} \right) \right) = \mathrm{Vec}\left(\mathbf{B} \right)$

dan $\widehat{\mathrm{Var}}\left(\mathrm{Vec}\left(\hat{\mathbf{B}} \right) \right) = \hat{\boldsymbol{\Sigma}} \otimes \left(\mathbf{X}^T\mathbf{X} \right)^{-1}$

Sehingga untuk pengamatan dalam model linear multivariat dengan *m outlier* dihilangkan, diperoleh:

$$\hat{\mathbf{B}}^{C}_{A_m} = \hat{\mathbf{B}} - \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T_{A_m}\left(\mathbf{Q}_{A_m} - \mathbf{I} \right)^{-1}\hat{\mathbf{E}}_{A_m}$$

dengan $E\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) \right) = \mathrm{Vec}\left(\mathbf{B}^{C}_{A_m} \right)$

Fungsi *Lagrange* (10) diturunkan terhadap $\boldsymbol{\Sigma}$ dan disamakan dengan nol, maka:

$$\frac{\partial F\left(\mathbf{B}, \boldsymbol{\Sigma}, \lambda \right)}{\partial \boldsymbol{\Sigma}} = \frac{\partial \ln L\left(\mathbf{B}, \boldsymbol{\Sigma} \right)}{\partial \boldsymbol{\Sigma}} = \mathbf{0}$$

Berdasarkan (8), maka diperoleh:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \right)^T\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \right)$$

Fungsi *Lagrange* (10) diturunkan terhadap λ dan disamakan dengan nol, maka diperoleh:

$$\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B} \right) \right)^T \left(\widehat{\mathrm{Var}}\left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) \right) \right)^{-1} \left(\mathrm{Vec}\left(\hat{\mathbf{B}}^{C}_{A_m} \right) - \mathrm{Vec}\left(\mathbf{B} \right) \right) \leq$$

Selanjutnya menentukan nilai estimasi dari parameter yang optimal dengan metode numerik yang menggunakan kondisi Karush-Kuhn-Tucker (KKT) pada program nonlinear.

Program nonlinear Metode LDL dengan kondisi KKT adalah:

Maksimumkan:

$$F\left(\mathbf{B}, \boldsymbol{\Sigma}, \lambda \right) = \ln L\left(\mathbf{B}, \boldsymbol{\Sigma} \right) - \left\{ F^T S F - K \right\}$$

Kendala:

$$F^T S F \leq K$$

$$\mathrm{Vec}\left(\hat{\mathbf{B}} \right) = \left(\mathbf{I}_q \otimes \left(\mathbf{X}^T\mathbf{X} \right)^{-1}\mathbf{X}^T \right) \mathrm{Vec}\left(\mathbf{Y} \right)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \right)^T\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \right)$$

$$\lambda\left\{ F^T S F - K \right\} = 0$$

Solusi terakhir akan mendapatkan $X_1,\ X_2, \ldots, X_p$ yang optimal.

Untuk kasus khusus $\boldsymbol{\theta}_1$ dari $\boldsymbol{\theta}$, maka LD dapat dimodifikasi sebagai:

$$\mathrm{LDL}_{A_m}\left(\boldsymbol{\theta}_1 \middle| \boldsymbol{\theta}_2 \right) = 2\left(\ln L\left(\hat{\boldsymbol{\theta}} \right) - \ln L\left(\hat{\boldsymbol{\theta}}_{1 A_m}, \hat{\boldsymbol{\theta}}_{2 A_m} \right) \right)$$

dimana $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\Sigma})$, $\boldsymbol{\theta}_1 = \mathbf{B}$, $\boldsymbol{\theta}_2 = \boldsymbol{\Sigma}$, sehingga $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{B}}, \hat{\boldsymbol{\Sigma}})$, dan diperoleh:

$$\mathrm{LDL}_{A_m}(\mathbf{B}|\boldsymbol{\Sigma}) = 2\left(\ln L(\hat{\mathbf{B}}\hat{\boldsymbol{\Sigma}}) - \ln L\left(\hat{\mathbf{B}}^{C}_{A_m}\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{A_m}\right)\right)\right)$$

Fungsi *likelihood* dengan kendala sebanyak *m* pengamatan yang diduga *outlier* adalah:

$$L\left(\hat{\mathbf{B}}\hat{\boldsymbol{\Sigma}}, \hat{\mathbf{B}}\left(\hat{\ }^{C}_{A_m}\right)\right) = (2\pi)^{\frac{mm}{2}}\left|\hat{\mathbf{B}}\left(\ ^{C}_{A_m}\right)\right|^{-\frac{n}{2}}\exp\ (\mathrm{A})$$

dimana:

$$\mathrm{A} = -\frac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Sigma}\left(\mathbf{B}^{C}_{A_m}\right)\right)^{-1}\left(\mathbf{Y} - \mathbf{X}^{C}_{A_m}\mathbf{B}^{C}_{A_m}\right)^{T}\left(\mathbf{Y} - \mathbf{X}^{C}_{A_m}\mathbf{B}^{C}_{A_m}\right)\right\}$$

$$\hat{\mathbf{B}}^{C}_{A_m} = \hat{\mathbf{B}} - \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}_{A_m}\left(\mathbf{Q} - \mathbf{E}^{-1}\right)\hat{\ } \text{ dan}$$

$$\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}\right) = \frac{1}{n} + \left(\hat{\mathbf{E}}_{A_m}\right)^{-1}\mathbf{Q}\left(\mathbf{Q}\right)\mathbf{I}$$

Setelah mendapatkan $\hat{\mathbf{B}}^{C}_{A_m}$ dan $\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{A_m}\right)$, selanjutnya ditentukan fungsi *Likelihood* untuk pengamatan yang ada *outlier* yaitu:

$$\mathrm{LDL}_{A_m} = \mathrm{LDL}_{A_m}(\mathbf{B}|\boldsymbol{\Sigma})$$

$$= 2\left(\ln L\left(\hat{\mathbf{B}}\hat{\boldsymbol{\Sigma}}\right) - \ln L\left(\hat{\mathbf{B}}^{C}_{A_m}\hat{\boldsymbol{\Sigma}}\left(\hat{\mathbf{B}}^{C}_{A_m}\right)\right)\right) \qquad (11)$$

Dengan melogaritmanaturalkan dan membuang $\hat{\mathbf{B}}^{C}_{A_m}$ pada (11) di atas, diperoleh:

$$\mathrm{LDL}_{A_m} = n\left[\frac{n\hat{\Sigma} + \frac{1}{n}\hat{\mathbf{E}}^{T}_{A_m}\mathbf{C}_{A_m}\hat{\mathbf{E}}_{A_m}}{n\left|\hat{\Sigma}\right|}\right] \qquad (12)$$

dimana $\mathbf{C}_{A_m} = \left(\mathbf{Q} - \mathbf{Q}_{A_m}\right)^{-1}\mathbf{I}_{A_m}\left(\mathbf{Q} - \ _{A_m}\right)^{-1}$.

LDL_{A_m} didekati dengan $\mathrm{LDL}_{A} = \sum_{i=1}^{m} Z^2$,

sehingga $\mathrm{LDL}_{A} \sim F_{v_1 v_2 \alpha}$,

dimana $v_1 = p$ dan $v_2 = n - p - 1$.

Penentuan *outlier* dilakukan dengan membandingkan $\mathrm{LDL}_{A\ hitung}$ pada (12) dan F_{tabel} dengan uji hipotesis adalah:

H_0 : A_m bukan *outlier* dan

H_1 : A_m adalah *outlier*.

Jika $\mathrm{LDL}_{A\ hitung} > \lambda.F_{tabel}$, maka tolak H_0, artinya pengamatan tersebut adalah *outlier*.

KESIMPULAN

Penentuan dan pendeteksian *outlier* pada model linear multivariat menggunakan pengganda *Lagrange* yang disebut Metode LDL. Pengganda *Lagrange* yang digunakan adalah daerah kepercayaan dari vektor parameter dimana pengamatan yang *outlier* telah dihilangkan dari model. Penentuan nilai estimasi dari parameter yang optimal dilakukan melalui metode numerik dengan menggunakan kondisi KKT pada program nonlinear. Metode ini membandingkan nilai LDL hitung dengan nilai tabel *F* yang akan mengikuti distribusi *F*.

DAFTAR PUSTAKA

Adnan R, Mohamad MN & Setan H. 2003. Multiple Outliers Detection Procedures in Linear Regression. *Matematika*. **1**:29-45.

Bazaara MS, Sherali HD & Shetty CM. 1993. *Nonlinear Programming: Theory and Algorithms*. 2rd edition. John Wiley & Sons. New York.

Christensen R. 1991. *Linear Model for Multivariate, Time Series, and Spatial Data*, Springer-Verlag. New York.

Cook RD. 2000. Detection of Influential Observation in Linear Regression. *Technometrics*. **42**(1):65-68.

Diaz-Garcia JA, Gonzalez-Farias G & Alvarado-Castro V. 2007. Exact Distributions for Sensitivity Analysis in Linear Regression. *Applied Mathematical Sciences*. **22**:1083-1100.

Hawkins DM. 1980. *Identifications of Outliers.* Chapman and Hall. New York.

Makkulau, Linuwih S, Purhadi & Mashuri M. 2008. Prosedur Pendeteksian Outlier pada Model Linear Multivariat dengan Metode Likelihood Displacement Statistic. *Prosiding Seminar Nasional Matematika IV*. Jurusan Matematika FMIPA ITS. Desember 2008. Surabaya.

Makkulau, Linuwih S, Purhadi & Mashuri M. 2009. Pendeteksian Outlier Model Linear Multivariat pada Produksi Gula dan Tetes Tebu. *Prosiding Seminar Nasional Matematika*. Jurusan Matematika FMIPA Universitas Jember, Februari 2009. Jember.

Rencher AC & Schaalje GB. 2008. *Linear Models in Statistics.* 2nd edition. John Wiley & Sons. New York.

Rousseeuw PJ & Hubert M. 1997. Recent Developments in PROG-RESS, dalam L1-Statistical Procedure and Related Topics, edited by Y. Dodge. *Institute of Mathematical Statistics Lecture Notes and Monograph Series.* Hayward, California. **31**:201-214.

Srivastava. MS & Von Rosen D. 1998. Outliers in Multivariate Regression Models. *Journal of Multivariate Analysis*. **65**:195-208.

Xu J, Abraham B & Steiner SH. 2005. *Outlier Detection Methods in Multivariate Regression Models*.http://www.bisrg.uwaterloo.ca/archive/RR-06-07.pdf [4 April 2007].