PENDETEKSIAN *OUTLIER* DENGAN METODE REGRESI *RIDGE*

Sri Harini

Jurusan Matematika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
e-mail: sriharini21@yahoo.co.id

Abstrak

Dalam analisis regresi linier berganda adanya satu atau lebih pengamatan pencilan (outlier) akan menimbulkan dilema bagi para peneliti. Keputusan untuk menghilangkan pencilan tersebut harus dilandasi alasan yang kuat, karena kadang-kadang pencilan dapat memberikan informasi penting yang diperlukan. Masalah outlier ini dapat diatasi dengan berbagai metode, diantaranya metode regresi ridge (ridge regression). Untuk mengetahui kekekaran regresi ridge perlu melihat nilai-nilai R^2, PRESS, serta leverage (h_{ii}), untuk metode regresi ridge dengan berbagai nilai tetapan bias k yang dipilih.

Kata kunci: *outlier*, *PRESS*, *regresi ridge*, R^2, *Leverage* (h_{ii})

1. Pendahuluan

Pada analisis regresi berganda sering ditemui satu atau lebih pengamatan tidak sesuai dengan model yang digunakan pada sebagian besar pengamatan lainnya. Hal ini dapat terjadi karena kesalahan dalam pencatatan pengamatan-pengamatan tersebut, kesalahan alat ukur, atau karena ketidakcocokan model yang digunakan. Pengamatan semacam itu disebut pencilan *(outlier)*. Pencilan bisa dihilangkan bila ada penjelasan tentang kasus pencilan yang menunjukkan situasi khusus yang tercakup dalam model.

Pencilan dalam data regresi berganda dapat berpengaruh pada hasil analisis statistik. Pengamatan pencilan mungkin menghasilkan residual yang besar dan sering berpengaruh terhadap fungsi regresi yang dihasilkannya. Untuk itu perlu dilakukan identifikasi terhadap pencilan ini guna melihat kesalahan sampel observasi. (Walker dan Birch, 1988).

2. Kajian Pustaka

Pendeteksian Pencilan *(Outlier)*

Seringkali model regresi dibentuk dari data yang banyak mengandung kekurangan, diantaranya adalah adanya pencilan yaitu pengamatan dengan residual yang besar. Pencilan sering menyebabkan kesalahan dalam pemilihan model, dan biasanya dihilangkan. Kenyataannya, beberapa pencilan dapat memberi informasi yang berarti, misalnya pencilan timbul dari kombinasi keadaan yang tidak biasa yang mungkin penting dan perlu diselidiki lebih lanjut. Oleh karena itu adanya pencilan dalam data perlu diselidiki secara seksama, barangkali dapat diketahui ada alasan dibalik keganjilan itu. Pencilan dapat disebabkan oleh kesalahan dalam data atau status fisik yang ganjil dari obyek yang dianalisis. Kesalahan dalam data berupa gangguan, penyimpangan instrumen, kesalahan operator, atau kesalahan pencetakan (Retnaningsih, 2001).

Pendeteksian pencilan terhadap nilai-nilai variabel x, dapat menggunakan matrik topi yang didefinisikan sebagai $\mathbf{H}=X(X'X)^{-1}X'$. Unsur ke-i pada diagonal utama matrik topi disebut *leverage* (h_{ii}). Unsur h_{ii} dapat diperoleh dari $h_{ii}=x_i(X'X)^{-1}x_i'$. Nilai diagonal h_{ii} terletak antara 0 dan 1 dan jumlahnya sama dengan p, yaitu banyak parameter regresi di dalam fungsi regresi termasuk suku intersep (Neter, Wasserman, dan Kutner, 1990).

Nilai *leverage* yang besar menunjukkan pencilan dari nilai-nilai variabel x untuk pengamatan ke-i. Hal ini disebabkan, bahwa h_{ii} adalah ukuran jarak antara nilai x untuk pengamatan ke-i dengan rata-rata nilai x untuk semua pengamatan. Sehingga, nilai h_{ii} yang besar menunjukkan pengamatan ke-i berada jauh dari pusat semua pengamatan variabel x. Suatu nilai h_{ii} dianggap besar apabila nilainya lebih besar dari 2p/n dan dapat berpotensi sebagai pengamatan yang berpengaruh.

Regresi *Ridge (Ridge Regression)*

Regresi *ridge* merupakan salah satu metode yang dianjurkan untuk memper-baiki masalah multikolinearitas dengan cara memodifikasi metode kuadrat terkecil, sehingga dihasilkan penduga koefisien regresi lain yang bias (Neter, *et al.,* 1990). Modifikasi metode kuadrat terkecil tersebut dilakukan dengan cara menambah tetapan bias, k, yang relatif kecil pada diagonal matriks **X'X**, sehingga penduga koefisien regresi dipengaruhi oleh besarnya tetapan bias, k. Pada umumnya nilai k berkisar antara 0 dan 1. Untuk menentukan penduga ridge, dimulai dari asumsi model linier secara umum, yaitu :

$$y = \mathbf{X}\beta + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

dimana : y adalah vektor pengamatan pada variabel respon yang berukuran (nx1)
X adalah matrik yang berukuran nx(p+1) dari p variabel bebas
β adalah vektor berukuran (p+1)x1 dari koefisien regresi
ε adalah vektor berukuran (nx1) dari error

Dalam regresi *ridge* variabel bebas x dan variabel tak bebas y ditransformasikan dalam bentuk variabel baku Z dan y*, dimana transformasi variabel bebas dan tak bebas ke bentuk variabel baku diperoleh dari $Z=\frac{x-\bar{x}}{s_x}$ dan $y^*=\frac{y-\bar{y}}{s_y}$.

Selanjutnya $Z'Z=\left(\frac{x-\bar{x}}{s_x}\right).\left(\frac{x-\bar{x}}{s_x}\right)$ dan $Z'y=\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$. Sementara itu rumus dari korelasi $r_{xx}=\frac{(x-\bar{x})(x-\bar{x})}{s_x s_x}$. Sehingga persamaan normal kuadrat terkecil (X'X)*b*=X'y akan berbentuk $(r_{xx})b=r_{xy}$, dengan r_{xx} adalah matrik korelasi variabel x dan r_{xy} adalah vektor korelasi antara y dan masing-masing variabel x. Akibat dari transformasi matrik **X** ke **Z** dan vektor y ke y*, maka akan menjadikan persamaan normal regresi *ridge* berbentuk : $(r_{xx}+kI)\hat{b}^*=r_{xy.}$ Penduga koefisien regresi *ridge* menjadi :

$$\hat{b}^*=(r_{xx}+k\,\mathbf{I})^{-1}\ r_{xy} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

dimana : $\hat{b}^*$ adalah vektor koefisien regresi *ridge*
r_{xx} adalah matrik korelasi variabel x yang berukuran pxp
r_{xy} adalah vektor korelasi antara variabel x dan y berukuran px1
k adalah tetapan bias
I adalah matrik identitas berukuran pxp.

Masalah yang dihadapi dalam regresi *ridge* adalah penentuan nilai dari *k*. Prosedur yang cukup baik untuk menentukan nilai *k* ini adalah dengan menggunakan nilai statistik C_p-*Mallows*, yaitu C_k. Statistik C_p-*Mallows* adalah suatu kriteria yang berkaitan dengan rata-rata kuadrat error (*mean square error*) dari nilai kesesuaian model. Nilai *k* yang terpilih adalah yang meminimumkan nilai C_k (Myers, 1990). Nilai C_k dapat dirumuskan sebagai berikut :

$$C_k = \frac{JKR_k}{\hat{\sigma}^2} - n + 2 + 2\,tr[H_k] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

dimana : JKR_k adalah jumlah kuadrat residual dari regresi *ridge*
n adalah banyak pengamatan

$\mathbf{H}_k = [Z(Z'Z+kI)^{-1}Z']$ dengan I adalah matrik identitas

tr $[H_k]$ adalah teras dari matrik $\mathbf{H}_k$

$\hat{\sigma}^2$ adalah penduga varian metode kuadrat terkecil

Acuan lain yang digunakan untuk memilih nilai k adalah dengan melihat nilai *VIF* (Myers, 1990). Nilai *VIF* untuk koefisien regresi *ridge* $\hat{b}^*$ didefinisikan sebagai diagonal dari matrik $(r_{xx}+kI)^{-1}r_{xx}(r_{xx}+kI)^{-1}$. Rumusan ini didapat dengan serangkaian proses sebagai berikut :

Jika di metode kuadrat terkecil diketahui nilai koefisien penduga $\hat{b}$ dan varian($\hat{b}$):

$$\hat{b} = (X'X)^{-1} X'y \text{ dengan } y=X\hat{b}$$
$$\text{Varian}(\hat{b}) = \sigma^2 (X'X)^{-1}$$

Dalam regresi *ridge* harga $\hat{b}^*$ dan varian($\hat{b}^*$) diketahui sebagai :

$$\hat{b}^* = (X'X+kI)^{-1} X'y$$
$$= (X'X+kI)^{-1} X' Xb$$
$$\text{Varian}(\hat{b}^*) = \sigma^2 (X'X+kI)^{-1} (X'X) (X'X)^{-1} (X'X) (X'X+kI)^{-1}$$
$$= \sigma^2 (X'X+kI)^{-1} (X'X) (X'X+kI)^{-1}$$

Sehingga *VIF* merupakan diagonal matrik $(X'X+kI)^{-1} (X'X) (X'X+kI)^{-1}$. Bila x dibakukan, maka *VIF* dari regresi *ridge* adalah diagonal dari matrik $(r_{xx}+kI)^{-1}r_{xx}(r_{xx}+k\ I)^{-1}$.

Leverage dalam Regresi *Ridge*

Ketika teknik bias digunakan pada regresi *ridge*, untuk mengurangi efek dari multikolinearitas, maka rumus pencilan di dalam data tersebut dapat dimodifikasi. Seperti halnya di dalam metode regresi kuadrat terkecil, maka pencilan dalam regresi *ridge* dapat diukur dengan nilai *leverage* (h_{ii}). Untuk itu nilai h_{ii} pada regresi kuadrat terkecil berubah sebagai fungsi dari k, guna mendapatkan nilai h_{ii} pada regresi *ridge* (Retnaningsih,2001).

Dengan memakai penduga (2.2), maka nilai-nilai vektor dugaan y adalah :

$$\hat{y}_i^* = Z\ b^*$$
$$= Z(Z'Z+kI^*)^{-1}Z'y$$

Oleh karena itu, matrik **H** untuk regresi *ridge* menjadi $\mathbf{H}^*=Z(Z'Z+kI^*)^{-1}Z'$, dan unsur ke-i pada diagonal utama matrik $\mathbf{H}^*$ adalah $h_{ii}^* = z_i (Z'Z+kI^*)^{-1} z_i'$. Matrik $\mathbf{H}^*$ berperan sama seperti matrik **H** pada metode kuadrat terkecil. Sehingga, nilai dugaan ke-i dapat ditulis dalam bentuk elemen $\mathbf{H}^*$ sebagai berikut (Walker dan Birch, 1988):

$$\hat{y}_i^* = \sum_{j=1}^{n} h_{ij}^* y_j$$

Unsur diagonal matrik topi *ridge* h_{ii}^* dapat diinterpretasikan sama sebagai *leverage* pada diagonal matrik topi pada metode kuadrat terkecil.

Lichtenstein dan Velleman (1983) dalam Walker dan Birch (1988) mengungkapkan beberapa fakta penting dari sifat unsur diagonal matrik H*. Pertama, untuk k>0, maka nilai $h_{ii}^* < h_{ii}$ dengan i = 1, 2, …, n. Dengan demikian, untuk setiap pengamatan, nilai *leverage* regresi *ridge* lebih kecil dari *leverage* regresi kuadrat terkecil. Kedua, *leverage* menurun secara monoton sejalan dengan kenaikan k. Ketiga, laju penurunan *leverage* tergantung pada posisi baris tertentu dari Z sepanjang sumbu utama. Artinya, *leverage* dari baris yang terletak di sumbu utama yang berpadanan dengan akar karakteristik besar akan berkurang lebih sedikit dari pada *leverage* dari baris yang terletak di sumbu utama yang berpadanan dengan akar karakteristik kecil.

3. Pembahasan

Penurunan Rumus dari Regresi *Ridge*

Jika $\hat{\beta}^*$ adalah penduga dari vektor β, maka jumlah kuadrat residual dapat ditulis (Hoerl dan Kennard, 1970) sebagai berikut :

$$\begin{aligned}\phi &= (\text{Y-X}\hat{\beta}^*)'(\text{Y-X}\hat{\beta}^*)\\ &= (\text{Y-X}\hat{\beta})'(\text{Y-X}\hat{\beta})+(\hat{\beta}-\hat{\beta}^*)'\text{X'X}(\hat{\beta}-\hat{\beta}^*)\\ &= \phi_{\min}+\phi(\hat{\beta}^*)\end{aligned}$$

dimana $\hat{\beta}$ adalah penduga kuadrat terkecil dari β. Untuk ϕ tetap, maka dipilih nilai $\hat{\beta}^*$ dan dibuat meminimumkan $\hat{\beta}^{*\prime}\hat{\beta}^*$ dengan kendala $(\hat{\beta}-\hat{\beta}^*)'\text{X'X}(\hat{\beta}-\hat{\beta}^*)=\phi_0$, sehingga masalah *Lagrange* (Hoerl dan Kennard, 1970) menjadi :

$$F = \hat{\beta}^{*\prime}\hat{\beta}^*+\frac{1}{k}[(\hat{\beta}^*-\hat{\beta})'\text{X'X}(\hat{\beta}^*-\hat{\beta})-\phi_0]$$

$$\frac{\partial F}{\partial \hat{\beta}^*} = 2\hat{\beta}^*+\frac{1}{k}[2(\text{X'X})\ \hat{\beta}^*-2(\text{X'X})\ \hat{\beta}] = 0$$

$$\hat{\beta}^*\ [1+\frac{1}{k}(\text{X'X})] - \frac{1}{k}\ (\text{X'X})\ \hat{\beta}]$$

$$\hat{\beta}^* = [k\text{I}+(\text{X'X})]^{-1}\ (\text{X'X})\ \hat{\beta}$$

Jadi penduga regresi *ridge* adalah : $\hat{\beta}^* = [(\text{X'X}+k\text{I})]^{-1}\ \text{X'Y}$

Adapun sifat-sifat Regresi *Ridge* sebagai berikut (Marquardt, 1970):

1. Penduga $\hat{\beta}^*$ adalah transformasi linier dari $\hat{\beta}$, dan transformasi hanya tergantung pada X dan k.

   $\hat{\beta}^* = [k\text{I}+(\text{X'X})]^{-1}\ \text{X'Y}$, tetapi $\text{X'Y} = (\text{X'X})\ \hat{\beta}$

   Maka, $\hat{\beta}^* = [(\text{X'X}+k\text{I})]^{-1}\ (\text{X'X})\ \hat{\beta} = \text{Z}_k\hat{\beta}$

   $\text{E}(\hat{\beta}^*) = \text{Z}_k\hat{\beta}$

   Sehingga $\hat{\beta}^*$ adalah penduga bias dari $\hat{\beta}$

2. Varian $\hat{\beta}^*$ adalah :

   $\text{V}(\hat{\beta}^*) = \sigma^2[k\text{I}+(\text{X'X})]^{-1}\ (\text{X'X})\ [k\text{I}+(\text{X'X})]^{-1}$

   $\text{V}(\hat{\beta}^*) = \text{var}\ [k\text{I}+(\text{X'X})]^{-1}\ \text{X'Y} = [k\text{I}+(\text{X'X})]^{-1}\ \text{X'}\ \sigma^2\text{IX}[k\text{I}+(\text{X'X})]^{-1}$

   $\text{V}(\hat{\beta}^*) = \sigma^2[k\text{I}+(\text{X'X})]^{-1}\ (\text{X'X})\ [k\text{I}+(\text{X'X})]^{-1}$

3. *MSE (Mean Square Error)* dari $\hat{\beta}^*$: $\text{E}(\text{L}^2) = \text{Tr}[\text{V}(\hat{\beta}^*)]+\hat{\beta}'(\text{Z}_k\text{-I})'(\text{Z}_k\text{-I})\ \hat{\beta}$

   $= \text{Varian} + (\text{bias})^2$

   $\text{E}(\text{L}^2) = \text{E}[(\hat{\beta}^*-\hat{\beta})'(\hat{\beta}^*-\hat{\beta})]$

   $$= \sigma^2\sum_{j=1}^{p}\frac{\lambda_i}{(\lambda_j+k)^2}+k^2\beta'(\text{X'X}+k\text{I})^{-2}\ \beta\ =\ \text{V}(\hat{\beta}^*)+k^2\beta'(\text{X'X}+k\text{I})^{-2}\ \beta$$

4. Jika $\text{k}\geq 0$, dan misal $\hat{\beta}^*$ memenuhi persamaan $\hat{\beta}^* = [(\text{X'X}+k\text{I})]^{-1}\ \text{X'Y}$, maka $\hat{\beta}^*$ meminimumkan jumlah kuadrat residual : $\Phi(\hat{\beta}^*) = (Y-X\hat{\beta}^*)'(Y-X\hat{\beta}^*)$.

5. Jika $\hat{\beta}^*$ adalah solusi dari $[k\text{I}+(\text{X'X})]\ \hat{\beta}^* = \text{X'Y}$ untuk nilai k yang diberikan, maka $\|\hat{\beta}^*\|$ adalah fungsi monoton turun kontinyu dari k, sedemikian hingga pada saat $k \to \infty$, $\|\hat{\beta}^*\| \to 0$.

6. Jika $\beta'\beta$ terbatas, maka ada tetapan k>0 sedemikian hingga *MSE* dari $\hat{\beta}*$ kurang dari *MSE* penduga kuadrat terkecil
7. Dalam persamaan (X'X+*k*I) $\hat{\beta}*$= X'Y, g = X'Y adalah vektor gradien dari $\Phi(\beta)$. Misal γ_k adalah sudut antara $\hat{\beta}*$ dan g, maka γ_k adalah fungsi monoton turun kontinyu dari k, sedemikian hingga k $\to \infty, \gamma_k \to 0$.

Pemilihan nilai tetapan bias k merupakan sesuatu yang tidak dipisahkan dalam regresi *ridge*. Untuk itu perlu dirunut dari mana asal nilai tetapan bias k tersebut. Untuk melihat $\hat{\beta}*$ dari sudut pandang *MSE*, maka Hoerl dan Kennard (1970) meng-ekspresikan hal tersebut ke dalam bentuk E(L^2), dimana :

$$E(L^2) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_i}{(\lambda_j + k)^2} + k^2 \beta'(X'X+kI)^{-2} \beta$$

$$= \gamma_{(1)}(k) + \gamma_2(k)$$

Elemen kedua, yaitu $\gamma_2(k)$, adalah jarak kuadrat dari Zβ ke β. Elemen $\gamma_2(k)$ akan bernilai nol, jika *k*=0, sehingga $\gamma_2(k)$ dapat dipandang sebagai bias kuadrat. Elemen pertama, yaitu $\gamma_1(k)$, merupakan total varian dari dugaan parameter. Total varian dari semua $\hat{\beta}*_j$ adalah jumlah diagonal elemen $\sigma^2 Z(X'X)^{-1} Z'$. Total varian turun seiring dengan kenaikan *k*, sementara bias kuadrat naik seiring dengan kenaikan *k*.

Total varian $\gamma_1(k)$ adalah kontinyu, merupakan fungsi monoton turun dari *k*.

$$\frac{d\gamma_1}{dk} = \sum -2\sigma^2 \lambda_j (\lambda_j + k)^{-3}.1$$

$$= -2\ \sigma^2 \sum \frac{\lambda_j}{(\lambda_j + k)^3}$$

Bias kuadrat $\gamma_2(k)$ adalah kontinyu, merupakan fungsi monoton naik dari k.

$$\frac{d\gamma_2}{dk} = \sum \frac{2\alpha_j^2 k(\lambda_j + k)^2 - \alpha_j^2 k^2 (2(\lambda_j + k))}{(\lambda_j + k)^4}$$

$$= \sum \frac{2\alpha_j^2 k\lambda_j + 2\alpha_j^2 k^2 - 2\alpha_j^2 k^2}{(\lambda_j + k)^3} \qquad = \sum \frac{2\alpha_j^2 k\lambda_j}{(\lambda_j + k)^3}$$

$$\frac{d\gamma_1}{dk} + \frac{d\gamma_2}{dk} = \sum_{j=1}^{p} 2\left[k \frac{\lambda_j\ \alpha_j^2}{(\lambda_j + k)^3} - \sigma^2 \frac{\lambda_j}{(\lambda_j + k)^3} \right] = 0$$

$$k\lambda_j \alpha_j^2 - \sigma^2 \lambda_j = 0$$

$$k = \frac{\sigma^2}{\alpha_j^2}$$

Sedangkan untuk rumusan C_k yang digunakan sebagai alternatif pemilihan nilai tetapan bias k dapat diturunkan sebagai berikut (Myers, 1990):

$$\sum_{i=1}^{n} Var\hat{y}_i^* + \sum_{i=1}^{n} [Bias\hat{y}_i^*]^2$$

$$\frac{\sum_{i=1}^{n} Var\hat{y}_i^*}{\sigma^2} = tr[A_k]^2$$

$$\sum_{i=1}^{n}(Bias\hat{y}_i^*)^2 = JKR_k - \sigma^2 tr(I - A_k)^2$$

$$\frac{\sum_{i=1}^{n}(Bias\hat{y}_i^*)^2}{\sigma^2} = \frac{JKR_k - \sigma^2 tr(I - A_k)^2}{\sigma^2}$$

$$= \frac{JKR_k}{\sigma^2} - tr(I - A_k)^2$$

Jadi penduga dari $\sum_{i=1}^{n} Var\hat{y}_i^* + \sum_{i=1}^{n}[Bias\hat{y}_i^*]^2$ diberikan oleh :

$$C_k = E[\hat{y}_i - E(y_i)]^2$$

$$= [E(y_i) - E(\hat{y}_i)]^2 + V(\hat{y}_i)$$

$$C_k = \frac{\sum_{i=1}^{n} Var\hat{y}_i^* + \sum_{i=1}^{n}[Bias\hat{y}_i^*]^2}{\sigma^2}$$

$$= tr(A_k)^2 + \frac{JKR_k}{\sigma^2} - tr(I - A_k)^2$$

$$= \frac{JKR_k}{\sigma^2} - n + 2tr(A_k)$$

Bila $H_k = X^*(X^{*\prime}X^* + kI)^{-1}X^{*\prime}$

dan $tr(A_k) = tr(H_k) + 1$

$$C_k = \frac{JKR_k}{\hat{\sigma}^2} - n + 2[tr(H_k) + 1]$$

$$C_k = \frac{JKR_k}{\hat{\sigma}^2} - n + 2 + 2tr(H_k)$$

Identifikasi Pencilan

Leverage (h_{ii}) adalah elemen-elemen diagonal dari matrik proyeksi *least squares* yang disebut matrik topi, **H** = X(X'X)$^{-1}$X', yang menjelaskan pendugaan atau nilai-nilai dugaan, karena : $\hat{y} \equiv Xb = Hy$. Elemen-elemen diagonal H merupakan jarak antara x_i dan $\bar{x}$. Oleh karena H adalah matrik proyeksi, maka dia simetris dan idempoten (H^2=H). Elemen-elemen dari matrik topi yang dipusatkan adalah :

$\tilde{h}_{ij} = h_{ij} - (1/n)$. Hal ini berimplikasi, bahwa $(1/n) \le h_i \le 1$. Jumlah akar karakteristik dari matrik proyeksi tidak nol sama dengan rank dari matrik. Dalam hal ini rank(**H**) = rank(**X**)=p dan trace **H**=p, atau karena **X** rank penuh, maka $\sum_{i=1}^{n} h_i$ = p.

Ukuran rata-rata elemen diagonal adalah p/n. Data yang diinginkan adalah yang jauh dari pengamatan berpengaruh, dimana masing-masing pengamatan mempunyai h_i dekat dengan rata-rata p/n. Untuk itu perlu beberapa kriteria untuk memutuskan kapan nilai h_i cukup besar atau cukup jauh dari rata-ratanya. Jika variabel-variabel bebas didistribusikan secara independen, maka dapat dicari distribusi eksak dari fungsi-fungsi tertentu dari h_i. Belsley, Kuh, dan Welsch (1980) mengambil teori distribusi untuk mencari batas kritis dari nilai *leverage* sebagai berikut :

Statistik Λ Wilks untuk dua grup, dimana 1 grup terdiri dari titik tunggal :

$$\Lambda(\tilde{x}_i) = \frac{\det(\tilde{X}'\tilde{X} - (n-1)\tilde{x}'(i)\tilde{x}(i) - \tilde{x}_i'\tilde{x}_i)}{\det(\tilde{X}'X)}$$

$$= \frac{1-\frac{n}{n-1}\tilde{x}_i(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i')\det(\tilde{X}'\tilde{X})}{\det(\tilde{X}'\tilde{X})} = 1-\frac{n}{n-1}\tilde{h}_i$$

$$= 1-\frac{n}{n-1}(h_i-\frac{1}{n}) = \frac{n}{n-1}(1-h_i)$$

$$\frac{n-p}{p-1}\left[\frac{1-\Lambda(\tilde{x}_i)}{\Lambda(\tilde{x}_i)}\right] \sim F_{p-1,n-p}$$

Sehingga dapat ditulis : $\frac{n-p}{p-1}\left[\frac{h_i-(\frac{1}{n})}{(1-h_i)}\right] \sim F_{p-1,n-p}$

Untuk p besar (lebih dari 10) dan n-p besar (lebih dari 50), maka pada tabel F nilai-nilainya kurang dari 2 sehingga nilai 2(p/n) merupakan batas yang cukup bagus. Selanjutnya, pengamatan ke-i adalah titik *leverage* ketika h_i melebihi 2(p/n).

Penurunan Rumus *PRESS*

Rumus *PRESS* umumnya adalah $y_i - \hat{y}_{i,-i}$. Rumus ini bisa ditulis sebagai berikut (Myers, 1990) : $e_{i,-i} = y_i - x_i'b_{-i}$

$$e_{i,-i} = y_i - x_i'\left[(X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_{ii}}\right]X_{-i}'y_{-i}$$

$$= y_i - x_i'(X'X)^{-1}X_{-i}'y_{-i} - \frac{h_{ii}x_i'(X'X)^{-1}X_{-i}'y_{-i}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i(1-h_{ii})x_i'(X'X)^{-1}X_{-i}'y_{-i} - h_{ii}x_i'(X'X)^{-1}X_{-i}'y_{-i}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i'(X'X)^{-1}X_{-i}'y_{-i}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i'(X'X)^{-1}(X'y - x_iy_i)}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - \hat{y}_i + h_{ii}y_i}{1-h_{ii}} = \frac{y_i-\hat{y}}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}$$

4. Kesimpulan

Masalah pencilan ini dapat diatasi degan berbagai metode, diantaranya metode regresi *ridge (ridge regression)*. Hal ini ditinjau dari ketepatan model, dimana metode regresi *ridge* memberikan hasil yang relatif lebih baik dibandingkan dengan metode kuadrat terkecil.

Daftar Pustaka

Belsley, D.A., Kuh, E., dan Welsch, R.E, (1980), *Regression Diagnostics.* John Wiley. New York.

Hoerl, A.E dan Kennard, R.W., (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* Vol. 12, no. 1.

Marquardt, D.W., (1970), Generalized Inverses, Ridge Regression, Biased LinearEstimation, and Nonlinier Estimation. *Technometrics,* Vol. 12, no. 3.

Mason, R.L. dan Gunst, R.F., (1985), Outlier-Induced Collinearities. *Technometrics,* Vol. 2, no. 4.

Myers, R.H, (1990), *Classical and Modern Regression with Applications.* 2nd Edition. PWS-KENT, Boston.

Neter, J., Wasserman, W. dan Kutner, M.H., (1990), *Applied Linear Statistical Models, Regression, Analysis of Variance & Experimental Design*, Richard D. Irwin Inc. Illinois. Toppan Company. LTD, Tokyo.

Retnaninsih, E., (2001), Studi Perbandingan Metode Regresi Ridge dengan Kuadrat Trekecil Parsial Pada Struktur Ekonomi dan Tingkat Kesra Penduduk Indonesia. Tidak Dipublikasikan, *Tesis Program Master*, ITS, Surabaya.

Walker, E. dan Birch, J.B., (1988). Influence Measure in Ridge Regression. *Technometrics,* 25: 221-227.