JURNAL MATEMATIKA DAN KOMPUTER
Vol. 5. No. 3, 161 - 167, Desember 2002, ISSN : 1410-8518

PENGARUH SUATU DATA OBSERVASI DALAM MENGESTIMASI PARAMETER MODEL REGRESI

Herni Utami, Ruri I, dan Abdurakhman

Jurusan Matematika FMIPA UGM

Abstrak

Observasi yang mempengaruhi model regresi sedemikian hingga elipsoid konfidensi untuk estimasi parameter regresinya menjadi kecil apabila observasi tersebut "dihilangkan" adalah observasi penting. Sehingga observasi penting tersebut bisa merupakan observasi berpengaruh sesungguhnya atau bisa juga sebagai outlier. Salah satu cara menentukan observasi ke-i penting atau tidak, melihat elipsoid konfidensi parameter model regresi linear dengan "menghilangkan" observasi tersebut.

Kata kunci : elipsoid konfidensi

1. PENDAHULUAN

Berbicara mengenai regresi secara umum, berarti membicarakan proses bagaimana kita menghubungkan antara variabel eksplanatori (independen) dengan variabel respon (dependen) dari suatu himpunan data (*data set*) dengan harapan diperoleh suatu model yang sesuai untuk bentuk hubungan variabel-variabel tadi. Setelah diperoleh model yang sesuai, muncul suatu pertanyaan mengenai apakah suatu observasi penting yang mempengaruhi model tersebut? Pengertian observasi penting di sini adalah observasi yang mempengaruhi model regresi sedemikian hingga elipsoid konfidensi untuk estimasi parameter regresinya menjadi kecil apabila observasi tersebut "dihilangkan". Sehingga observasi penting tersebut bisa merupakan observasi berpengaruh sesungguhnya atau bisa juga sebagai outlier.

Sebelum melakukan analisa regresi ganda, uji yang biasa dilakukan adalah melihat ada atau tidaknya kejanggalan (outlier atau gap) pada distribusi univariat setiap variatnya dengan menggunakan plot diagram scatter, meskipun dengan cara ini tidak dapat mendeteksi observasi multivariat yang tidak sesuai.

Setelah terbentuk model regresinya, maka kebanyakan prosedur deteksi yang digunakan terfokus pada residual, nilai prediksi (*fitted value*, $\hat{y}$), dan variabel eksplanatori. *Studentized residual*, t_i, banyak direkomendasikan sebagai alat deteksi adanya outlier. Behnken dan Draper (1972) menggambarkan estimasi variansi $\hat{y}$ (ekuivalen dengan estimasi variansi residual, $\hat{V}(R_i)$) dengan plot residual atau studentized residual memberikan informasi lebih. Lebih spesifik mereka mengatakan *"Suatu variasi yang luas di dalam variansi residual mencerminkan suatu keanehan dari matriks* $\mathbf{X}$*, yaitu suatu jarak yang tak homogen dari observasi-observasi dan akan menunjukkan data yang defisiensi"*. Huber (1975) juga menyatakan bahwa variansi-variansi tersebut memiliki informasi lebih. Setelah observasi-observasi penting terdeteksi menggunakan ukuran-ukuran di atas, dengan menguji efek "penghapusan" (*deleting*) observasi tertentu merupakan satu langkah lebih lanjut.

Tujuan dari penulisan makalah ini, adalah untuk menunjukkan salah satu metode untuk mendeteksi observasi berpengaruh dalam model regresi linear.

2. MODEL REGRESI LINEAR

Jika variabel bebas dinotasikan dengan $\mathbf{X}$ dan variabel tidak bebas dinotasikan dengan $\mathbf{Y}$, maka model regresi linear dinyatakan dalam bentuk:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

dimana:

$\mathbf{Y}$: vektor observasi dengan order Nx1

$\boldsymbol{\beta}$: vektor parameter dengan order px1

$\mathbf{X}$: matriks yang elemen-elemennya diketahui, dengan order Nxp

$\boldsymbol{\varepsilon}$: vektor kesalahan dengan order Nx1, yang setiap elemennya diasumsikan berdistribusi normal independen dengan mean nol dan variansinya σ^2.

Dalam makalah ini, diasumsikan model diatas adalah model linear rank penuh, sehingga rank dari $\mathbf{X}$ adalah p.

JURNAL MATEMATIKA DAN KOMPUTER
Vol. 5. No. 3, 161 - 167, Desember 2002, ISSN : 1410-8518

Karena $\boldsymbol{\beta}$ belum diketahui, sehingga diestimasi dari data. Salah satu cara untuk mencari estimasi vektor $\boldsymbol{\beta}$ dengan menggunakan metode kuadrat terkecil (*least square*) yang diperoleh dengan jalan meminimumkan jumlah kuadrat kesalahannya.

Dari persamaan (1) dapat ditulis kembali menjadi

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

sehingga jumlah kuadrat kesalahannya adalah

$$\boldsymbol{\varepsilon}'\,\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

$$= \mathbf{X}'\mathbf{Y} - 2\boldsymbol{\beta}\boldsymbol{\beta}\,\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\,\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Untuk $\frac{d(\boldsymbol{\varepsilon}'\,\boldsymbol{\varepsilon})}{d\boldsymbol{\beta}} = 0$, maka $-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \quad \text{…….....…(2)}$$

Dari persamaan (2) diperoleh estimasi $\boldsymbol{\beta}$, yaitu:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \text{……(3)}$$

Berdasarkan persamaan (3) yang merupakan penyelesaian dari persamaan normal (2), maka:

1. vektor residu model (1):

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \text{….....…(4)}$$

2. covariansi dari $\hat{\mathbf{Y}}$:

$$V(\hat{\mathbf{Y}}) = V(\mathbf{X}'\hat{\beta})$$

$$= V(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 \quad \text{……..……(5)}$$

3. covariansi dari $\mathbf{R}$:

$$V(\mathbf{R}) = V((\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y})$$

$$= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2 \quad \text{…….……..(6)}$$

4. ellipsoid konfidensi $(1-\alpha)100\%$ untuk $\boldsymbol{\beta}$, jika diketahui himpunan semua vektor β^*
   (estimasi $\boldsymbol{\beta}$ yang lain) adalah

$$\frac{(\boldsymbol{\beta}^{*}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}^{*}-\hat{\boldsymbol{\beta}})}{ps^{2}} \le F(p,n-p,1-\alpha)$$

dengan $s^2 = \mathbf{R}'\mathbf{R}/(n-p)$ dan $F(p,n-p,1-\alpha)$ adalah persentil $1-\alpha$ distribusi F dengan derajat bebas p dan n-p.

3. MENDITEKSI PENGARUH SEBUAH DATA OBSERVASI DALAM MENGESTIMASI PARAMETER MODEL

Untuk menentukan derajat pengaruh data ke-i dalam memperoleh $\hat{\boldsymbol{\beta}}$, langkah pertama adalah menghitung estimasi $\boldsymbol{\beta}$ dengan menghapus titik tersebut, sehingga diperoleh $\hat{\boldsymbol{\beta}}_{(-i)}$, yaitu estimasi kuadrat terkecil $\boldsymbol{\beta}$ dengan menghapus data ke-i. Selanjutnya dihitung:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(-i)}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(-i)}-\hat{\boldsymbol{\beta}})}{ps^{2}} \text{ ……..…(7)}$$

Dengan menentukan tingkat signifikan α, akan diperoleh batas konfidensi (1-α)100% untuk $\boldsymbol{\beta}$ berdasarkan $\hat{\boldsymbol{\beta}}$.

Untuk menghitung D_i, secara mudah akan ditunjukkan di bawah ini:

$$(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)}) = (\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})^{-1}\mathbf{x}_i(\mathbf{Y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) \text{ ……..…..(8)}$$

dimana $\mathbf{X}_{(-i)}$: matriks yang diperoleh dengan menghilangkan baris ke-i dari matriks $\mathbf{X}$

Y_i : observasi ke-i

$\mathbf{x}_i$: baris ke-i dari matriks $\mathbf{X}$.

Selanjutnya jika $v_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$, dan diasumsikan $v_i < 1$ maka diperoleh

$$(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}) = (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}/(1-v_i)$$

$$(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})\mathbf{x}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i/(1-v_i)$$

$$(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})\mathbf{x}_i = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i(1-v_i) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i)/(1-v_i)$$

$$(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)})\mathbf{x}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i/(1-v_i) \text{ …….. .....(9)}$$

Dengan mensubstitusikan persamaan (9) ke persamaan (8), akan diperoleh:

$$(\hat{\beta} - \hat{\beta}_{(-i)}) = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1-v_i}(Y_i - \mathbf{x}_i'\hat{\beta})$$

sehingga persamaan (7) bisa ditulis:

$$D_i = \left(\frac{Y_i - \mathbf{x}_i'\hat{\beta}}{s\sqrt{1-v_i}}\right)^2 \frac{v_i}{p(1-v_i)} \quad \text{…….(10)}$$

Dari persamaan (10) tampak bahwa D_i bergantung pada 3 hal, yaitu: jumlah parameter (p), $t_i = \left(\frac{Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{s\sqrt{1-v_i}}\right)$, dan rasio antara covariansi nilai prediksi ke-i $\left(V(\hat{Y}_i) = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\sigma^2 = v_i\sigma^2\right)$ dan covariansi residu ke-i $\left(V(R_i) = \sigma^2(1-v_i)\right)$. Dengan demikian persamaan (10) dapat ditulis dalam bentuk yang sederhana menjadi:

$$D_i = \frac{t_i^2}{p}\frac{V(\hat{Y}_i)}{V(R_i)} \quad \text{……...….(11)}$$

Jelas, t_i^2 merupakan ukuran untuk mengetahui bahwa observasi ke-i dapat dikatakan sebagai outlier dari model yang diasumsikan. Sedangkan $V(\hat{Y}_i)/V(R_i)$ merupakan ukuran sensitif relatif estimasi, yaitu $\hat{\boldsymbol{\beta}}$, terhadap data yang potensial terpencil dari himpunan data. Untuk nilai rasio yang besar memberikan indikasi bahwa data yang bersangkutan memberikan bobot yang besar dalam menentukan $\hat{\boldsymbol{\beta}}$. Kombinasi t_i^2 dan $V(\hat{Y}_i)/V(R_i)$ dalam persamaan (11) menghasilkan suatu ukuran pengaruh menyeluruh dari sembarang data dalam menentukan estimasi parameter dengan metode least square error. Dalam suatu analisis, untuk informasi tambahan bisa dilakukan uji t_i dan $V(\hat{Y}_i)/V(R_i)$ secara terpisah.

4. CONTOH KASUS

Longley(1967) memberikan himpunan data yang menghubungkan antara 6 variabel ekonomi dengan total tenaga kerja yang dibutuhkan dari tahun 1947 sampai 1962. Tabel 1 memuat t_i, $V(\hat{Y}_i)/V(R_i)$, D_i, dan tahun. Dari tabel tampak

bahwa D_i terbesar diperoleh untuk tahun 1951. Penghilangan data tahun ini ternyata akan merubah estimasi kuadrat terkecil, yaitu $\hat{\boldsymbol{\beta}}$ ke batas daerah konfidensi 35% untuk $\hat{\boldsymbol{\beta}}$. Sedang D_i terkecil kedua tahun 1962 dan ini penghilangan data tahun ini akan merubah estimasi $\hat{\boldsymbol{\beta}}$ ke batas daerah konfidensi 15%. Jelas, tahun 1951 dan 1962 mempunyai pengaruh yang besar dalam menentukan $\hat{\boldsymbol{\beta}}$.

Tahun	$\lvert t_i \rvert$	$V(\hat{Y}_i)/V(R_i)$	D_i
1947	1.15	0.74	0.14
1948	0.48	1.30	0.04
1949	0.19	0.57	*
1950	1.70	0.59	0.24
1951	1.64	1.60	0.61
1952	1.03	0.59	0.09
1953	0.75	0.97	0.08
1954	0.06	1.02	*
1955	0.07	0.84	*
1956	1.83	0.49	0.23
1957	0.07	0.56	*
1958	0.18	0.93	*
1959	0.64	0.60	0.04
1960	0.32	0.30	*
1961	1.42	0.59	0.17
1962	1.21	2.21	0.47

* : lebih kecil dari 5.10^{-3}

5. KESIMPULAN

Dari data hasil observasi, dapat dideteksi bagaimana pengaruh observasi ke-i dalam memperoleh estimasi parameter model regresi linear. Dengan melihat elipsoid konfidensi parameter model regresi linear, dapat ditentukan apakah jika observasi ke-i dihilangkan akan diperoleh elipsoid konfidensi yang lebih kecil, yang berarti observasi ke-i merupakan observasi penting.

DAFTAR PUSTAKA

1. Dennis C. R, *Detection of Influential Observation in Linear Regression*, Technometrics, 2000, 42 : 65-68.

2. David A. B, Edwin, K, and Roy, E.W, *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, Wiley, New York, 1980.

3. Peter, J. H, *Robust Statistics*, Wiley, New York, 1981.