ARTIKEL PENELITIAN KELOMPOK

DANA RUTIN UNAND 2003

Kontrak No. 107/J.16/PL/RUTIN/V/2003

ANALISIS PERBANDINGAN UJI-UJI PENCILAN PADA ANALISIS REGRESI

Oleh:

Admi Nazra, S.Si. M.Si — **Ketua**

Yuni Silvyani — **Anggota**

Drs. I Made Arnawa, M.Si — **Pembimbing**

Fakultas Matematika dan Ilmu Pengetahuan Alam

Departemen Pendidikan Nasional

Lembaga Penelitian Universitas Andalas

Padang 2003

Dibiayai oleh Dana Rutin UNAND 2003

ANALISIS PERBANDINGAN UJI-UJI PENCILAN PADA ANALISIS REGRESI

Abstrak

Dalam tulisan ini dibahas mengenai uji pencilan dengan simpangan baku, uji pencilan dengan sisa ter-student dan uji pencilan dengan kriteria informasi. Ketiga jenis uji ini belum tentu memberikan kesimpulan yang sama untuk menentukan data mana yang dikategorikan sebagai data pencilan pada suatu masalah regresi linier. Disamping itu uji pencilan dengan simpangan baku dan uji pencilan dengan sisa ter-student lebih sederhana untuk digunakan dibandingkan dengan uji pencilan kriteria informasi. Uji pencilan dengan sisa ter-student memberikan hasil yang relatif akurat dari dua uji lainnya.

PENDAHULUAN

Analisis regresi merupakan metoda statistika yang amat banyak digunakan, tidak hanya di lingkungan para peneliti di bidang matematika atau statistika namun juga di bidang-bidang lain seperti biologi, kimia, pertanian, ekonomi dan lain-lain. Umumnya analisis regresi digunakan dalam rangka mengolah data untuk menentukan hubungan antara dua atau lebih peubah sehingga diperoleh model atau hubungan fungsional antara peubah tersebut. Sehingga dengan model tersebut para peneliti dapat berusaha memahami, menerangkan, mengendalikan dan kemudian memprediksikan kelakuan sistem yang mereka teliti. Secara umum model merupakan penyederhanaan abstraksi dari keadaan alam yang sesungguhnya. Keadaan alam yang ingin diteliti biasanya amat rumit dan kemampuan kita menelitinya secara keseluruhan amat terbatas karena itu kita perlu menyederhanakannya sesuai dengan kemampuan akal kita menghadapinya. Model yang dimaksud disini akan selalu berbentuk fungsi, dan regresi merupakan alat yang ampuh dalam pembentukannya.

Model regresi secara umum dapat ditulis sebagai $\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{e}$ dengan $\boldsymbol{Y}$ merupakan vektor respon $n\times 1$, $\boldsymbol{X}$ menyatakan matriks peubah bebas $n\times(k+1)$ (tulis $k+1=p$), β vektor parameter $(k+1)\times 1$ dan $\boldsymbol{e}$ vektor galat $n\times 1$. Metoda yang digunakan untuk menaksir β adalah metoda kuadrat terkecil.

Kebaikan model dapat dilihat dari nilai R^2 (koefisien determinasi) dan pengujian hipotesis terhadap parameter. Sedangkan kococokan model dengan data di lihat dengan mengamati sisa (residual). Salah satu yang dapat dilihat dari sisa adalah pencilan (*outlier*). Pencilan adalah data yang tidak mengikuti pola umum model. Pada beberapa literatur telah dikemukakan beberapa uji untuk memeriksa apakah suatu data dapat dikategorikan sebagai pencilan. Secara kasar dapat diambil patokan yaitu yang sisanya berjarak dua simpangan baku atau lebih dari rata-ratanya dapat dikategorikan sebagai data pencilan (Sembiring,1995). Namum (Weisberg,1985) dan (Gentlemen & Wilk, 1975a-b) mengemukakan suatu cara yang sederhana ialah dengan menggunakan sisa ter-*student* dengan derajat bebas dk=n-p-1. Dimana bila sisa untuk data yang bersangkutan lebih besar dari nilai $t(n-p-1,\alpha)$ dari table-t maka dianggap data tersebut terpencil. Pynnonen (1992) menggunakan kriteria informasi untuk mendeteksi pencilan. Ada dua kriteria yaitu $AIC=n\log(1-R^2)-2\log(n-k)!+2m$ dan $BIC= n\log(1-R^2)-2\log(n-k)!+m\log n$.

Dari beberapa uji pencilan yang disebutkan di atas, maka muncul permasalahan, apakah data yang dikategorikan sebagai pencilan oleh suatu uji tertentu dapat juga terdeteksi sebagai pencilan oleh uji yang lain. Artinya apakah setiap uji-uji tersebut memberikan hasil yang sama dalam menentukan suatu data apakah pencilan atau tidak. Jika memberikan hasil yang sama, uji manakah yang lebih sederhana atau lebih efektif dan efisien untuk digunakan. Hal ini sangat penting, sebab jika dalam penelitian ini nantinya ditemukan bahwa uji pencilan ini tidak memberikan informasi pencilan yang sama untuk berbagai metoda atau uji yang disebutkan di atas, maka tentunya seorang peneliti harus berhati-hati dalam menggunakan atau memilih uji-uji tersebut.

Pendeteksian pencilan ini tidak hanya penting dalam rangka untuk memperbaiki model yang kita cari, namum juga dengan diketahuinya suatu data adalah pencilan maka seorang peneliti dapat menelusuri data tersebut untuk mengetahui dan mempelajari lebih mendalam mengenai data pencilan yang bersangkutan.

TUJUAN DAN MANFAAT PENELITIAN

Tujuan dari penelitian ini adalah:

1. Untuk mencoba berbagai uji pencilan yaitu uji simpangan baku, uji sisa ter-*student*, dan uji kriteria informasi untuk kasus yang sama.
2. Untuk mengetahui apakah data yang dikategorikan sebagai data pencilan oleh uji tertentu, dapat juga terdeteksi sebagai pencilan oleh uji-uji yang lain.
3. Untuk mengetahui uji manakah yang lebih sederhana atau lebih efektif dan efisien untuk digunakan.

Manfaat Penelitian.

Dari hasil penelitian ini seorang peneliti khususnya peneliti yang akan menggunakan regresi dalam pengolahan datanya, dapat mengetahui dan menggunakan uji-uji pencilan yang lebih cocok dan tepat. Jadi hasil ini akan memberikan kontribusi penting bukan hanya untuk para peneliti namum juga bagi pengembangan konsep analisis regresi khususnya analisa data pencilan.

METODA PENELITIAN

Metoda yang digunakan dalam penelitian ini adalah metoda simulasi dengan menggunakan program komputer (SAS, MINITAB, MATLAB), dan komputer yang digunakan adalah yang berkecepatan tinggi. Langkah-langkah yang dilakukan dalam penelitian ini adalah:

1. Membangkitkan data dari populasi yang berdistribusi normal dan beberapa data yang bukan berdistribusi normal (yang diduga sebagai pencilan).
2. Melakukan berbagai uji pencilan seperti yang disebutkan di atas.
3. Membandingkan dan menganalisa hasil yang diperoleh.
4. Mencobakan uji-uji ini pada beberapa kasus yang diambil dari beberapa literatur serta dibandingkan dengan hasil simulasi.

TINJAUAN PUSTAKA

Model umum persamaan regresi (Sembiring,1995) dalam bentuk matriks adalah

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\ \underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{e}} \quad ; \quad n \geq p$$

dimana:

1. $\boldsymbol{Y}$ adalah vektor random terobservasi dengan komponennya y_t yaitu variabel random terobsevasi (peubah tak bebas pada pengamatan ke-t t=1,2,...,n)..
2. $\boldsymbol{X}$ matriks disain yang nilainya diketahui dan rank $(\boldsymbol{X}) = k$ dengan komponennya x_{tk} yaitu variabel nonrandom terobservasi yang diketahui nilainya (peubah bebas ke-k k=1,2,...,p pada pengamatan ke-t),
3. $\boldsymbol{\beta}$ vektor parameter dengan komponennya β_k yaitu koefisien regresi(parameter) ke-k yang akan ditaksir,
4. $\boldsymbol{e}$ vektor galat/gangguan dengan komponennya e_t yaitu variabel random yang tak terobservasi (galat/gangguan) dengan : $E(e_t)=0$, $E(e_t^2) = \sigma^2$, $E(e_t e_s) = 0$ $t \neq s$, atau $E(\boldsymbol{e})= 0$, $E(\boldsymbol{ee}') = \sigma^2 \boldsymbol{I}_n$ σ^2 tak diketahui.

Dengan metoda kuadrat terkecil diperoleh $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ dan $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ serta vektor sisa $\boldsymbol{e}=\hat{\boldsymbol{Y}} - \boldsymbol{Y}$.

Seperti yang sudah dijelaskan pada bagian pendahuluan bahwa data yang sisanya berjarak dua simpangan baku dari rata-ratanya dapat dikategorikan sebagai pencilan (Sembiring,1995). Artinya jika sisa data ke-i $e_i \geq 2s$ dimana

$$s^2=\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y - \bar{y})^2}=\hat{\sigma}^2$$

Kemudian (Weisberg,1985) dan (Gentlemen & Wilk, 1975a-b) menggunakan sisa ter-*studen* dengan derajat bebas dk=n-p-1 untuk menentukan data pencilan. Dimana bila $e_i^* = \frac{e_i}{s\sqrt{1-h_{ii}}} \geq t(n\text{-}p\text{-}1,\alpha)$ dari table-t maka dianggap data tersebut terpencil. Dimana h_{ii} adalah unsur diagonal ke-i pada matriks $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Pynnonen (1992) menggunakan kriteria informasi untuk mendeteksi pencilan. Ada dua kriteria yaitu AIC=nlog(1-R^2)-2log(n-k)!+2m dan BIC=nlog(1-R^2)-2log(n-k)!+mlog(n) dengan $R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$ dan k=m=jumlah pencilan. Data-data yang dikategorikan sebagai pencilan berdasarkan metoda atau uji ini adalah apabila nilai AIC dan BIC terkecil diantara data-data lain.

HASIL DAN PEMBAHASAN

Untuk melihat perbedaan hasil uji pencilan untuk berbagai uji pencilan yaitu uji simpangan baku, uji sisa terstudent dan uji kriteria informasi maka digunakan tiga ilustrasi berikut ini.

Ilustrasi pertama diambil dari Barnett (1983) dengan model regresi $y=\beta_0+\beta_1 x+\varepsilon$. Datanya sebagai berikut:

(x)	4	5	7	9	11	14	17	20	23	26	30	35
(z)	110	81	90	74	20	30	37	22	38	25	18	9
(y =logz)	4.7	4.4	4.5	4.3	3.0	3.4	3.6	3.1	3.6	3.2	2.9	2.2

Dengan data ini diperoleh model dan hasil-hasil lainnya sebagai berikut:

Y = 4.66 - 0.0645x S = 0.4103 R-Sq = 73.8%

sisa	sisa ter-student
0.298925	0.83080
0.058370	0.15949
0.292684	0.78142
0.225892	0.59228
-0.953487	**-2.46645**
-0.354593	-0.90591
0.048558	0.12361
-0.277888	-0.71097
0.462085	1.19887
0.236805	0.62909
0.166207	0.46383
-0.204558	-0.63082

Dengan data ini kita peroleh nilai AIC dan BIC sebagai berikut:

Jumlah observasi yang dikategorikan sebagai pencilan k=m	Data Pencilan	R^2	AIC	BIC
0	-	0.7379	-56.04	-56.04
1	(11;3)	0.8974	**-60.33**	**-59.84**
2	(11;3)dan (14;3,4)	0.9338	-58.41	-57.82
3	(11;3)(14;3,4)dan (20;3,1)	0.9606	-58.41	-56.96

Dari tabel diatas nilai AIC dan BIC yang terkecil adalah data (x;y)=(11;3).

Dari tabel sisa, terlihat bahwa harga mutlak sisa yang lebih besar dari 2s=0.8206 adalah sisa untuk data (x;y)=(11;3) yaitu -0.953487. Untuk uji ter-student dimana data dengan harga mutlak sisa lebih besar dari nilai tabel t(9;0,05)=1,833 adalah data (x;y)=(11;3)dengan sisa -2.46645.

Ketiga jenis uji meberikan hasil yang sama dimana data pencilan adalah data (x;y)=(11;3). Grafik 1 pada lampiran memperlihatkan juga bahwa data (x;y)=(11;3) terlihat memencil dari data-data lainnya.

Sebagai ilustrasi kedua untuk membandingkan ketiga jenis uji pencilan digunakan data dari Chatterjee S. and B. Price (1977)

Datanya adalah sebagai berikut:

X	Y
2.5	3.80
2.7	4.10
2.9	5.80
3.1	4.80
3.3	5.70
3.5	4.40
3.7	4.80
3.9	3.60
4.1	5.50
4.3	4.15
4.5	5.80
4.7	3.80
4.9	4.75
5.1	3.90
5.3	6.20
5.5	4.35
5.7	4.15
5.9	4.85
6.1	6.20
6.3	3.80
6.5	7.00
6.7	5.40
6.9	6.10
7.1	6.50
7.3	6.10
7.5	4.75
2.5	1.00
2.7	1.20
7.3	9.50
7.5	9.00

Grafik data ini dapat dilihat pada lampiran (grafik 2).

Dengan menggunakan model $y=\beta_0+\beta_1 x+\varepsilon$ diperoleh hasil-hasil sebagai berikut.

```
Y = 1.71 + 0.665 X     S = 1.402        R-Sq = 38.6%

Sisa
ter-student   sisa
0.32518       0.43006
0.44844       0.59699
1.61582       2.16392
0.76564       1.03085
1.32892       1.79778
0.26846       0.36471
0.46326       0.63163
-0.51287      -0.70144
0.77705       1.06549
-0.30391      -0.41758
0.79888       1.09935
-0.75043      -1.03373
-0.15730      -0.21680
```

-0.87060	-1.19987
0.70203	0.96706
-0.73832	-1.01601
-0.98186	-1.34508
-0.57042	-0.78216
0.31789	0.43477
-1.53896	-2.09830
0.71302	0.96863
-0.56508	-0.76444
-0.14670	-0.19752
0.09183	0.06941
-0.34828	-0.46366
-1.47195	-1.94673
-1.79194	-2.36994
-1.72994	-2.30301
2.20567	**2.93634**
1.74153	2.30927

Nilai AIC dan BIC yang diperoleh dari data ini adalah sebagai berikut:

Jumlah observasi yang dikategorikan sebagai pencilan k=m	Data Pencilan	R^2	AIC	BIC
0	-	0.396	129,58	129,59
1	(7,3;9,5)	0.355	**-153.67**	**-81.01**
2	(2,7;1,2) (7,3;9,5)	0.309	-142.87	2.44
3	(2,5;1) (2,7;1,2) (7,3;9,5)	0.239	-131.31	86.67
4	(2,5;1) (2,7;1,2) (7,3;9,5) (7,5;9)	0.161	-119.79	170.84

Nilai AIC dan BIC yang terkecil terjadi jika data pencilannya adalah (x;y)=(7,3;9,5) yaitu observasi 29. Hasil ini juga bersesuai dengan uji simpangan baku dimana data dengan sisa lebih besar dari 2s=2,804 adalah data ke-29 juga. Namun kedua uji diatas memberikan hasil yang berbeda dengan uji sisa ter-student. Dimana dengan uji ini data pencilan adalah data dengan sisanya lebih besar dari t(27;0,05)=1,703 yaitu data observasi ke-27, 28, 29 dan 30.

Berikut ditampilkan data simulasi y=5+2x dengan data x adalah 2.0 6.0 3.0 8.0 5.0 2.0 4.0 7.0 5.0 2.0 1.0 6.0 8.0 4.0 dan 9.0. Untuk memunculkan data yang diasumsikan sebagai data pencilan maka ditambah pasangan data (x;y) (5,5;17) dan (2,5;8). Grafik dari data ini dapat dilihat pada gambar 3. di lampiran.

Dengan model $y=\beta_0+\beta_1 x+\varepsilon$ diperoleh hasil-hasil sebagai berikut.

y = 4.68 + 2.06 x S = 0.5570 R-Sq = 98.8%

sisa	sisa ters-student
0.20863	0.40316
-0.01282	-0.02396
0.15327	0.28844
-0.12365	-0.24412
0.04254	0.07877
0.20863	0.40316
0.09790	0.18171
-0.06819	-0.13013
0.04254	0.07877
0.20863	0.40316

```
0.26400         0.53157
-0.01282        -0.02396
-0.12355        -0.24412
0.09790         0.18171
-0.17892        -0.37216
1.01486         1.88497
-1.81905        -3.46314
```

Jumlah observasi yang dikategorikan sebagai pencilan k=m	Data Pencilan	R^2	AIC	BIC
0	-	0.988	-82.21	-82.21
1	(2,5;8)	0.997	**-103.60**	**-104.56**
2	(2,5;8) dan (5,5;17)	1	$+\infty$	$+\infty$

Berdasarkan hasil perhitungan AIC dan BIC diatas diperoleh kesimpulan bahwa data pencilan adalah (2,5;8). Hasil ini sama dengan uji simpangan baku. Sedangkan dengan uji sisa ter-student diperoleh bahwa data pencilan adalah (2,5;8) dan (5,5;17).

Melihat kepada hasil ketiga ilustrasi diatas dapat diperoleh hasil bahwa ketiga uji pencilan belum tentu memberikan kesimpulan yang sama. Jadi ada data yang oleh suatu uji dapat diketegorikan sebagai data pencilan namun dengan uji lain data tersebut tidak dapat dikategorikan sebagai data pencilan. Kalau dibandingkan uji-uji yang dipakai dengan grafik data aslinya, terlihat dari ilustrasi ini bahwa data-data yang dikategorikan sebagai data pencilan menurut uji sis ter-student, hampir bersesuaian juga dengan posisi data tersebut dilihat dari grafik, yang memang cukup memencil dibandingkan dengan data-data lainnya. Melihat hasil ini sepertinya uji sisa ter-student memberikan hasil yang relatif akurat dibandingkan dua uji lainnya walaupun dari ilustrasi ini uji simpangan baku memberikan hasil yang sama dengan uji AIC dan BIC.

Dilihat dari kesederhanan dan keefektifannya, uji sisa ter-student ini lebih mudah untuk digunakan dibandingkan dengan uji AIC dan BIC. Karena saat ini perangkat lunak statistika selalu menyediakan menu untuk menghitung sisa ter-student ini. Sedangkan uji AIC dan BIC belum tersedia fasilitas untuk itu. Disamping itu dengan uji AIC dan BIC ini kita harus memprediksi dulu data-data yang mungkin dikategorikan sebagai data pencilan dan kita harus melakukan regresi beberapa kali tergantung kapada berapa macam kelompok data yang diasumsikan sebagai data pencilan. Penggunaan ketiga uji pencilan ini tidak tergantung kepada pola-pola datanya.

KESIMPULAN

Dari pembahasan dan penjelasan diatas dapat diambil beberapa kesimpulan sebagai berikut:

1. Uji pencilan dangan simpangan baku, uji pencilan dengan sisa ter-student dan uji pencilan dengan kriteria informasi, belum tentu memberikan kesimpulan yang sama untuk menentukan data mana yang dikategorikan sebagai data pencilan pada suatu masalah regresi linier.
2. Uji pencilan dangan simpangan baku dan uji pencilan dengan sisa ter-student lebih sederhana untuk digunakan dibandingkan dengan uji pencilan kriteria informasi.
3. Uji pencilan dengan sisa ter-student memberikan hasil yang relatif akurat dari dua uji lainnya.

DAFTAR PUSTAKA

Barnett, V. Principles and methods for handling outliers in data sets. *Statis-tical Methods and The Improvement of Data Quality*, pp. 131-166, 1983

Chatterjee S. and B. Price, Regression Analysis by Example, John Wiley & Sons, New York, 1977.

Gentleman, J.F. dan Wilk, M.B., Detecting Outliers in a two-way table I, *Technometrics* , 17,h.1-14, 1975a.

Gentleman, J.F. dan Wilk, M.B., Detecting Outliers II , *Biometrics* , 31,h.387-400, 1975b.

Pynnonen, S. Detection of Outlier in Regression Analysis by Information Criteria, *www.uwasa.fi/~sjp* , 1992.

Ryan, T.P., *Modern Regression Methods*, John Wiley & Sons, New York, 1997

Sembiring, R. K., *Analisis Regresi*, ITB, 1995.

Weisberg, S., *Applied Linear Regression* ,ed-2. John Wiley & Sons, New York, 1985