MATEMATIKA, 1990, Jilid 6, no.1, hlm 1-13
© Jabatan Matematik, UTM.

MENGESAN DATA TERPENCIL MENERUSI REGRESI NORMA-L1

Mokhtar bin Abdullah
Jabatan Statistik
Pusat Pengajian Kuantitatif
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor.

ABSTRAK

Satu tatacara pengujian berasaskan peminimuman hasil tambah reja mutlak (dikenali juga sebagai tatacara norma-L1) disarankan bagi mengesan data terpencil di dalam model regresi ringkas. Nilai-nilai genting bagi statistik ujian ditentukan daripada kajian simulasi. Anggaran norma-L1 dihitung menggunakan alkhwarizmi yang dicadangkan oleh Barrodale dan Roberts (1973).

1. Pengenalan

Masalah mengesan data terpencil di dalam model regresi telah banyak dibincangkan oleh para penyelidik. Banyak statistik ujian telah dicadangkan dan kebanyakannya berasaskan reja kuasa dua terkecil (lihat sebagai contoh, Srikantan (1961) dan Tietjen, Moore dan Beckman (1973)).

Telah diketahui bahawa penganggar yang berasaskan tatacara kuasa dua terkecil amat mudah dipengaruhi oleh data terpencil. Dalam suasana sedemikian penganggar yang berasaskan median memberikan alternatif yang lebih baik (atau lebih kukuh). Menurut Harvey (1977) penganggar yang meminimumkan hasil tambah reja regresi mutlak (dikenali juga sebagai norma-L1) bagi model regresi ringkas merupakan pengitlakan daripada median sampel bagi model lokasi. Sorotan ringkas tentang regresi norma-L1 diberikan di dalam Bahagian 2.

Di dalam kertas ini tatacara ujian norma-L1 dicadangkan bagi mengesan data terpencil tunggal bagi model regresi ringkas. Perumusan statistik ujian dibincangkan di dalam Bahagian 3. Pendekatan yang serupa seperti yang disarankan oleh Tietjen, Moore dan Beckman (1973) digunakan dalam menentukan nilai-nilai genting. Iaitu, nilai-nilai genting ditentukan daripada anggaran nilai-nilai persentil bagi statistik ujian tersebut. Kesan pembolehubah bebas terhadap nilai-nilai genting juga ditinjau. Dalam Bahagian 4 contoh penggunaan tatacara ujian norma-L_1 ditunjukkan berdasarkan data Forbes (lihat Forbes(1857)). Kuasa ujian juga ditinjau dan dibincangkan dalam Bahagian 5. Pengitlakan tatacara ujian ini kepada masalah mengesan beberapa data terpencil di dalam model regresi ringkas dan juga model regresi berganda dibentangkan di dalam Bahagian 6.

2. Regresi Norma-L1

Pertimbangkan model regresi linear ringkas

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \; i = 1,2,\ldots,n$$

dengan $(x_i \; , \; y_i)$ sebagai pasangan cerapan bagi pembolehubah merdeka dan bersandar dan ε ialah ralat rawak. β_0 dan β_1 merupakan parameter anu yang akan dianggar.

Regresi norma-L1 (y ke atas x) dijalankan dengan meminimumkan

$$Q = \sum_{i=1}^{n} | y_i - \beta_0 - \beta_1 x_i | \qquad (2.1)$$

Penggunaan tatacara norma-L1 berkait rapat dengan perkembangan kaedah penganggaran kukuh (lihat Huber (1973) atau Andrews (1974)). Regresi kuasa dua terkecil mempunyai sifat optimum apabila ralatnya tertabur normal. Menurut Andrews(1974) walaupun ralatnya tertabur normal, alternatif kepada kaedah kuasa terkecil masih diperlukan terutamanya apabila bentuk model sebenar tidak diketahui dengan tepat. Ekblom(1974) mencadangkan penggunaan tatacara norma-L1 apabila ralat tertabur secara taburan normal terkontaminasi. Telah diketahui bahawa anggaran norma-L1 boleh diperolehi daripada kaedah kebolehjadian maksimum dan mempunyai sifat optimum apabila ralatnya tertabur secara taburan Laplace (iaitu eksponen ganda dua)(lihat Kendall dan Stuart, 1973). Penganggaran norma-L1 menerusi pendekatan pengaturcaraan linear telah dikemukakan oleh Wagner (1959), Barrodale dan Roberts (1973),dan Amstrung dan Kung(1978).

3. Tatacara ujian norma-L1

Katalah r_i, i=1,2,...,n sebagai reja norma-L1 yang ditakrifkan sebagai

$$r_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i$$

dengan $\tilde{\beta}_0$ dan $\tilde{\beta}_1$ sebagai anggaran norma-L1 bagi β_0 dan β_1.

Juga, andaikan

$r_{(n)} = \text{maks} \, |r_i|, \; i=1,2,\ldots,n$ sebagai reja mutlak terbesar dengan $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(n)}$.

$r_{(n)}$ yang cukup besar berbanding dengan reja-reja yang lain mungkin menandakan adanya data terpencil tunggal di dalam set data. Reja ini juga boleh memberikan sumbangan yang bererti kepada hasil tambah reja mutlak Q. Sumbangan ini dapat dilihat apabila ia disingkirkan daripada Q.

Katalah

$$Q_n = Q - r_{(n)} \tag{2.2}$$

sebagai hasil tambah reja mutlak terlaras apabila reja mutlak terbesar itu digugurkan.

Dengan berasaskan (2.2) kita dapat takrifkan statistik ujian

$$T_n = Q_n / Q \tag{2.3}$$

bagi menguji sama ada terdapat data terpencil (yang sepadan dengan reja mutlak terbesar) di dalam set data. Nilai T_n yang cukup kecil memberi pertanda bahawa wujud data terpencil di dalam set data tersebut.

3.1 Penentuan nilai genting

Nilai genting bagi T_n, katalah h_α, ditentukan daripada hubungan

$$Pr\,[\,T_n < h_\alpha\,] = \alpha$$

pada paras keertian α . Taburan T_n sukar diperolehi secara analisis. Oleh itu, nilai-nilai h_α ditentukan menerusi kajian simulasi seperti berikut;

L sampel rawak dijanakan dengan setiap sampel bersaiz n cerapan. Untuk setiap sampel, didapatkan $\tilde{\beta}_0$ dan $\tilde{\beta}_1$ dan seterusnya T_n dihitung daripada (2.3). Daripada nilai-nilai T_n yang dihitung itu, susun nilai-nilai tersebut mengikut tertib menaik. Seterusnya, nilai-nilai persentil bagi T_n ditentukan daripada susunan tertib tersebut. Nilai-nilai persentil ini digunakan sebagai mewakili nilai-nilai genting h_α dan persentil ke-1, ke-5 dan ke-10 masing-masingnya mewakili paras keertian α = 0.01, 0.05 dan 0.10.

Penentuan nilai-nilai h_α dilakukan untuk setiap $n = 5(1), 20(10), 50$ seperti yang terpapar dalam Jadual 1. Dalam kajian simulasi ini rutin-NAG G05DDF digunakan bagi menjanakan nilai-nilai variat normal ε.

Rutin-NAG E02GAF (berasaskan alkhwarizmi Barrodale dan Roberts(1973)) digunakan untuk menghitung anggaran-anggaran regresi norma-L1. Semua aturcara komputer di tulis dalam bahasa FORTRAN dan dilaksanakan oleh sistem komputer IBM 4341 UKM.

3.2 Kesan nilai x ke atas nilai genting

Salah satu kesukaran dalam menentukan nilai genting h_α ialah taburan T_n bergantung kepada nilai-nilai x sehinggakan penjadualan secara tepat titik-titik persentilnya sukar dilakukan.

Untuk melihat bagaimana berbagai susunan nilai x mempengaruhi nilai-nilai genting, pendekatan yang serupa dengan yang disarankan oleh Tietjen, Moore dan Beckman (1973) digunakan. Iaitu, empat susunan nilai x yang melampau dicuba untuk setiap nilai n.

I. $x_1 = x_2 = 1,\ x_i = 0,\ i=3,4,\ldots,n$
II. $x_1 = x_2 = 0,\ x_i = 0.5,\ i=3,4,\ldots,n$
III. $[n/2]$ nilai $x_i = 1$, x_i lain $= 0$.
IV. Nilai $x_i \in [0,1]$ dan berjarak seragam.

Nilai-nilai genting untuk $\alpha = 0.01,\ 0.05$ dan 0.10 dihitung untuk setiap $n = 5(1),20(10),50$ bagi setiap susunan di atas.

Didapati bahawa purata nilai genting susunan I menghampiri purata nilai genting susunan IV manakala purata nilai genting susunan II mendekati purata nilai genting susunan III. Dengan itu, empat susunan di atas boleh dikumpulkan dalam dua susunan sahaja, iaitu, kumpulan A terdiri daripada susunan I dan IV manakala kumpulan B terdiri daripada susunan II dan III. Purata nilai genting kumpulan A kemudiannya dibandingkan dengan purata nilai genting kumpulan B dengan menggunakan ujian tanda. Didapati perbezaannya tidak bererti pada paras 5%. Dengan

itu, kita simpulkan bahawa perbezaan kedua-dua purata nilai genting tersebut boleh diabaikan dan seterusnya kesandaran nilai-nilai genting kepada nilai x tidaklah begitu kuat. Nilai genting hampiran kemudiannya didapati dengan mengambil purata nilai-nilai genting daripada keempat-empat susunan tersebut. Nilai-nilai genting hampiran ini dipaparkan dalam Jadual 1.

4. Contoh

Data yang digunakan dalam bahagian ini diambil daripada Forbes(1857). Analisis terperinci ke atas data ini menerusi kaedah kuasa dua terkecil boleh didapati daripada Weisberg(1980). Forbes mengkaji hubungan di antara titik didih air dan tekanan udara dalam usahanya untuk menganggar altitud di atas paras laut. Pembolehubah bersandar (y) ialah titik didih (dalam ^{0}F) manakala pembolehubah merdeka (x) ialah 100xlog tekanan (dalam Hg). Data yang diperolehi adalah seperti yang terpapar di dalam Jadual 2.

Penggunaan rutin-NAG E02GAF memberikan anggaran norma-L1

$$\tilde{\beta}_0 = -41.844 \quad ; \quad \tilde{\beta}_1 = 0.893$$

Sebagai perbandingan, kaedah kuasa dua terkecil memberikan

$$\hat{\beta}_0 = -42.131 \quad ; \quad \hat{\beta}_1 = 0.895$$

dan kedua-dua reja norma-L1 dan kuasa dua terkecil ditunjukkan dalam dua lajur terakhir di dalam Jadual 2.

Weisberg (1980) telah menghasilkan plot reja di antara reja kuasa dua terkecil dan nilai suaian kuasa dua terkecil. Plot ini memperlihatkan yang cerapan 12 adalah data terpencil dan ini disahkan lagi apabila ia digugurkan daripada set data.

α	.01	.05	.10
n		h_α	
5	.374	.473	.525
6	.493	.579	.617
7	.521	.603	.639
8	.604	.671	.699
9	.622	.680	.709
10	.678	.727	.751
11	.683	.735	.758
12	.728	.769	.788
13	.732	.773	.793
14	.762	.790	.798
15	.763	.801	.818
16	.792	.823	.837
17	.793	.825	.839
18	.813	.840	.850
19	.814	.842	.855
20	.831	.854	.866
30	.884	.901	.908
40	.913	.924	.929
50	.929	.938	.942

Jadual 1 : Nilai-nilai genting bagi T_n

i	x_i	y_i	Reja Norma-L1	Reja kuasa dua terkecil
1	194.5	131.79	-0.192	-0.25
2	194.3	131.79	-0.013	-0.07
3	197.9	135.02	0.000	-0.06
4	198.4	135.55	0.083	0.02
5	199.4	136.46	0.099	0.03
6	199.9	136.86	0.022	-0.04
7	200.9	137.82	0.118	0.50
8	201.1	138.00	0.120	0.05
9	201.4	138.06	-0.088	-0.16
10	201.3	138.05	-0.009	-0.08
11	203.6	140.04	-0.074	-0.15
12	204.6	142.44	1.432	1.36
13	209.5	145.47	0.083	0.00
14	208.6	144.34	-0.243	0.32
15	210.7	146.30	-0.160	-0.24
16	211.9	147.54	0.008	-0.08
17	212.2	147.80	0.000	-0.09

Jadual 2 : Reja Norma-L1 dan reja kuasa dua terkecil bagi data Forbes(1857).

Anggaran kuasa dua terkecil tanpa cerapan 12 ialah

$$\hat{\beta}_0^* = -41.302 \qquad \hat{\beta}_1^* = 0.891$$

Melihat kepada reja norma-L1 bagi cerapan 12, kesimpulan yang serupa boleh dibuat tentang cerapan ini. Malah, nilai reja norma-L1 ternyata lebih besar daripada nilai reja kuasa dua terkecil. Ini memperkuatkan lagi kesimpulan tentang terpencilnya cerapan ini daripada cerapan-cerapan yang lain.

Secara formalnya kita ingin menguji adakah cerapan 12 benar-benar data terpencil. Kita gunakan tatacara ujian norma-L1 yang disarankan dengan $n = 17$. Mula-mula hitung Q, iaitu hasil tambah reja mutlak.

Daripada Jadual 2, kita dapati

$$Q = \sum_{i=1}^{17} |r_i| = 2.744$$

Reja mutlak terbesar ialah $r_{(17)} = 1.432$. Dengan itu, hasil tambah reja mutlak terlaras menjadi

$$Q_{17} = Q - r_{(17)} = 1.312$$

Dengan statistik ujian (2.3), kita dapati

$$T_{17} = Q_{17} / Q = 0.478$$

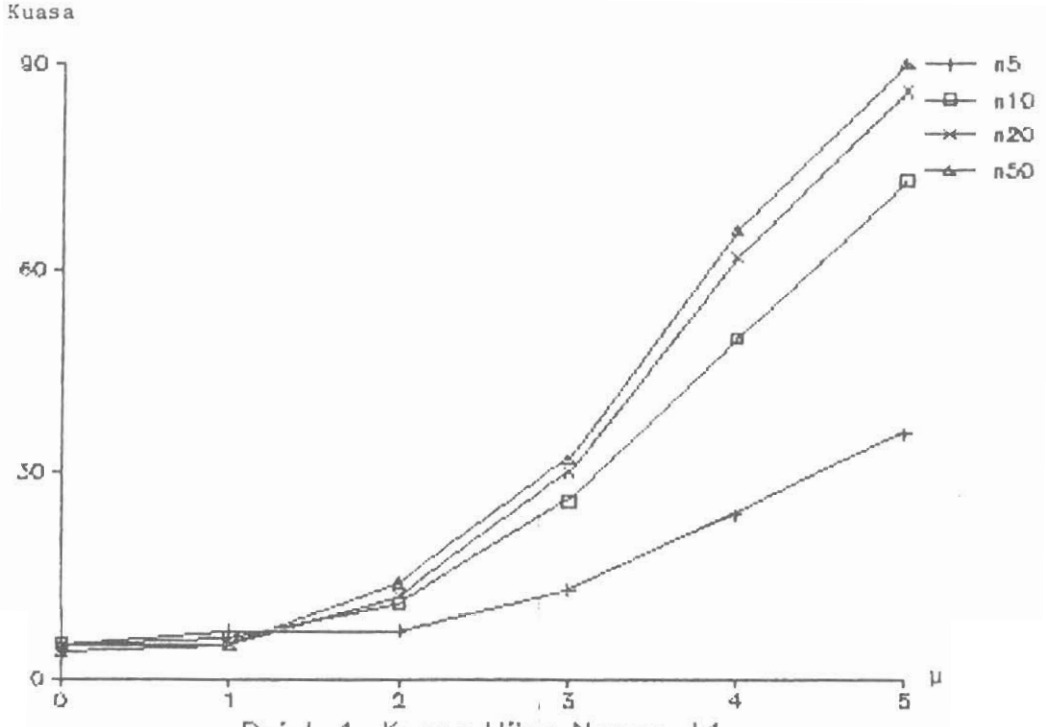
Oleh kerana T_{17} lebih kecil daripada kedua-dua $h_{0.05}$(=0.825) dan $h_{0.01}$(=0.793), maka kita simpulkan bahawa cerapan 12 adalah data terpencil pada kedua-dua paras 5% dan 1%.

5. 'Kuasa' ujian norma-L1

'Kuasa' ujian norma-L1 diukur dalam sebutan peratusan sampel yang data terpencil di dalam setiap sampel berjaya dikesan oleh tatacara ujian ini. Kajian dijalankan berdasarkan 2000 sampel yang setiapnya bersaiz 5, 10, 20 dan 50 dan $\beta_0 = \beta_1 = 1$.

Data terpencil tunggal (yang sepadan dengan ε_r, $r \neq i$, $i=1,2,\ldots,n$) daripada taburan normal $N(\mu,1)$ dengan $\mu = 1(1)5$ digunakan untuk mengkontaminasikan sampel. ε_i, $i=1,2,\ldots,n$ yang tidak terkontaminasi diambil daripada taburan normal $N(0,1)$.

Rajah 1 memperlihatkan kuasa ujian norma-L_1 dalam sebutan peratusan sampel yang data terpencil berjaya dikesan oleh tatacara ujian ini.



Rajah 1: Kuasa Ujian Norma-L1

6. Kesimpulan

Kita telah mengemukakan statistik ujian yang berasaskan tatacara norma-L1 untuk menguji data terpencil di dalam regresi linear ringkas. Idea di sebalik tatacara ujian ini ialah sesuatu cerapan disifatkan sebagai data terpencil apabila penyingkirannya daripada hasil tambah reja mutlak mengakibatkan penurunan nilai yang terbesar di dalam hasil tambah tersebut.

Bagaimanapun, penghitungan untuk mendapatkan anggaran norma-L1 adalah lebih sukar berbanding dengan kaedah kuasa dua terkecil. Tetapi dengan adanya beberapa alkhwarizmi yang cekap seperti yang dikemukakan oleh Barrodale dan Roberts (1973), Amstrung dan Kung (1978) dan beberapa lagi, kesukaran ini sudah boleh diatasi.

Tatacara ujian yang disarankan ini boleh diperluaskan dengan mudah kepada kes-kes yang membabitkan beberapa data terpencil. Untuk k > 1 data terpencil (2.2) boleh dituliskan sebagai

$$Q_{n,k} = Q - \sum_{j=n-k+1}^{n} r_{(j)}$$

dengan $r_{(j)}$ sebagai reja mutlak terbesar ke-j dan seterusnya statistik ujian (2.3) menjadi

$$T_{n,k} = Q_{n,k} / Q$$

Perluasan kepada regresi berganda juga dapat dilakukan secara langsung dengan

$$Q = \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p} \beta_j x_{ji} \right|$$

$$r_i = y_i - \sum_{j=0}^{p} \tilde{\beta}_j x_{ji}$$

dan $\tilde{\beta}_j$, j=0,1,...,p adalah anggaran-anggaran parameter regresi norma-L1.

RUJUKAN

[1] Andrews, D.F.(1974), A Robust Method for Multiple Linear Regression, Technometrics,16,523-551.

[2] Amstrong, R.D. & Kung, M.T.(1978), AS132: Least Absolute Value Estimates for a Simple Linear Regression Problem, Applied Statistics,27,363-366.

[3] Barrodale, I. & Roberts, F.D.K.(1973), An Improved Algorithm for Discrete L1 Linear Approximation, SIAM Journal of Numerical Analysis, 10, 839-848.

[4] Ekblom, H.(1974), L_p Methods for Robust Regression, BIT 14,22-23.

[5] Forbes, J.D.(1857), Further Experiments and Remarks on the Measurement of Heights by the Boiling Point of Water, Transaction of the Royal Society, Edinburgh,21,135-143.

[6] Harvey, A.C.(1977), A Comparison of Preliminary Estimators for Robust Regression, J.of Amer.Statist. Assoc.,72,910-913.

[7] Huber, P.J.(1973), Robust Regression: Asymptotics, Conjectures and Monte Carlo, Annals of Mathematical Statistics,1,799-821.

[8] Kendall, M.G. & Stuart, A.(1973), The Advanced Theory of Statistics, Vol. 2, Griffin, London.

[9] NAG Library Manuals Mark 8 Oxford (1981),Numerical algorithms Group.

[10] Srikantan, K.S.(1961), Testing for a Single Outlier in a Regression Model, Sankhya, Ser. A,23,251-260.

[11] Tietjen, G.L, Moore, R.L. & Beckman, R.J. (1973), Testing for a Single Outlier in Simple Linear Regression, Technometrics, 15, 717-721.

[12] Wagner, W.H. (1959), Linear Programming Techniques for Regression Analysis, J. of Amer. Statist. Assoc., 54, 206-212.

[13] Weisberg, S. (1980), Applied Linear Regression, John Wiley , New York.