PEKALI KORELASI: PENDEKATAN TEGUH

NEO SIEW PING

FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA TERENGGANU

2003

Pekali korelasi : pendekatan teguh / Neo Siew Ping.

Lihat sebelah

PEKALI KORELASI: PENDEKATAN TEGUH

Oleh
Neo Siew Ping

Projek Ilmiah Tahun Akhir ini diserahkan untuk memenuhi
sebahagian keperluan bagi
Ijazah Sarjana Muda Sains (Matematik Komputasi)

JABATAN MATEMATIK
FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA TERENGGANU
2009

JABATAN MATEMATIK
FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA TERENGGANU

PENGAKUAN DAN PENGESAHAN LAPORAN MAT 4499 B

Adalah ini diakui dan disahkan bahawa laporan penyelidikan bertajuk Pekali Korelasi: Pendekatan Teguh oleh Neo Siew Ping, No matrik: UK 13235 telah diperiksa dan semua pembetulan yang disarankan telah dilakukan. Laporan ini dikemukakan kepada Jabatan Matematik sebagai memenuhi sebahagian daripada keperluan memperolehi Ijazah Sarjana Muda Sains Matematik Komputasi, Fakulti Sains dan Teknologi, UMT.

Disahkan oleh :

.....

Penyelia
Nama: Dr. Muhammad Safiih Bin Lola
Cop Rasmi :

MUHAMAD SAFIIH BIN LOLA
Pensyarah
Jabatan Matematik
Fakulti Sains dan Teknologi
Universiti Malaysia Terengganu
21030 Kuala Terengganu

Tarikh: 30/04/09

Disahkan oleh :

.....

Ketua Jabatan Matematik
Nama : Dr Haji Mustafa bin Mamat
Cop Rasmi:

DR. HJ. MUSTAFA BIN MAMAT
Ketua
Jabatan Matematik
Fakulti Sains dan Teknologi
Universiti Malaysia Terengganu
21030 Kuala Terengganu

Tarikh : 30/04/2009

PENGAKUAN

Saya mengakui projek ilmiah tahun akhir yang bertajuk **Pekali Korelasi: Pendekatan Teguh** adalah hasil kerja saya kecuali nukilan dan ringkasan yang setiap satunya telah saya jelaskan sumbernya.

Tandangan : 
Nama : Neo Siew Ping
No. Matrik : UK13235
Tarikh : 30 April 2009

PENGHARGAAN

Terlebih dahulu, saya ingin merakamkan penghargaan ikhlas kepada penyelia Projek Ilmiah Tahun Akhir saya, Dr. Muhamad Safiih Bin Lola atas bimbingan dan dorongan yang diberikan oleh beliau sepanjang tempoh projek penyelidikan ini untuk memenuhi sebahagian keperluan bagi Sarjana Muda Sains (Matematik Komputasi). Saya ingin mengucapkan ribuan terima kasih kepada Dr. Safiih kerana menawarkan peluang berharga kepada saya untuk melibatkan diri dalam penyelidikan keteguhan pekali korelasi dalam bidang statistik. Dr. Safiih telah memberi panduan untuk strategi dan cara menjalankan penyelidikan dengan efektif. Di samping itu, beliau telah memberi nasihat kepada saya dalam menyediakan laporan Projek Ilmiah Tahun Akhir ini dan semasa persediaan persembahan laporan. Selain itu, beliau sanggup meluangkan dan mengorbankan masa beliau untuk menyelesaikan segala masalah yang dihadapi oleh pelajarnya sepanjang penyelidikan projek ini. Beliau juga sanggup berkongsi sagala pengalaman dan pengetahuan beliau untuk membantu pelajarnya. Pengorbanan beliau adalah amat dihargai sekali.

Seterusnya, saya ingin berterima kasih kepada Puan Nor Azlida binti Aleng@ Mohamad, penyelaras Projek Ilmiah Tahun Akhir (PITA) Jabatan Matematik, yang menyelaraskan pelajar untuk memastikan semua pelajar menjalankan satu penyelidikan dalam bidang Matematik untuk memenuhi jam kredit Sarjana Muda. Di samping itu, saya juga ingin merakamkan ribuan terima kasih kepada semua pensyarah Jabatan Matematik Universiti Malaysia Terengganu (UMT).

Pada akhirnya, penghargaan juga ingin ditujukan oleh saya kepada mereka terutamanya keluarga yang memberi galakan sepenuhnya dan rakan saya yang terlibat sama ada secara langsung atau tidak langsung dalam membantu penulis untuk menjayakan projek penyelidikan ini. Sekian, terima kasih.

ABSTRAK

Pekali korelasi Pearson's, pekali korelasi Spearman's rho dan pekali korelasi Kendall's tau dipengaruhi oleh kehadiran data ekstrim dalam sampel data di mana titik bendung pengaruh pekali korelasi turut dipengaruhi. Keputusan telah menunjukkan bahawa pekali korelasi ini adalah cukup teguh terhadap sebahagian bilangan data ekstrim sahaja. Pekali korelasi ini tidak mempunyai bendung pengaruh yang tinggi dan sensitif terhadap kehadiran data 'rosak'. Dalam kajian ini, satu pekali korelasi berasaskan penganggar-M telah diutarakan. Penganggar-M telah digabungkan dengan Spearman's rho untuk mengkaji bendung pengaruhnya. Adalah ditunjukkan bahawa pekali korelasi Spearman's rho berasaskan penganggar-M mempunyai bendung pengaruh yang tinggi dan kurang sensitif kepada data ekstrim apabila sampel data nyata dicemarkan berbanding dengan pekali korelasi Pearson's, Spearman's rho dan Kendall's tau. Kesimpulannya, penganggar-M boleh mengekalkan keteguhan pekali korelasi dengan titik bendung pengaruh yang tinggi.

ABSTRACT

The correlation coefficients, Pearson's, Spearman's rho and Kendall's tau are being affected by the existence of outliers in the data which can influence the breakdown point of the correlation coefficients. The result shows that these correlation coefficients are only sufficiently robust in the presence of some outliers. Moreover, these correlation coefficients do not have high breakdown points when there are more real data being contaminated. In this paper, a correlation coefficient based on the M-estimator is being proposed. M-estimator has been incorporated into Spearman's rho to investigate its breakdown point. It is shown that correlation coefficient Spearman's rho based on M-estimator has a higher breakdown point and is less sensitive to the existence of outliers when the real data are being contaminated compare to correlation coefficients Pearson's, Spearman's rho and Kendall's tau. As a conclusion, M-estimator can be used to remain the robustness of correlation coefficients with high breakdown point.

KANDUNGAN

SENARAI JADUAL

SENARAI RAJAH

BAB 1

PENDAHULUAN

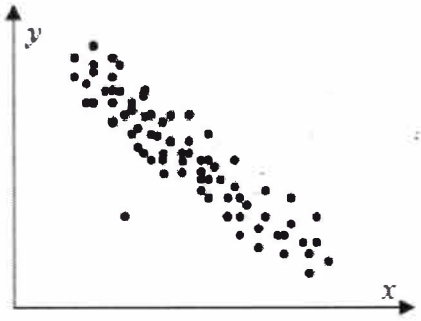
1.1 Pengenalan

1.1.1 Regresi Teguh

Dalam statistik, regresi teguh adalah salah satu bentuk analisis regresi yang direka untuk memintasi pembatasan kaedah parametrik dan bukan parametrik tradisional. Dengan spesifikasinya, penganggar kuasa dua terkecil untuk model regresi adalah tidak cukup teguh kepada data ekstrim.

1.1.2 Data Ekstrim Dalam Analisis Regresi

Salah satu sebab statistik teguh diperlukan adalah kewujudan data ekstrim dalam satu set data. Data ekstrim adalah data yang biasa terpisah daripada suatu taburan data yang besar (Wikipedia). Keadaan ini boleh disebabkan oleh kesalahan jumlah kasar. Contohnya kesalahan salinan atau kesalahan tempat perpuluhan, kesalahan skala ukuran yang diambil untuk satu jangka masa, pertukaran dua nilai yang berbeza maksud, kegagalan peralatan atau sebagainya. Apabla data adalah diambil daripada populasi bertaburan normal, kewujudan data ekstrim dalam suatu data boleh diminimumkan dengan menggunakan penganggar teguh yang mempunyai titik bendung pengaruh tinggi.



Rajah 1.1: Ruang dua dimensi dengan satu data ekstrim (sudut bawah kiri)

1.1.3 Pekali Korelasi

Pekali korelasi digunakan untuk mengukur kekuatan di antara pembolehubah bersandar dan pembolehubah tidak bersandar. Bentuk-bentuk pekali korelasi adalah seperti pekali korelasi hasil darab, Spearman's rho dan Kendall' s tau.

Biarkan $(x_1, y_1),...,(x_n, y_n)$ terhadap n pemerhatian daripada satu dwipembolehubah taburan normal parameter $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ di mana μ_x dan σ_x^2 adalah min dan variant bagi x, μ_y dan σ_y^2 adalah min dan variant bagi y. ρ adalah pekali korelasi antara x dan y dan diberi sebagai $\rho = \frac{\beta\sigma_x}{\sigma_y}$ dengan β adalah parameter kecerunan regresi y atas x. Sampel pekali korelasi yang biasa digunakan untuk menganggar ρ adalah pekali korelasi hasil darab Pearson's (r_p) yang ditakrifkan sebagai

$$r_p = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{[\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2]^{1/2}} \tag{1.1}$$

Dalam masalah nyata adalah selalu dinyatakan bahawa data ekstrim boleh berlaku sama ada pada pembolehubah bersandar y atau pembolehubah tidak bersandar x atau

pada kedua-dua pembolehubah tersebut. Pekali korelasi r_ρ dalam (1.1) adalah berdasarkan sampel min $\bar{x}$ dan $\bar{y}$ masing-masing yang sangat sensitif kepada kehadiran data ekstrim. Chernick (1982) mengillustrasikan ketakteguhan r_ρ dengan menunjukkan bahawa fungsi terpengaruhnya adalah tidak terbatas.

Sebagai alternatif kepada pekali korelasi hasil darab r_ρ kita boleh merujuk kepada pekali korelasi tidak berparametrik di mana pekali korelasinya adalah berdasarkan kepada peringkat pemerhatian (Mokhtar Abdullah, 1990). Tidak berparametrik bermaksud tidak ada andaian yang dibuat berkenaan dengan taburan fungsi di bawah pemerhatian. Dua pekali korelasi jenis ini yang terkenal adalah Spearman's rho (r_s) dan Kendall's tau (r_k). Spearman's rho boleh dikira dengan formula

$$r_s = 1 - \frac{6D^2}{n(n^2-1)} \tag{1.2}$$

dengan

$$D^2 = \sum_i^n (r_{yi} - r_{xi})^2 \tag{1.3}$$

dan r_{xi} dan r_{yi} adalah peringkat bagi x_i dan y_i masing-masing. Kendall's tau, r_k adalah diberi sebagai

$$r_k = 1 - \frac{4Q}{n(n-1)} \tag{1.4}$$

dengan Q adalah nombor penyongsangan antara peringkat x dan y. Satu penyongsangan adalah sebarang pasangan objek (i, j) di mana $r_i - r_j$ dan $r_i' - r_j'$ mempunyai tanda yang bertentangan.

Dengan menggantikan nilai pemerhatian dengan peringkatnya, pengaruh bagi sesetengah pemerhatian ekstrim boleh dikurangkan. Oleh itu, kita boleh menjangka pekali korelasi r_s dan r_k adalah kurang sensitif kepada data ekstrim berbanding dengan pekali korelasi hasil darab r_p.

1.1.4 Statistik Teguh

Statistik teguh memberi pendekatan alternatif kepada kaedah statistik yang klasik. Statistik teguh menyumbangkan kepada kaedah yang mencontohi kaedah klasik tetapi dipengaruhi oleh data ekstrim atau keberangkatan kecil daripada model andaian (Wikipedia). Contohnya, median dan median lencongan mutlak adalah penilaian teguh. Median mempunyai titik bendung pengaruh sebanyak 50 peratus. Contoh statistik tidak teguh adalah seperti min dan julat. Min hanya mempunyai titik bendung pengaruh 0 peratus.

Soalan yang masih timbul adalah sama ada korelasi berperingkat ini adalah cukup teguh terhadap bilangan data ekstrim yang teguh dan kuat iaitu sama ada mereka mempunyai bendung pengaruh yang tinggi. Titik analisis, ε^* adalah pembahagian terkecil bagi data ekstrim yang boleh menyebabkan satu penganggaran mengambil nilai sembarangan. Satu penganggar teguh adalah sesuatu dengan titik bendung pengaruh yang tinggi. Donoho dan Huber (1983) telah menunjukkan bahawa satu penganggar pekali korelasi yang berada dalam selang [-1,1] gagal apabila penganggar dapat bergerak kepada titik akhir selang tersebut atau dalam sesetengah kes kalau pencemaran boleh menyebabkan satu korelasi bukan kosong kepada kosong. $\varepsilon^* = 50$ peratus adalah yang terbaik dijangka dapat diperolehi dengan satu penganggar kerana apabila pencemaran menjadi semakin banyak, ini akam menyebabkan kesusahan membezakan di antara data 'baik' dan data 'rosak' bahagian-bahagian tertentu dalam sampel.

1.2 Pernyataan Masalah

Kebiasaannya, data ekstrim wujud dalam aplikasi lapangan. Kajian ini menggunakan beberapa jenis pekali korelasi bagi mendapatkan hasil dan data sampling yang terbaik. Dengan menggunakan penganggar teguh ini, ketinggian titik bendung pengaruh pekali korelasi yang terlibat dapat ditentukan. Persoalannya adalah untuk menentukan atau meramalkan ketinggian titik bendung pengaruh apabila pencemaran data sampling yang berikutnya menggunakan pekali korelasi teguh.

1.3 Batasan Kajian

Kajian ini tertumpu kepada permodelan matematik dan komputer secara serentak terhadap kewujudan dan pengaruh data ekstrim dalam data persampelan dan kewujudan data pincang dalam model. Tiga jenis pekali korelasi yang dikaji iaitu pekali korelasi Pearson's, Spearman's rho, Kendall's tau dan hanya model teguh Spearman berasaskan penganggar-M yang digunakan. Pengaruh peratus pencemaran data terhadap pekali korelasi ini diperhatikan.

1.4 Objektif Kajian

Antara objektif-objektif kajian untuk penyelidikan projek ini termasuklah untuk:

(i). Membangunkan model pekali korelasi teguh.

(ii). Mengkaji dan membandingkan sifat bendung pengaruh terhadap pekali korelasi dan penganggar teguh terhadap pekali korelasi Spearman's teguh.

(iii). Menentukan pekali korelasi teguh iaitu penganggar-M adalah pekali korelasi yang mempunyai bendung pengaruh yang lebih baik dalam kewujudan data ekstrim.

BAB 2

SOROTAN KAJIAN

2.1 Pengenalan

Dalam bab ini kita akan membincangkan tentang peramalan dan kajian-kajian lepas yang berkaitan dengan pekali korelasi teguh. Data ekstrim boleh menyebabkan keputusan anggaran tidak tepat dan jitu. Pekali korelasi teguh adalah sangat penting dalam meminimumkan pengaruh data ekstrim dalam satu taburan data persampelan.

2.2 Kajian Lepas Terhadap Pekali Korelasi Teguh

Mokhtar Bin Abdullah (1990) telah menjalankan kajian dalam pekali korelasi teguh. Beliau telah mengkaji beberapa pekali korelasi seperti pekali korelasi Pearson's, pekali korelasi Spearman's dan pekali korelasi Kendall's. Keputusan berangka menunjukkan bahawa pekali korelasi ini cukup teguh terhadap sejumlah data ekstrim yang besar. Walau bagaimanapun pekali korelasi ini tidak mempunyai bendung pengaruh yang tinggi. Dengan itu, beliau telah mengkaji pekali korelasi kuasa dua median terkecil yang didapati mempunyai bendung pengaruh yang tinggi berbanding dengan pekali korelasi Pearson's.

Massart et. al, (1986) telah membuat kajian mengenai kaedah teguh kuasa dua median terkecil dan pengesanan kesalahan model dalam regresi dan penentu ukuran. Dalam kajian ini, kuasa dua terkecil digunakan untuk membuat perbandingan dengan penganggar kuasa dua median terkecil. Daripada hasil perbincangan dan keputusan yang telah dibuat, mereka membuat keputusan bahawa kuasa dua median terkecil adalah satu kaedah teguh yang tidak sentitif kepada data ekstrim atau perlanggaran anggapan yang lain untuk model biasa dan normal. Ini adalah bertentangan dengan kaedah regresi ynag konvensional yang meminimumkan penambahan kuasa dua. Mereka telah menunjukkan bahawa pekali regresi ini boleh digunakan untuk mengesan atau membetulkan data ekstrim atau kesalahan model dalam aplikasi penentu ukuran.

Ortiz et. al, (2005) telah membuat kajian mengenai teknik regresi teguh yang merupakan alternatif dalam mengesan data ekstrim dalam analisis kimia. Mereka mengemukakan bahawa kehadiran data ekstrim mempunyai pengaruh yang penting dalam penentuan sensitiviti, julat linear atau kebolehupayaan mengesan antara lain apabila nombor merit ini dikembangkan dengan kaedah tidak teguh. Sesetengah kaedah teguh yang digunakan dalam penentu ukuran dalam kimia analisis seperti penganggar-M Huber, penganggar-GM Tukey dan Welsh, penganggar kabur, penganggar-M berkekangan dan kuasa dua trim terkecil. Kajian mendapati bahawa regresi kuasa dua median terkecil mempunyai kebolehan yang teguh dalam mengesan data ekstrim dalam analisis kimia. Analisis perbandingan telah dilakukan dengan mempraktikan kaedah teguh kepada data sebenar dan sintetik. Seterusnya, penggunaan regresi teguh telah dicadangkan terhadap ISO 5725-5.

Estelberger et. al, (1995) telah mengkaji pekali korelasi berperingkat dalam pentafsiran data makmal. Dalam kajian ini, dua pekali korelasi telah digunakan untuk perbandingan iaitu pekali korelasi Pearson's product-moment dan pekali korelasi Spearman's rho. Dalam bahagian pertama, mereka mengkaji hubungan kelinearan bagi kedua-dua pekali korelasi tersebut. Satu ujian makmal baru ('atribut y') diambil yang

mempunyai julat kelinearan berbanding dengan ujian piawai ('atribut x'). Didapati pekali korelasi Pearson's dan kuasa duanya menurun apabila hubungan kelinearan muncul antara x dan y. Bagi pekali korelasi berperingkat pula, mengekalkan nilai maksimum 1.0 akibat kekuatan sifat monotonik antara dua atribut. Jadi, pekali korelasi berperingkat juga membantu dalam mengesan hubungan kelinearan antara atribut. Dalam bahagian dua kajian mereka, mereka menganalisis satu set data makmal. Mereka membuat keputusan bahawa sesuatu data ekstrim adalah berpengaruh terhadap pekali korelasi Pearson's dan dalam analisis regresi linear tetapi pekali korelasi berperingkat adalah tidak terpengaruh secara pratikal. Ketidaksensitiviti pekali korelasi adalah akibat semula jadi yang disebabkan oleh pengukuran nilai kepada berperingkat.

Wisnowski et. al, (2001) telah menjalankan kajian ke atas teknik untuk mengesan pelbagai data ekstrim dalam regresi kelinearan dengan menggunakan simulasi Monte Carlo. Prosedur ini termasuk kaedah secara langsung daripada algoritma atau kaedah tidak langsung daripada penganggar regresi teguh. Mereka menghuraikan kesan ketumpatan data ekstrim dan geometri, pembolehubah dimensi regresi dan jarak data ekstrim. Antara kaedah penganggar teguh yang digunakan ialah LMS, LTS, M, MM, penganggar-M piawai, Coakley dan Hettmansperger dan Simpson dan Montgomery. Kertas kajian mereka menunjukkan penganggar Simpson dan Montgomery dan kaedah Rousseuw dan Van Zomeron dengan nilai terputus bersimulasi menunjukkan keputusan yang paling baik dalam kajian mereka. Mereka membuat keputusan bahawa kaedah mengesan berfungsi dengan baik apabila dimensi rendah, peratus pencemaran data yang rendah, jarak baki terpencil yang besar dan jumlah pelbagai titik kepulan yang besar.

BAB 3

METODOLOGI

3.1 Pengenalan

Pengiraan model teguh Spearman's rho adalah berdasarkan peringkat sampel min $R(\bar{x})$ dan $R(\bar{y})$. Peringkat sampel min mudah dipengaruhi oleh data ekstrim. Kajian ini akan mengutarakan satu penganggar teguh bagi pekali korelasi yang dipercayai mempunyai titik bendung pengaruh yang tinggi untuk mengurangkan pengaruh pekali korelasi oleh data ekstrim. Ini adalah berdasarkan prosedur gabungan regresi penganggar-M. Rousseuw telah mengungkapkan satu penganggar regresi yang teguh dengan kumungkinan titik bendung pengaruh yang paling tinggi. Contohnya $\varepsilon^* = 50$ peratus. Pekali korelasi dan penganggar teguh akan dibandingkan (Pearson's, Spearman's rho, Kendall's tau dan model teguh Spearman's rho berasaskan penganggar-M).

3.2 Kawasan Pensampelan

Dalam kajian ini, data yang digunakan adalah data sekunder mengenai kajian antara masa ulangkaji dengan markah Matematik yang diperolehi oleh pelajar.

3.3. Regresi Teguh

Menurut Huber (1964), penganggar-M ditakrifkan sebagai penganggar yang meminimumkan fungsi objektif, ie.

$$Q = \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) \tag{3.1}$$

dengan e_i adalah reja bagi cerapan ke-i, s sebagai penganggar skala. Fungsi berpengaruh ditakrifkan sebagai $\psi(t)$ adalah diperolehi dengan membezakan fungsi objektif $\rho(t)$ dan fungsi pemberat $w(t)$ adalah ditakrifkan seperti berikut:

$$w(t) = \frac{\psi(t)}{t} \tag{3.2}$$

keseimbangan antara keteguhan dan tahap efisien adalah sangat penting dalam memilih $\psi(t)$.

Fungsi objektif Huber:

$$\rho(t) = \begin{cases} \dfrac{t^2}{2} & , jika\ |t| \le c \\ c|t| - \dfrac{1}{2}c^2 & , jika\ |t| > c \end{cases} \tag{3.3}$$

dengan membezakan $\rho(t)$ akan memperolehi $\psi(t)$

$$\psi(t) = \begin{cases} t & , jika\ |t| \le c \\ c\ sign(t) & , jika\ |t| > c \end{cases} \tag{3.4}$$

dan $w(t)$ adalah

$$w(t) = \begin{cases} 1 & , jika \ |t| \leq c \\ \dfrac{c}{|t|} & , jika \ |t| > c \end{cases} \tag{3.5}$$

di mana c adalah pemalar tala. Nilai untuk pemalar tala, c dipilih sebagai 1.345 supaya dapat memperolehi 95 peratus keefisienan.

Bagi linear regresi ringkas,

$$y_i = \beta_0 + \beta_i x_i + \varepsilon_i \tag{3.6}$$

dengan $\hat{\beta}_j, j = 0,1$ sebagai parameter regresi dan ε_i ralat rawak, maka reja,

$$e_i = y_i - \hat{\beta}_o - \hat{\beta}_1 x_i \tag{3.7}$$

dengan $\hat{\beta}_0$ dan $\hat{\beta}_1$ masing-masing anggaran bagi β_0 dan β_1.

3.3.1 Perbandingan antara empat pekali korelasi teguh

Dalam kajian ini, tiga pekali korelasi teguh telah dikaji bagi membandingkan ketinggian bendung pengaruh dengan penganggar-M. Penganggar-M akan digabungkan ke dalam pekali korelasi model teguh Spearman's rho untuk mengkaji sifat titik bendung pengaruh pekali korelasi. Sampel data nyata akan dirosakkan sebanyak 10%, 20%, 30%, 40% dan 50% dan sensitiviti pekali korelasi kepada data ekstrim akibat pencemaran data nyata diperhatikan.

3.3.1.1 Pearson's

Persamaan pekali korelasi untuk pengiraan ρ adalah:

$$r_p = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]^{1/2}} \tag{3.8}$$

Penjelasan:

x_i adalah pemerhatian sampel data x yang ke i

y_i adalah pemerhatian sampel data y yang ke i

$\bar{x}$ adalah nilai min untuk sampel data x

$\bar{y}$ adalah nilai min untuk sampel data y

3.3.1.2 Spearman's rho

Persamaan pekali korelasi untuk pengiraan ρ adalah:

$$r_s = 1 - \frac{6D^2}{n(n^2 - 1)} \tag{3.9}$$

dengan

$$D^2 = \sum_{i}^{n} (r_{yi} - r_{xi})^2 \tag{3.10}$$

Penjelasan:

n adalah jumlah pemerhatian dalam sampel data

D adalah pembezaan peringkat antara nilai x dan y

D^2 adalah kuasa dua bagi pembezaan peringkat antara nilai x dan y

r_{xi} adalah peringkat bagi x_i

r_{yi} adalah peringkat bagi y_i

3.3.1.3 Kendall's tau

Persamaan pekali korelasi untuk pengiraan ρ adalah:

$$r_k = 1 - \frac{4Q}{n(n-1)} \tag{3.11}$$

Penjelasan:

n adalah jumlah pemerhatian dalam sampel data

Q adalah jumlah penyongsangan antara peringkat x dan y

3.1.1.4 Spearman's Rho Teguh

$$r_s = \frac{\sum_{w_i}[R(x_i) - \overline{R}(\overline{x}_{w_i})][R(y_i) - \overline{R}(\overline{y}_{w_i})]}{[\sum_{w_i}[R(x_i) - \overline{R}(\overline{x}_{w_i})]^2 \sum_{w_i}[R(y_i) - \overline{R}(\overline{y}_{w_i})]^2]^{1/2}} \tag{3.12}$$

Penjelasan:

$R(x_i)$ adalah peringkat bagi nilai x_i

$R(y_i)$ adalah peringkat bagi nilai y_i

$\overline{R}(\overline{x}_w)$ adalah peringkat bagi nilai x model teguh

$\overline{R}(\overline{y}_w)$ adalah peringkat bagi nilai y model teguh

Dalam model teguh Spearman's rho, pemberat penganggar-M dimasukkan iaitu w_i.

BAB 4

KEPUTUSAN DAN PERBINCANGAN

4.1 Pengenalan

Dapatan kajian yang akan dibincangkan dan dihuraikan termasuklah penganggar-M iaitu pekali korelasi yang mempunyai bendung pengaruh yang lebih baik dalam kewujudan data ekstrim.

4.2 Pemerihalan Data

Data yang digunakan adalah data hubungan antara markah yang dicapai dengan jumlah masa ulangkaji dalam seminggu di mana pembolehubah (y) mewakili masa ulangkaji dan pembolehubah (x) mewakili markah yang diperolehi oleh pelajar (Mokhtar Abdullah, 1994). Data tersebut dengan bendung pengaruh sehingga 50 peratus digunakan untuk memerhatikan nilai pekali korelasi. Data yang digunakan akan diberi dalam Jadual 4.1.

Jadual 4.1 Data masa ulangkaji melawan markah

Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)	Markah (y_i)
1	33	62
2	14	66
3	17	68
4	21	72
5	24	60
6	14	64
7	17	64
8	15	62
9	21	62
10	21	70
11	23	66
12	10	50
13	45	78
14	28	70
15	25	62
16	30	60
17	18	64
18	28	62
19	10	60
20	12	60
21	6	48
22	14	46
23	28	58
24	8	64
25	8	60

26	10	54
27	21	72
28	28	70
29	17	48
30	18	54
31	12	52
32	12	56
33	13	62
34	18	70
35	22	62
36	10	70
37	14	62
38	20	58
39	23	64
40	42	80

4.3 Analisis Kajian

Dalam konteks ini, pekali korelasi Pearson's, Spearman's rho, Kendall's tau dan penganggar-M telah digunakan untuk mengkaji sampel data sebanyak 40 dalam Jadual 4.1. Sampel data dalam Jadual 4.1 kemudiannya dicemarkan bermula dengan 10%. Seterusnya Jadual 4.3 4.4, 4.5 dan 4.6 masing-masing menunjukkan data dicemarkan sebanyak 20%, 30%, 40% dan 50%.

Jadual 4.2 Pencemaran data x_i sebanyak 10%

Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)
37	1400
38	2000
39	2300
40	4200

Jadual 4.3 Pencemaran data x_i sebanyak 20%

Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)
31	1200
32	1200
33	1300
34	1800
37	1400
38	2000
39	2300
40	4200

Jadual 4.4 Pencemaran data x_i sebanyak 30%

Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)
27	2100
28	2800
29	1700
30	1800
31	1200
32	1200
33	1300
34	1800
37	1400
38	2000
39	2300
40	4200

Jadual 4.5 Pencemaran data x_i sebanyak 40%

Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)
19	1000
20	1200
22	1400
26	100
27	2100
28	2800

29	1700
30	1800
31	1200
32	1200
33	1300
34	1800
37	1400
38	2000
39	2300
40	4200

Jadual 4.6 Pencemaran data x_i sebanyak 50%

Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)	Pelajar (i)	Masa ulangkaji (dalam jam/minggu) (x_i)
1	3.3	29	1700
2	1.4	30	1800
3	1.7	31	1200
4	2.1	32	1200
19	1000	33	1300
20	1200	34	1800
22	1400	37	1400
26	100	38	2000
27	2100	39	2300
28	2800	40	4200

4.4 Perbincangan

Berdasarkan Jadual 4.7 sebanyak 40 pemerhatian data diambil untuk mengkaji pekali korelasi Pearson's, Spearman's rho, Kendall's tau dan model pekali korelasi berasaskan penganggar-M memberikan keputusan seperti yang dijadualkan dalam Jadual 4.7. Jadual 4.7 menunjukkan nilai pekali korelasi Pearson's (rp), Spearman's rho (rs), Kendall's tau (rk) dan pekali korelasi berasaskan penganggar-M (rw) apabila pemerhatian yang 'baik' digantikan dengan 10%, 20%, 30%, 40% dan 50% data 'rosak'.

Jadual 4.7 Nilai pekali korelasi rp, rs, rk dan rw bagi 40 pemerhatian data

Pencemaran (%)	**Pearson's (rp)**	**Spearman's Rho (rs)**	**Kendall's Tau (rk)**	**Model teguh Spearman's Rho (rw)**
0	0.5606168	0.4270063	0.3000000	0.7608819
10	0.2861574	0.3704834	0.2564103	0.8344376
20	0.2414087	0.2374380	0.1615385	0.8001460
30	0.2180830	0.2046854	0.1307692	0.7775156
40	0.1253826	-0.0306818	-0.0461539	0.7989837
50	0.1242005	-0.0907950	-0.0897436	0.7856444

Dalam Jadual 4.7, kolum pertama adalah nilai pecemaran, i.e baik, 10%, 20%, 30%, 40% dan 50%. Kolum kedua, tiga dan empat adalah pekali korelasi masing-masing bagi Pearson's, Spearman's rho dan Kendall's tau manakala kolum terakhir ialah model teguh Spearman's rho berasaskan penganggar-M.

Jadual tersebut menunjukkan bahawa apabila data adalah baik (tanpa pencemaran), nilai pekali korelasi masing-masing adalah 0.5606168, 0.4270063 dan 0.3 manakala nilai pekali korelasi model teguh Spearman's rho berasaskan penganggar-M ialah 0.7608819. Ini menunjukkan bahawa model teguh Spearman's rho berasaskan penganggar-M adalah lebih baik berbanding dengan pekali korelasi yang lain.

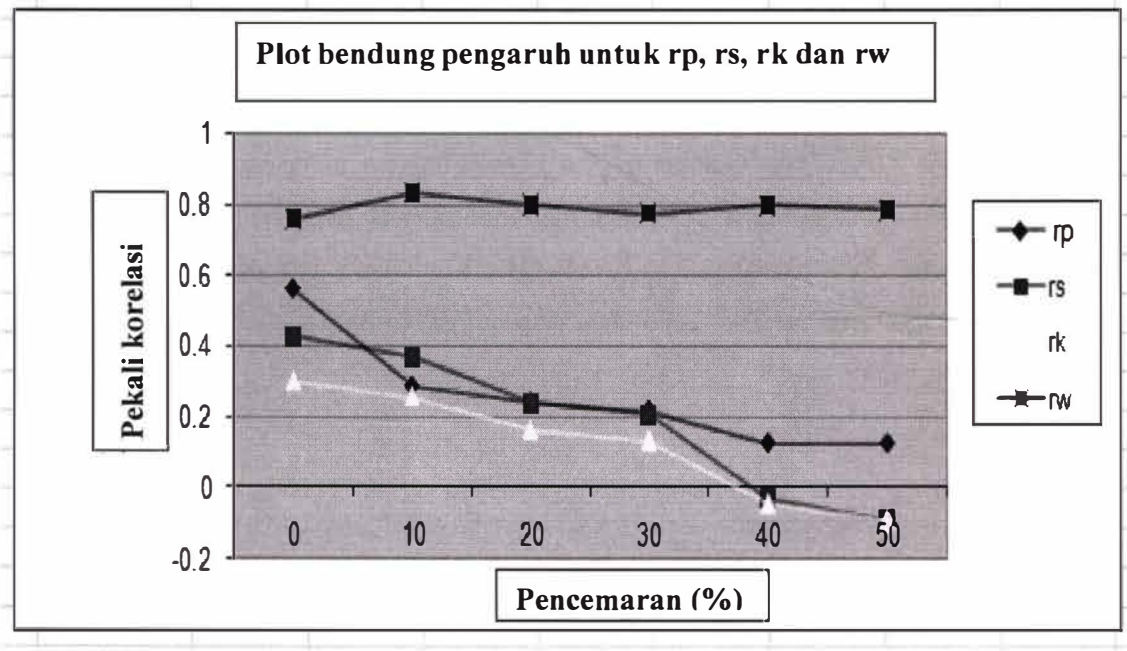
Apabila berlaku pencemaran data sebanyak 10 peratus, iaitu terdapat sebanyak 4 data 'rosak', nilai pekali korelasi bagi Pearson's adalah 0.2861574, Spearman's rho, 0.3704834 dan Kendall's tau, 0.2564103 manakala nilai pekali korelasi Spearman's rho berasaskan penganggar-M adalah 0.8344376. Dengan ini, dapat dilihatkan bahawa model teguh Spearman's berasaskan penganggar-M masih adalah baik dibandingkan dengan pekali korelasi yang lain.

Apabila terdapat 8 data nyata yang dirosakkan (pencemaran 20 peratus), Jadual 4.7 menunjukkan nilai pekali korelasi masing-masing adalah 0.2414087, 0.2374380, 0.1615385 manakala nilai model teguh Spearman's rho pula memberi nilai 0.8001460 apabila pencemaran data sebanyak 20 peratus berlaku. Walaupun pencemaran data telah bertambah sehingga 20 peratus, model teguh Spearman's rho berasaskan penganggar-M masih adalah terbaik berbanding dengan pekali korelasi yang lain dalam kes ini.

Seterusnya, kita memerhatikan keputusan untuk pencemaran data sebanyak 30 peratus. Nilai pekali korelasi Pearson's adalah 0.2180830, nilai Spearman's rho diberi 0.2046854 dan nilai pekali korelasi Kendall's lebih rendah berbanding Pearson's dan Spearman's, iaitu hanya 0.1307692. Untuk Spearman's rho yang telah digabungkan dengan penganggar-M, iaitu model teguh Spearman's, nilai pekali korelasinya adalah 0.7775156. Sekali lagi dapat dilihat bahawa model teguh Spearman's rho masih memberi nilai yang terbaik berbanding dengan pekali korelasi lain walaupun nilai pekali korelasinya telah menurun.

Dalam keadaan pencemaran data sebanyak 40 peratus, diperhatikan bahawa nilai pekali korelasi masing-masing adalah 0.1253826, -0.0306818, dan -0.0461539. Nilai pekali korelasi Spearman's rho dan Kendall's tau telah menurun kepada nilai negatif. Model teguh Spearman's rho pula memberi nilai 0,7989837 di mana nilainya telah meningkat berbanding dengan nilainya dalam pencemaran data sebanyak 30 peratus dan pekali korelasi ini masih memberi nilai yang baik berbanding dengan pekali korelasi yang lain.

Akhirnya, di mana sampel data telah dirosakkan separuh, pencemaran sebanyak 50 peratus atau kehadiran 20 data 'rosak', pekali korelasi Pearson's memberi nilai 0.1242005. Pekali korelasi Spearman's rho dan Kendall's tau mengekalkan nilai pekali korelasinya dalam negatif, iaitu -0.0907950 dan -0.0897436 manakala pekali korelasi Spearman's rho berasaskan penganggar-M telah menurun sedikit memberi nilai 0.7856444. Walaupun pencemaran data telah meningkat sehingga 50 peratus, model teguh Spearman's rho masih memberi nilai yang lebih baik berbanding dengan pekali korelasi yang lain.



Rajah 2: Plot bendung pengaruh rp, rs, rk dan rw bagi 40 pemerhatian data

Daripada Rajah 2, diperhatikan bahawa nilai rp berbeza jauh daripada nilai asal data 'baik' apabila peratus pencemaran data berlaku semakin tinggi. Apabila peratus pencemaran data 'baik' semakin meningkat, nilai pekali korelasi rp, rs dan rk telah menurun dan nilai bagi pekali korelasi rs dan rk telah berubah daripada nilai positif kepada nilai negatif manakala untuk rp dan rw kedua-duanya tetap kekal dalam nilai positif. Adalah didapati bahawa pekali korelasi berperingkat rs dan rk yang mengalami bendung pengaruh pada masa yang sama iaitu apabila enam belas data nyata telah dicemarkan menjadi data 'rosak' walaupun pada mula-mulanya pekali korelasi ini adalah teguh. Kewujudan data 'rosak' telah menyebabkan pekali korelasi rs dan rk disifatkan sebagai pekali korelasi yang mempunyai bendung pengaruh yang rendah dan sangat sensitif kepada pengaruh pencemaran data berbanding dengan pekali korelasi berperingkat rp. Nilai pekali korelasi rs dan rk bertukar menjadi nilai negatif apabila keteguhan pekali korelasi ini semakin kurang dengan kehadiran bilangan data 'rosak' yang semakin banyak. Hubungan x dan y menjadi semakin jauh.

Walau bagaimanapun, pekali korelasi berasaskan penganggar-M mempunyai bendung pengaruh yang tinggi berbanding dengan pekali korelasi rp, rs dan rk kerana apabila data ekstrim berlaku, nilai pekali korelasinya menurun dan menaik, tetapi nilai pekali korelasi masih kekal dalam nilai positif. Pengaruh pencemaran data nyata kepada pekali korelasi berasaskan penganggar-M adalah tidak besar. Pekali korelasi ini tidak sensitif walaupun terdapat kewujudan data eksrim. Penganggar-M sebagai pemberat model teguh Spearman's membantu mengekalkan keteguhan pekali korelasinya dengan mengawal bendung pengaruh pekali korelasi walaupun terdapat data 'rosak'.

BAB 5

KESIMPULAN DAN CADANGAN

5.1 Kesimpulan

Daripada keputusan yang diperolehi daripada penyelikan ini, didapati bahawa untuk sampel data yang 'baik' atau apabila sampel data tidak mengalami pencemaran data, pekali korelasi Pearson's menunjukkan nilai pekali korelasi yang kurang baik daripada model teguh Spearman's rho (pekali korelasi berdasarkan penganggar-M). Pekali korelasi Pearson's adalah kurang sensitif kepada kehadiran data 'rosak' kerana pekali korelasi Pearson's mempunyai bendung pengaruh yang lebih tinggi berbanding dengan pekali korelasi Spearman's rho dan Kendall's tau. Pekali korelasi berperingkat iaitu Spearman's rho dan Kendall's tau mempunyai ketinggian bendung pengaruh yang lebih kurang sama. Kedua-dua pekali korelasi ini sangat sensitif terhadap data ekstrim berbanding pekali korelasi Pearson's. Walau bagaimanapun, keteguhan pekali korelasi ini semakin menurun dan mencapai kerosakan apabila pencemaran data nyata semakin meningkat. Keputusan menunjukkan bahawa pekali korelasi yang berasaskan penganggar-M adalah pekali korelasi yang memberi nilai yang baik walaupun pencemaran data berlaku atau apabila terdapat data ekstrim dalam sampel data.

Keputusan yang didapati telah menunjukkan bahawa pekali korelasi yang telah diubahsuai, iaitu pekali korelasi Spearman's rho yang digabungkan dengan penganggar-

M sebagai pemberat melambangkan satu pembaikan yang berpotensi terhadap keteguhan berbanding dengan pekali korelasi Pearson's, Spearman's rho dan Kendall's tau apabila pencemaran berlaku terhadap pembolehubah x dan y dalam sampel data. Pekali korelasi berasaskan penganggar-M mempunyai bendung pengaruh yang tinggi dan teguh terhadap kewujudan data ekstrim. Pekali korelasi ini tidak sensitif terhadap kewujudan data ekstrim dalam satu sampel.

5.2 Cadangan

Dalam kajian ini, hanya nilai pekali korelasi untuk data nyata yang dikajikan. Bendung pengaruh pekali korelasi berasaskan penganggar-M hanya dipertimbangkan dalam data nyata yang mempunyai data ekstrim. Adalah dicadangkan bahawa sifat bendung pengaruh pekali korelasi Pearson's, Spearman's rho, Kendall's tau dan model teguh Spearman's rho dikaji secara lanjut dengan memerhatikan nilai kecondongan, variant, kesalahan piawai, min kesalahan piawai dan punca kuasa dua min kesalahan piawai pekali korelasi tersebut. Data yang diambil kira boleh digenerasikan sebanyak 500 kali percubaan. Sebanyak 50 data dan 100 data diambil dengan $\rho = 1$. Data yang digenerasikan juga dicemarkan seperti dalam kajian yang dijalankan dengan pencemaran data sebanyak 10%, 20%, 30%, 40% dan 50% untuk memerhatikan sifat bendung pengaruh.

Di samping itu, kajian terhadap keteguhan pekali korelasi juga boleh dilihat dengan menggunakan penganggar-MM yang dipercayai mempunyai sifat keteguhan dan bendung pengaruh yang lebih baik.

RUJUKAN

Chernick, M. R. 1982. The influence function and its application to data validation. *America Journal of Mathematical and Management Science* (2): 263-288.

Donoho, D. L. & Huber P. J. 1983. The notion of breakdown point, in. P. Bickel, K. Doksum & J. L. Hodges, Jr (Eds) *A Festschrift for Erich L. Lehmann* (Belmont, CA, Wadsworth).

Estelberger, W. Reibnegger, G. 1995. The rank correlation coefficient: an additional aid in the interpretation of laboratory data. *Clinica Chimica Acta* (239): 203-207.

Huber, P. J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* (35): 73-101.

Massart, D. L., Kaufman, L., Rousseeuw, P. J., Leroy, A. M. 1986. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta* (187): 171-179.

Mokhtar Bin Abdullah. 1990. On a robust correlation coefficient. *The Statistician* (39): 455-460.

Mokhtar Abdullah. 1994. *Analisis regresi.* Kuala Lumpur: Dewan Bahasa dan Pustaka Kementerian Pendidikan Malaysia.

Ortiz, M. C., Sarabia, L. A., Herrero, A. 2005. Robust regression techniques a useful alternative for the detection of outlier data in chemical analysis. *Talanta* (70):499-512.

Wikipedia. 2009. Robust Statistics. http://en.wikipedia.org/wiki/Robust_Statistic [20 Februari 2009].

Wikipedia. 2009. Outliers. http://en.wikipedia.org/wiki/Outliers [22 Februari 2009].

Wisnowski, J. W., Montgomery, D. C., Simpson, J. R. 2001. A comparative analysis of multiple outlier detection regression model. *Computational Statistic & Data Analysis* (36): 351-382.

BIODATA PENULIS

Nama	: Neo Siew Ping
Alamat tetap	: 55, Jalan 3/2, Taman Sri Kluang, 86000 Kluang, Johor
Nombor telefon	: 0177113114
Email	: nsiewping86@yahoo.com
Tarikh lahir	: 10 Mac 1986
Tempat lahir	: Melaka
Kewarganegaraan	: Malaysia
Bangsa	: Cina
Jantina	: Perempuan
Agama	: Buddha
Pendidikan	: Sarjana Muda Sains (Matematik Komputasi)
Anugerah	: Senarai Dekan untuk semester Julai 06/07, Disember 06/07, Julai 07/08, Disember07/08, Julai 08/09

PEKALI KORELASI: PENDEKATAN TEGUH - NOE SIEW PING