Jurnal Penelitian Sains ; hal 59 - 73
No.4. Oktober 1998 ISSN: 1410-7058

BEBERAPA ALTERNATIF KOEFISIEN KORELASI SEDERHANA UNTUK DATA TERPENCIL

Robinson Sitepu
Jurusan Matematika FMIPA Universitas Sriwijaya

ABSTRAK

Penelitian ini bertujuan untuk menentukan koefisien determinasi (R^2) untuk data terpencil atau ekstrim, karena koefisien determinasi (R^2) merupakan salah satu ukuran untuk menyatakan kecocokan model. Apabila terdapat data yang terpencil atau ekstrim maka koefisien determinasi tersebut tidak dapat dipergunakan. Dengan metode kuadrat median terkecil akan dimodifikasi beberapa koefisien determinasi untuk data yang terpencil

PENDAHULUAN

Analisis regresi berkenaan dengan studi ketergantungan suatu (variabel tak bebas) pada satu atau lebih variabel lainnya (variabel bebas/variabel yang menjelaskan) dengan tujuan untuk memperkirakan dan atau meramalkan nilai rata-rata variabel tak bebas apabila nilai variabel bebas diketahui. Analisa regresi sederhana atau dua variabel didalam mempelajari ketergantungan satu variabel pada hanya satu variabel yang menjelaskan, tetapi kalau kita mempelajari ketergantungan satu variabel pada lebih dari satu variabel yang menjelaskan maka analisisnya disebut analisis regresi majemuk (multiple regression analysis).

Hubungan fungsional antara variabel-variabel telah diuraikan dalam regresi linier sederhana. Didalam regresi linier ditinjau bagaimana hubungan antara variabel bebas dan tak bebasnya. Permasalahan berikutnya dirasakan perlu ialah berapa kuat variabel-variabel tersebut saling berhubungan. Dengan kata lain perlu ditentukan derajat hubungan antara variabel-variabel tersebut. Statistika yang membahas tentang derajat hubungan antara variabel-variabel dikenal dengan Analisa Korelasi. Ukuran yang dipakai untuk mengetahui derajat hubungan, terutama untuk data kuantitatif dinamakan Koefisien Korelasi. Yang harus diperhatikan bahwa koefisien korelasi memperlihatkan hubungan linier antara x dan y, hubungan yang tidak linier tidak berlaku bagi koefisien korelasi.

Dalam regresi linier sederhana koefisien korelasi dilambangkan dengan r. Untuk menghitung koefisien korelasi sederhana begitu banyak rumus yang dapat dipergunakan.

Beberapa penulis: Cohen and Cohen (1983), Draper and Smith (1981), Edwards (1976), Goulden (1952), Guilford and Fructher (1957), Kendall and Stuart (1947), Kerlinger and Pedhazur (1973), Kleinbaum, Kuper and Muller (1988), Kualseth (1985), Neter and Kutner (1976), Rodgers and Nicewander (1988), Schilling (1984), Snedecor and Cohran (1956), Weisberg (1980) dan Williams (1959) memaparkan beberapa alternatif untuk menghitung koefisien korelasi sederhana.

a. Kovarians

$$r_{yx} = \frac{S_{xy}}{S_x S_y}$$

b. Koefisien Regresi

$$r_{yx} = b_{yx} \frac{S_x}{S_y} \text{ atau } r_{yx} = b_{xy} = \frac{S_y}{S_x}$$

c. Hasil kali Koefisien Regresi

$$r_{yx} = \sqrt{b_{yx} b_{xy}}$$

d. Koefisien Jalur

$$r_{yx} = \sqrt{p_{yx} p_{xy}}$$

e. Rasio dari dua varians

$$r_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \text{ atau } r_{yx} = \sqrt{\frac{JK_{Regresi}}{JK_{Total}}}$$

f. Fungsi dari Statisti Uji

$$r_{yx} = \frac{t_{1-\frac{\alpha}{2}:n-2}}{\sqrt{t^2{}_{1-\frac{\alpha}{2}:n-2} + n - 2}}$$

g. Varians dari variabel yang standard

$$r_{yx} = 1 - \frac{1}{2} S^2{}_{y-x} \quad \text{atau} \quad r_{yx} = \frac{1}{2} S^2{}_{y+x} - 1$$

h. Hasil perkalian dua variabel standard

$$r_{yx} = \frac{1}{n} \sum_{i=1}^{n} Z_x Z_y$$

i. Selisih antara variabel yang distandardkan

$$r_{yx} = 1 - \frac{1}{2}\left\{\frac{\sum_{i=1}^{n}(Z_x - Z_y)^2}{n}\right\}$$

j. Variabel Y dan $\hat{Y}$

$$r_{yx} = \frac{\sum_{j=1}^{n}(Y_j\hat{Y}_i - n\overline{Y})^2}{\sqrt{(\sum_{i=1}^{n}Y_i^2 - n\overline{Y}^2)(\sum_{i=1}^{n}\hat{Y}_i^2 - n\overline{Y}^2)}}$$

k. Bentuk Deviasi

$$r_{yx} = \frac{\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}d_i^2}{2\sqrt{(\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}$$

l. Nilai taksiran $\hat{Y}$ dengan rata-rata $\hat{Y}$

$$r_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

m. Nilai taksiran $\hat{Y}$ dengan rata-rata Y

$$r_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

n. Nilai residu

$$r_{yx} = \sqrt{1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

o. Nilai residu dengan rata-rata nilai residu

$$r_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \overline{e})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

p. Jumlah kuadrat Y terhadap $\hat{Y}$ dengan jumlah kuadrat y

$$r_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(Y - \hat{Y})^2}{\sum_{i=1}^{n} Y_i^2}}$$

q. Jumlah kuadrat $\hat{Y}$ dengan jumlah kuadrat y

$$r_{yx} = \sqrt{\frac{\sum_{i=1}^{n} \hat{y}_i^{\,2}}{\sum_{i=1}^{n} Y_i^{\,2}}}$$

Nilai-nilai r_{yx} untuk cara ke l, ke m, ke n, ke o dan ke p untuk model regresi yang mempunyai intercept, sedangkan cara ke q untuk model regresi yang tidak mempunyai intercept.

r. Moment

$$r_{yx} = \frac{m_{yx} - m_x m_y}{\sqrt{(m_{x^2} - m_x^{\,2})(m_{y^2} - m_y^{\,2})}}$$

s. Aturan Balon

$$r_{yx} = \sqrt{1 - (\frac{h}{H})^2}$$

t. Sudut antara dua vektor acak

r_{yx} = Cos □

Titik Terpencil

Asumsi yang sangat penting dalam analisis regresi adalah model yang dibangun berlaku untuk setiap titik dari data yang dimiliki, artinya apabila diperoleh sampel atas variabel acak, maka diasumsikan semua titik data itu mengikuti model yang telah ditentukan, apabila data seperti itu ditemukan, maka titik seperti ini dinamakan titik pencilan. Sehingga kajian mengenai titik pencilan ini perlu dilakukan dalam setiap analisis regresi.

Setelah menggambarkan hasil pengamatan dalam diagram pencar dan sudah bisa menentukan polanya merupakan pola garis lurus, maka langkah selanjutnya adalah

Jurnal Penelitian Sains ; hal 59 - 73
No.4. Oktober 1998 ISSN: 1410-7058

memperhatikan apakah pada diagram pencar ada titik yang letaknya terpencil. Pertanyaan yang muncul apakah titik tersebut merupakan titik terpencil atau tidak. Apabila titik tersebut merupakan sebuah titik terpencil, maka titik itu harus dikeluarkan dari analisis. Cara mengeluarkannya tidak boleh begitu saja tetapi harus menggunakan test for outlier in regression analysis.

Perumusan pada pengujian titik terpencil dalam analisis regresi berbentuk :

H_0 : Titik tersebut bukan merupakan titik terpencil

H_1 : Titik tersebut merupakan titik terpencil

Statistik uji yang digunakan adalah:

$$t = \frac{Y - \hat{Y}}{S_{Y-\hat{Y}}}$$

statistik di atas mengikuti distribusi student t dengan v = n – 2 (Weisbeerg, 1972).

$$S^2{}_{Y-\hat{Y}} = S^2{}_{Y|X}\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\left(\sum_{i=1}^{n} X_i{}^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)}\right]$$

dan $S^2{}_{Y|X} = \frac{n-1}{n-2}(S^2{}_Y - b_1^2 S_X^2)$ dalam hal ini :

$$S^2{}_Y = \frac{\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}}{n-1} \quad \text{dan} \quad S^2{}_X = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}$$

apabila diambil dalam bentuk deviasi terhadap rata-rata, yaitu

$$S_Y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n-1}; S_X^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} \text{ dan } b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

maka $S^2_{Y|X}$ dapat juga dihitung dengan menggunakan rumus

$$S^2_{Y|X} = \frac{1}{n-2}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right]$$

Kriteria penolakan, tolah H_0 jika $|t| > t_{1-\square/2;n-2}$, maka titik yaang mencurigakan dianggap sebagai titik terpencil dan harus dikeluarkan dari analisis.

METODOLOGI

Data yang dipergunakan dalam contoh pemakaian ini, diambil Naval Research Laboratory Anacostia Station, Washington, DD.C (dalam Croxton and Cowden, 1961), yaitu mengenai hubungan antara Hardness dengan Tensile Strength.

HASIL DAN PEMBAHASAN

Dari hasil penelitian ini diperoleh perumusan untuk menghitung koefisien korelasi sederhana untuk data terpencil.

Koefisien Korelasi untuk Data Terpencil

Rousseeuw dan Leroy (1987) memaparkan Koefisien Korelasi dalam regresi linier Median terkecil adalah:

$$R_m^2 = \left(1 - \frac{\text{med}\,|r_i|}{\text{med}\,y_i}\right)^2$$

apabila persamaan regresinya tidak mempunyai konstanta (intercept), jika mempunyai intercept, maka koefisien korelasinya adalah:

$$R_m^2 = \left(1 - \frac{\text{med}\,|r_i|}{\text{mad}\,Y_i}\right)^2$$

dengan :

$$\text{mad}\,(Y_i) = \text{med}\,|(Y_i - \text{med}\,Y_i)|$$

$$r_i = (Y_i - \text{med}\,Y_i)$$

Alternatif Koefisien Korelasi untuk Data Terpencil

Rousseeuw & Leroy (1987) memaparkan bahwa apabila tanda jumlah (Σ) dan nilai rata-rata pada semua rumus yang diperoleh melalui metoda kuadrat terkecil diganti dengan nilai mediannya, maka diperoleh perumusan yang didasarkan pada metode kuadrat median terkecil.

Berdasarkan penjelasan tersebut, maka berikut ini diberikan alternatif untuk menghitung koefisien korelasi sederhana yang ditentukkan berdasarkan metode kuadrat terkecil, yaitu:

1. Pearson

$$r_{yx} = \frac{\mathrm{Med}(X_i - \mathrm{Med}\,X_i)(Y_i - \mathrm{Med}\,Y_i)}{\sqrt{[\mathrm{Med}\,|\,(X_i - \mathrm{Med}\,X_i)\,|]^2[\mathrm{Med}\,|\,(Y_i - \mathrm{Med}\,Y_i)\,|]^2}}$$

1. Kovarians

$$r_{yx} = \frac{s_{YX}}{s_X s_Y}$$

dalam hal ini

$$s_{YX} = \frac{\mathrm{Med}\,|\,(X_i - \mathrm{Med}X_i)(Y_i - \mathrm{Med}Y_i)\,|}{n-1}$$

$$s_X = \frac{\sqrt{[\mathrm{Med}\,|\,(X_i - \mathrm{Med}\,X_i)\,|]^2}}{n-1}$$

dan

$$s_Y = \frac{\sqrt{[\mathrm{Med}\,|\,(Y_i - \mathrm{Med}\,Y_i)\,|]^2}}{n-1}$$

2. Koefisien regresi

$$r_{yx} = b_{yx}\frac{s_X}{s_Y}$$

atau

$$r_{yx} = b_{xy}\frac{s_Y}{s_X}$$

dalam hal ini

$$b_{YX} = \frac{\text{Med}\,|\,(X_i - \text{Med}\,X_i)(Y_i - \text{Med}\,Y_i)\,|}{[\text{Med}\,|\,(X_i - \text{Med}\,X_i)\,|]^2}$$

dan

$$b_{XY} = \frac{\text{Med}\,|\,(X_i - \text{Med}\,X_i)(Y_i - \text{Med}\,Y_i)\,|}{[\text{Med}\,|\,(Y_i - \text{Med}\,Y_i)\,|]^2}$$

3. Hasil Kali Koefisien Regresi

$$r_{yx} = \sqrt{b_{YX}b_{XY}}$$

4. Rasio Dua Varians

$$r_{YX} = \sqrt{\frac{[\text{med}\,|\,(Y_i - \text{med}\hat{Y}_i\,|]^2}{[\text{med}\,|\,(Y_i - \text{med}Y_i\,|]^2}}$$

atau

$$r_{YX} = \sqrt{\frac{\text{JK Regresi}}{\text{JKTotal}}}$$

JK Regresi = $b_{YX}\text{Med}[(X_i\text{-Med } X_i)(Y_i\text{-Med } Y_i)]$

JK Total = $[\text{Med}|(Y_i\text{-Med}Y_i)|]^2$

5. Variabel Y dengan $\hat{Y}$

$$r_{yx} = \frac{[\text{Med} \mid (Y_i - \text{Med } Y_i)(\hat{Y}_i - \text{Med } Y_i)]}{\sqrt{[\text{Med} \mid (X_i - \text{Med } X_i) \mid]^2 [\text{Med} \mid (Y_i - \text{Med } Y_i) \mid]^2}}$$

6. Nilai Taksiran $\hat{Y}$ dengan Median Y

$$r_{YX} = \sqrt{\frac{[\text{med} \mid (\hat{Y}_i - \text{med}Y_i \mid]^2}{[\text{med} \mid (Y_i - \text{med}Y_i \mid]^2}}$$

Pengolahan data

Data yang diperoleh, selanjutnya diolah untuk mendapatkan nilai koefisien korelasi berdasarkan rumus-rumus yang sudah dimodifikasi, dan dalam hal ini sebagai variabel bebas adalah Hardness (X) sedangkan variabel tak bebas adalah Tensile Strength (Y).

Tabel 1. Data Hasil Pengukuran Mengenai Hardness dan Tensile Strength

Hardness	Tensile Strength	Hardness	Tensile Strength	Hardness	Tensile Strength	Hardness	Tensile Strength
144	70.00	212	106,9	156	72,75	171	79,15
183	91.40	169	81,8	149	74,25	160	76
154	78.75	173	84,45	199	100,4	164	83,45
146	69.90	159	7,15	167	76,95	180	85
210	104.75	176	88,9	165	79,15	156	76,45
156	79.50	165	80,8	176	84,6	212	102
188	93.15	171	94,81	162	78,85	151	72,5
146	70.25	159	78	168	81,4	158	80,8
146	72.15	168	80,6	181	89,25	159	75,3
202	95.00	174	81,25	164	79,35	154	76,95
159	75.00	146	72,25	168	79,5	181	88,65

194	98.15	173	81,65	164	81,1	153	74,65
149	74.10	160	79,75	170	83,75	183	90,6
178	83.25	166	78	158	77,9	167	77,25
152	75.75	173	80,1	194	91	163	77,25
166	77.65	157	75,4	177	86,35	179	88,5
151	75.65	162	78,25	171	82,25	188	89,9
162	79.65	187	84,9	171	86,7	150	76,9
148	71.90	160	79	169	82,25	183	89
147	72.90	160	78,25	132	86,5	168	85,75
150	73.75	170	83,1	160	80,25	170	77,65
150	69.75	161	80,25	150	71	162	79,6
156	73.15	165	80,6	177	85,25	171	85,15
203	100.00	176	84,75	165	79,65	187	91,9
151	73.90	163	77,6	171	84,95	152	73,25
155	72.90	164	76,15	176	85,4	163	81,1
152	71.10	173	86,25	192	94,5	189	88,5
154	75.40	164	82,25	173	82,5	196	100,25

Dari hasil perhitungan diperoleh koefisien korelasi :

a. Untuk data yang lengkap dan menggunakan metode kuadrat terkecil

$r_{YX} = 0,9463007$

b. Untuk data yang titik terpencil sudah dikeluarkan

$r_{YX} = 0,955472183$

c. Untuk data yang lengkap dan menggunakan metode kuadrat median terkecil:

1. Rousseeuw & Leroy

$R_m = 0,921584$

2. Pearson

$r_{YX} = 0,954383$

3. Kovarians

$r_{YX} = 0{,}954383$

4. Koefisien Regresi

$r_{YX} = 0{,}954383$

5. Hasil kali Koefisien Regresi

$r_{YX} = 0{,}954383$

6. Rasio dari dua varians

$r_{YX} = 0{,}954383$

7. Variabel Y dengan $\hat{Y}$

$r_{yx} = 0{,}954369$

KESIMPULAN

Berdasarkan hasil perhitungan pada contoh pemakaian maka dapat diambil kesimpulan sebagai berikut:

1. Apabila terdapat data (titik) pencilan, maka titik tersebut tidak dikeluarkan, tetapi tetap dipergunakan dalam analisis.
2. Dari rumus-rumus yang dimodifikasi ternyata memberikan hasil koefisien korelasi yang lebih besar dari koefisien korelasi yang disarankan oleh Rousseuw & Leroy.

DAFTAR PUSTAKA

Chatillon, G., 1984, *The Balon Rules For a Raugh Estimate of the Correlation Coeficient*, The American Statistician, February, Vol 38, No. 1.

Cohen, J. and Cohen, P., *Analysis for the Behavorial Sciences*, Second edition, Lawrence Erlbaum Associate, Inc.

Draper, N., Smith, H., 1981, *Applied Regression Analysis*, Second edition, John Willey and Sons, Inc.

Edwards, A.L., 1976, *An Introduction to Linear Regression and Correlation*, W.H. Freeman and Company San Francisco.

Goulden, C.H., 1952., *Analysis Methods of Statistical*, Second Edition, John Wiley and Sons, Inc., Proprietor.

Kerlinger, F.N., Pedhazur, E.J., 1973., *Multiple Regression In Behavior*, Holt Rinehart and Winston, Inc, New York.

Kendall, M.G. and Stuart, A., 1947, *The Advanced Theory of Statistic*, Second Edition, Hafer Publishing Comppany, New York.

Kleinbaum, D.G., Kuper, L.L., and Muller, K.E., 1988, *Applied Regression and Other Multivariable Methods*, Second Edition, PWS-KENT PublishingCompany, Boston.

Neter, J., Wasserman, W. and Kutner, M.H., 1976, *Applied Linear Statistical Model, Regeression Analysis of Variance, and Experimental Designs*, Second Edition, Richard D. Irwin, Inc.

Rousseeuw, P.T., and Leroy, A.M., 1987, *Robust Regression and Outlier Detection*, John Willey N Sons, Inc.

Sudjana, *Metoda Statistika, E*disi ke 4; Penerbit Tarsito Bandung.

Snedecor, G.W., and Cohran, W.G., 1956, *Statistical Methods*, Fifth Edition, The Iowa State University Press.