Masalah Pencilan Dalam Regresi Linier Sederhana

Yeyen Muniar[1], Sigit Nugroho[2], dan Fachri Faisal[2]
[1]Alumni Jurusan Matematika Fakultas MIPA Universitas Bengkulu
[2]Staf Pengajar Jurusan Matematika Fakultas MIPA Universitas Bengkulu

ABSTRAK

Analisis regresi adalah teknik statistika yang digunakan untuk mencari hubungan fungsional dari satu atau beberapa variabel yang mempengaruhi (independent variable) terhadap satu variabel yang dipengaruhi (dependent variable). Tujuan dari penelitian ini adalah: 1). Untuk mengkaji pencilan dalam regresi linier sederhana. 2). Mempelajari pengaruh pencilan dalam regresi linier sederhana. 3). Memberikan penjelasan tentang cara mengatasi pencilan. Analisis data dilakukan dengan cara membangkitkan data yaitu dengan membangkitkan data simulasi dari program Microsoft Excel. Data baru yang mengakibatkan penurunan koefisien korelasi yang "cukup berarti" dapat dikategorikan sebagai data pencilan. Hasil penelitian menunjukkan adanya pengaruh pencilan terhadap sudut dan jarak yang dibentuk oleh garis regresi.

Kata Kunci : *Regresi Linier Sederhana, Pencilan, Analisis Regresi, Metode Kuadrat Terkecil, Koefisien Korelasi.*

PENDAHULUAN

Analisis regresi adalah teknik statistika yang digunakan untuk mencari hubungan fungsional dari satu atau beberapa variabel yang mempengaruhi *(independent variable)* terhadap satu variabel yang dipengaruhi *(dependent variable)*. Hubungan antara variabel bebas dengan variabel tak bebas tersebut merupakan hubungan linier.

Misalkan nilai suatu variabel X diduga mempengaruhi nilai variabel lain Y, dan kemudian perubahan nilai X digunakan menduga perubahan nilai Y. Dalam hal ini X dinamakan prediktor dan Y disebut respon.

Dalam regresi linier sederhana, bentuk fungsi f didekati oleh persamaan garis lurus. Model regresi yang paling sederhana adalah regresi linier dengan satu variabel penjelas. Penyelesaiannya menjadi lebih sederhana lagi dengan asumsi bahwa hanya variabel tak bebas Y yang bersifat sebagai variabel acak, sedangkan variabel bebas X dianggap sebagai variabel tetap.

Apabila hubungan antara X dan Y benar-benar bersifat linier maka hubungan ini dapat dirumuskan sebagai :

$$E(Y) = \beta_0 + \beta_1 X \qquad (1)$$

yang mana $E(Y)$ adalah nilai tengah atau nilai harapan Y, parameter β_0 adalah jarak perpotongan garis regresi dengan sumbu y dan β_1 merupakan parameter kemiringan garis terhadap sumbu x.

Dari uraian diatas, maka dalam regresi linier sederhana bentuk hubungan antara X_i dan Y_i dirumuskan oleh

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (2)$$

Analisis regresi sederhana memberikan sebuah persamaan yang dapat dipakai untuk mengestimasi *(estimate)* atau memprakirakan *(predict)* nilai sebuah variabel dari sebuah nilai tertentu lainnya. Jadi. regresi sederhana menghubungkan dua buah variabel, yaitu sebuah variabel bebas dan variabel tak bebas. Dalam regresi linier sederhana, persamaan taksiran *(estimating equation)* memiliki sebuah grafik yang merupakan sebuah garis lurus. Persamaan taksiran ditentukan dengan melakukan perhitungan atas data pengamatan (Bowen & Starr, 1982).

Dalam analisis regresi, hubungan antara variabel bebas dengan variabel tak bebas merupakan hubungan yang linier.

Analisis regresi linier sederhana mempunyai asumsi-asumsi sebagai berikut :

1. Variabel bebas dan variabel tak bebas mempunyai hubungan linier.
2. Persamaan liniernya dinyatakan dengan: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$; $i = 1,2,...,n$ (3)
3. Variabel tak bebas merupakan variabel random kontinu, sedangkan bebasnya merupakan serangkaian nilai yang ditentukan atau diketahui dan bukan random.
4. Variansi dari distribusi kondisional variabel tak bebas untuk berbagai nilai variabel bebas tertentu, semuanya sama (konstan) atau $\sigma_{\varepsilon_i}^2 = \sigma^2$ untuk setiap *i =1, 2, ..., n*.
5. Distribusi kondisional variabel tak bebas, untuk berbagai nilai variabel bebas tertentu, semua berdistribusi normal..
6. Nilai observasi yang satu dengan yang lain dari variabel random, tidak berkorelasi *(uncorrelated)*.

Pendugaan Parameter Model

Metode yang paling umum dalam analisis regresi untuk menduga parameter adalah Metode jumlah kuadrat galat terkecil. Metode kuadrat galat terkecil menekankan prosedur penilaian yang ditentukan dengan jumlah kuadrat galat terkecil (minimum) antara amatan dan dugaan.

β_0 dan β_1 adalah parameter regresi atau koefisien regresi yang tidak diketahui nilainya. Sedangkan ε_i adalah galat, nilai ε_i setiap pengamatan tidak sama. Meskipun tidak diketahui persis berapa nilainya tanpa memeriksa semua kemungkinan pasangan X dan Y, akan tetapi dapat digunakan informasi di dalam data contoh untuk menghasilkan nilai dugaan *(estimate)* b_0 dan b_1 bagi β_0 dan β_1 berturut-turut.

Jadi dapat dituliskan

$$\hat{Y} = b_0 + b_1 X \qquad (4)$$

$\hat{Y}$ melambangkan nilai taksiran Y untuk suatu X tertentu bila b_0 dan b_1 telah ditentukan.

Permasalahannya bagaimana mendapatkan penduga tersebut sehingga nilai dekat dengan nilai observasi Y. Kriteria penaksiran metode kuadrat terkecil (agar bebas dari asumsi-asumsi tentang ε_i) yaitu meminimumkan jumlah kuadrat simpangan, $\sum_{i=1}^{n} \varepsilon_i^2$. Dari persamaan (3), jumlah kuadrat semua simpangan garis yang sebenarnya adalah

$$S = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_1\right)^2 \tag{5}$$

Sebagai nilai dugaan akan dipilih b_0 dan b_1 yang apabila nilai tersebut disubstitusikan ke persamaan (5) akan dihasilkan S yang minimum (Draper and Smith, 1992). Secara intuitif dapat dimengerti bahwa semakin dekat titik-titik ke garis regresi maka semakin kecil jumlah kuadrat simpangan.

Penduga b_0 dan b_1 dapat ditentukan dengan mendiferensialkan persamaan (5) terhadap β_0 dan kemudian terhadap β_1 setelah itu menyamakan pendiferensialan itu dengan nol.

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2}{\partial \beta_0}$$

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right) \tag{6}$$

$$\frac{\partial S}{\partial \beta_1} = \frac{\partial \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2}{\partial \beta_1}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i \left(Y_i - \beta_0 - \beta_1 X_i\right) \tag{7}$$

Sehingga dapat digunakan untuk memperoleh nilai dugaan b_0 dan b_1 melalui melalui persamaan berikut:

$$\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right) = 0 \tag{8}$$

$$\sum_{i=1}^{n} X_i \left(Y_i - b_0 - b_1 X_i\right) = 0 \tag{9}$$

dengan demikian

$$b_1 = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} \tag{10}$$

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{11}$$

Persamaan (11) disubstitusikan kedalam persamaan (4) sehingga didapat bahwa

$$\hat{Y} = \overline{Y} + b_1 \left(X - \overline{X}\right) \tag{12}$$

ini menunjukkan bahwa garis regresi melalui rata-rata nilai Y observasi dan $\overline{X}$.

Pengujian Slope dimaksudkan untuk menentukan apakah parameter tersebut mencakup nilai-nilai tertentu.Jika empat asumsi untuk ε telah dipenuhi, maka distribusi sampling untuk $\hat{\beta}_1$ akan berdistribusi normal dengan rata-rata β_1 (slope sebenarnya) dan memiliki deviasi standar :

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{JK_{XX}}} \tag{13}$$

Dimana :

$\sigma_{\hat{\beta}_1}$ = Simpangan baku terhadap distribusi sampling $\hat{\beta}_1$

JK_{XX} = Jumlah Kuadrat

Uji hipotesis kegunaan model (Regresi Linier Sederhana) adalah sebagai berikut:

1. Hipotesis

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

2. Tingkat signifikansi α berhubungan dengan distribusi t dengan derajat bebas $(n-2)$.

3. Statistik Uji yang digunakan:

$$t_h = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s \Big/ \sqrt{JK_{XX}}} \tag{14}$$

4. Daerah Penolakan jika H_0 ditolak, bila $t_h > t_{\alpha/2}$ atau $t_h < -t_{\alpha/2}$.

5. Kesimpulan

Selang kepercayaan 100 $(1-\alpha)$ % untuk β_1 pada regresi linier sederhana adalah

$\hat{\beta}_1 \pm t_{(\alpha/2, n-2)} \cdot s\hat{\beta}_1$

dengan

$$s\hat{\beta}_1 = \frac{s}{\sqrt{JK_{XX}}} \tag{15}$$

Koefisien Determinasi

Kelayakan model regresi linier dapat diukur dengan menggunakan koefisien determinasi (R^2)

Koefisien determiasi didefinisikan sebagai

$$R^2 = \frac{JKR}{JKT} = 1 - \frac{JKG}{JKT} = \frac{JKT - JKG}{JKT} \tag{16}$$

Semakin besar JK regresi, R^2 semakin mendekati satu. Namun dalam data berulang R^2 tak mungkin bernilai satu karena adanya galat murni (Draper and Smith, 1992)

Pencilan dalam Regresi Linier Sederhana

Pada pencilan, selalu ada informasi yang harus dibuang. Dibutuhkan solusi untuk masalah itu yaitu dengan mengidentifikasi kemungkinan pencilan dan menaksir pengaruh yang terjadi. Sisaan yang merupakan pencilan adalah yang nilai mutlaknya jauh lebih besar daripada sisaan-sisaan lainnya dan bisa jadi terletak tiga atau empat simpangan baku atau lebih jauh lagi dari rata-rata sisaannya. Pencilan merupakan suatu keganjilan dan menandakan suatu titik data yang sama sekali tidak tipikal dibandingkan data lainnya. Oleh karenanya, suatu pencilan patut diperiksa secara seksama, barangkali saja alasan dibalik keganjilan itu dapat diketahui. Adakalanya pencilan memberikan informasi yang tidak bisa diberikan oleh titik lainnya, misalnya karena pencilan timbul dari kombinasi keadaan yang tidak biasa yang mungkin saja sangat penting dan perlu diselidiki lebih jauh (Draper & Smith, 1992).

Pencilan adalah hasil observasi (data pengukuran) dalam suatu kumpulan data yang nilainya sangat berbeda jika dibandingkan dengan sekumpulan data dari pengukuran lain. Penyebab pencilan ada tiga yaitu: data pengukuran tidak dicatat dan dimasukkan dalam komputer dengan benar, data pengukuran berasal dari populasi lain, dan data pengukurannya benar, tetapi mewakili peristiwa (keadaan) yang jarang terjadi (Santosa , 2004). Tanda dari pencilan adalah residunya yang besar (dalam harga mutlak) dibandingkan residu dari data yang lain.

Formulasi untuk pengujian pencilan yaitu:

$Y=\beta_0+\beta_1 X+\varepsilon \qquad ; i=1,2,...,n$

Misalkan data ke-i dicurigai sebagai outlier.

Untuk menguji kecurigaan ini, hipotesisnya adalah :

H_0 : data ke-i bukan outlier

H_1 : data ke-i adalah outlier

Hipotesis ini ekivalen dengan :

$$H_0 : Y_j = \beta_0 + \beta_1 X_j + \varepsilon_i \qquad j = 1,2,...,n$$

$$H_1 \begin{cases} Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j \\ Y_i = \beta_0 + \beta_1 X_i + \delta + \varepsilon_i \end{cases}$$

$$j = 1,2,...,n \quad i \neq j$$

Dengan demikian, yang akan diuji adalah:

$$H_0 : \delta = 0$$

$$H_1 : \delta \neq 0$$

Cara menguji pencilan dilakukan yaitu dengan langkah-langkah;

1). Buatlah prediksi variabel dummy X_2 sebagai berikut:

$$X_{2j} = \begin{cases} 1 & bila \quad j = i \\ 0 & bila \quad j \neq 1 \end{cases}$$

2). Dengan menggunakan variabel dummy X_2 ini, hipotesis diatas menjadi:

$$H_0 : Y_j = \beta_0 + \beta_1 X_j + \varepsilon_i$$

$$H_1 : Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j + \delta X_{2j}$$
$$j = 1,2,...n$$

3) Pengujian hipotesis ini adalah ekivalen dengan pengujian X_2.

4) Statistik uji yang digunakan adalah ;

$$F = \frac{(R_2^2 - R_1^2)(n-3)}{1 - R_2^2}$$

dengan

$R_1^2 = R^2$

jika modelnya $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j$

$R_2^2 = R^2$

Jika modelnya $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j + \delta X_{2j}$

5) H_0 ditolak, yang berarti data *ke-i* adalah pencilan, jika $F \geq F(\alpha : 1, n-3)$

Sisa memberikan keterangan tentang data yang tidak mengikuti pola umum model yang digunakan, ditandai oleh sisanya yang relatif besar. Sisa yang relatif besar dapat merupakan petunjuk bahwa modelnya belum cocok ataupun pengamatannya barangkali merupakan pencilan. Secara umum, pencilan adalah data yang tidak dapat mengikuti pola umum model dan secara kasar dapat diambil patokan yaitu yang sisanya berjarak tiga simpangan baku atau lebih dari rata-ratanya (Sembiring, 1995).

Tujuan pemeriksaan sisa, secara implisit, juga bearti apakah peubah bebas yang besar pengaruhnya sudah masuk kedalam model dan dalam bentuk (linier, kuadrat, log, dsb) yang sesuai. Secara lebih terperinci tujun pemeriksaan sisa adalah :

1. apakah sisa telah berpola acak;
2. apakah anggapan normal tidak dilanggar;
3. apakah variansi dapat dianggap tidak berubah (sama);
4. apakah ada data yang tidak mengikuti pola umum (pencilan);
5. apakah peubah yang masuk dalam model barangkali bukan berbentuk linier;
6. apakah peubah yang berpengaruh telah masuk kedalam model;

Simulasi

Sebagai suatu aplikasi, dibangkitkan data simulasi untuk beberapa ukuran sampel, yaitu n=26. Data dibangkitkan dari n=26 kemudian disimulasikan sebanyak n=100 untuk melihat perubahan nilai korelasi. Berbagai kemungkinan tambahan data (X,Y) mengakibatkan perubahan parameter regresi dan korelasi. Dari hasil yang disajikan pada tabel lampiran 1, akan dilihat pola perubahan khususnya korelasi dengan adanya tambahan data baru. Penggunaan perubahan korelasi lebih beralasan karena ukuran ini menunjukkan kepada keeratan hubungan dua variabel. Pola yang akan dilihat adalah jarak data baru (X,Y) terhadap $(\bar{X},\bar{Y})$ data lama, serta besarnya sudut yang dibentuk antara $(\bar{X},\bar{Y})$ terhadap (X,Y) dengan sudut persamaan regresi sebelum adanya tambahan data baru. Data baru

akan dapat disetarakan dengan pencilan jika mengakibatkan penurunan koefisien korelasi pada taraf tertentu.

Jika penurunan tersebut tidak lebih dari 7%, maka koefisien korelasi yang barunya masih diatas 0,8. Sedangkan apabila penurunan tersebut tidak lebih dari 18%, maka korelasi barunya masih diatas 0,7. Penurunan tidak lebih dari 29% mengakibatkan korelasi barunya tidak kurang dari 0,6. Pada taraf signifikansi korelasi baru 1%, sebetulnya memberikan toleransi penurunan koefisien korelasi hingga 40% yang masih membuat koefisien korelasi barunya diatas 0,5.

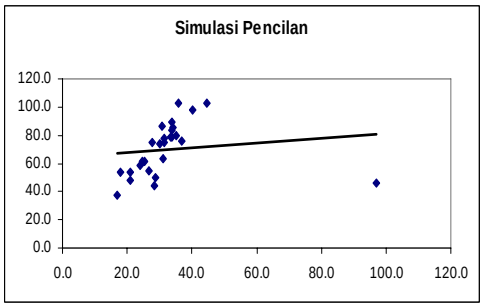
Dengan memperlihatkan pola pada tiap kasus penurunan koefisien korelasi diatas, dapat dilihat bahwa :

1). Apabila sudut yang dibentuk antara (X,Y) data baru dan $(\bar{X},\bar{Y})$ data lama terhadap garis regresi lamanya semakin mendekati 0° atau 180° semakin kecil penurunan koefisien korelasinya, bahkan pada kasus tertentu terjadi penambahan.

2). Meskipun sudut yang diberikan seperti dijelaskan pada poin (1), namun jarak antara (X,Y) data baru dan $(\bar{X},\bar{Y})$ data lama " tidak jauh ", maka penurunan koefisien korelasi itupun tidak terlalu berarti.

Tambahan data (X,Y) dengan sudut yang dibentuk terhadap $(\bar{X},\bar{Y})$ data lama dengan garis regresi yang mendekati 0° atau 180°, atau dengan kata lain data baru tersebut "dekat" dengan garis regresi tidak akan merubah koefisien korelasi secara berarti. Hal ini dapat dijelaskan bahwa apabila (X,Y) ada didekat garis regresi dan (X,Y) relatif disebelah kanan $(\bar{X},\bar{Y})$ maka pembilang dan penyebut pada koefisien regresi relatif secara sama berubah. Demikian juga apabila (X,Y) relatif disebelah kiri $(\bar{X},\bar{Y})$. Dengan melakukan proses sebaliknya, untuk menguji satu persatu dari data yang ada tersebut apakah data disebut sebagai pencilan atau bukan.

Dibawah ini terdapat berbagai simulasi pencilan berdasarkan $(\bar{X},\bar{Y})$ data lama



Gambar 1. Salah Satu Grafik Simulas Pencilan

KESIMPULAN

Apabila diberikan data bivariat yang cukup bagus untuk mendeskripsikan adanya hubungan yang erat dengan model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, penambahan data (X_i, Y_i) yang baru tidak akan menurunkan koefisien korelasi data lama jika :

1). Data baru tersebut letaknya dekat dengan garis regresi lama, atau sudut yang dibentuk antara (X_{i1}, Y_{i1}) dengan $(\bar{X}, \bar{Y})$ terhadap garis regresinya mendekati 0° atau 180°.

2). Jarak (X, Y) seperti disediakan diatas relatif dekat $(\bar{X}, \bar{Y})$ meskipun sudut yang dibentuk relatif besar.

Data baru yang mengakibatkan penurunan koefisien korelasi yang ”cukup berarti” dapat dikategorikan sebagai data pencilan.

DAFTAR PUSTAKA

1. Anonim. http://www.math.itb.ac.id/~ma291/rls.htm (27 Juli 2007).
2. Aunuddin. 2005. *Statistika: Rancangan dan Analisis Data*. IPB: Bogor.
3. Barnett, V. and Lewis. T. 1978. *Outliers in Statistical Data.* Chichester: Wiley.
4. Bowen, E.K. and M.K. Star. 1982. *Basic Statistics for Business and Economics,* McGraw-Hill Book Company, Singapore.
5. Draper, N.R. and H. Smith. 1992. *Analisis Regresi Terapan, Edisi kedua (Terjemahan).* Jakarta: Gramedia Pustaka Utama.
6. Irianto, A. 2004. *Statistika Konsep Dasar dan Aplikasinya*. Jakarta: Kencana.
7. Mangkuatmodjo, S. 2004. *Statistik Lanjutan*. Jakarta: Rineka Cipta.
8. Santosa, G.R. 2004. *Statistik*. Yogyakarta: ANDI.
9. Sembiring, R.K. 1995. *Analisis Regresi.* Penerbit ITB: Bandung.
10. Sudjana, M.A. 2001. *Teknik Analisis Regresi dan Korelasi Bagi Para Peneliti*. Bandung: Tarsito.
11. Walpole, R.E. 1992. *Pengantar Statistika (Terjemahan)*. Jakarta: Gramedia Pustaka Utama.
12. Walpole, R.E. and R.H. Myers. 1995. *Ilmu Peluang dan Statistika Untuk Insinyur dan Ilmuwan, edisi keempat (Terjemahan)*. ITB: Bandung.
13. Weisberg, S. 1980. *Applied Linear Regression*. New York: John Wiley & Sons.