Classification and Regression Tree (Cart) Analysis Pada Penderita Skizofrenia di RSJKO Soeprapto Daerah Bengkulu

Nurul Komariah[1], Sigit Nugroho[2], Jose Rizal[2]

[1]Alumni Jurusan Matematika Fakultas MIPA Universitas Bengkulu
[2]Staf Pengajar Jurusan Matematika Fakultas MIPA Universitas Bengkulu

ABSTRAK

Penelitian ini bertujuan untuk mengetahui deskripsi, model pohon klasifikasi dan model pohon regresi pada penderita skizofrenia di Rumah Sakit Jiwa dan Ketergantungan Obat (RSJKO) Soeprapto Daerah Bengkulu. Metode yang digunakan adalah Analisis CART (*Classification and Regresssion Tree*). Analisis statistik nonparametrik ini dirancang untuk menyajikan kaidah keputusan berbentuk pohon biner dengan menggunakan prosedur penyekatan berulang. Analisis ini memiliki keunggulan dalam menangani data dengan struktur yang kompleks, efisien dalam terminologi perhitungan, sangat tangguh dalam menangani pencilan, dan dapat menangani sembarang data kontinu/ numerik dan kategorik. Pohon klasifikasi dibangun dengan menggunakan dua kaidah, yaitu kaidah pemisahan *Gini* dan kaidah pemisahan *Twoing*, sedangkan pohon regresi dibangun dengan menggunakan metode *validasi silang*. Hasil penelitian menunjukkan bahwa 81.7% pasien merupakan pasien dengan diagnosa skizofrenia paranoid. Sisanya sebanyak 18.3% berturut-turut adalah skizofrenia residual, hebefrenik, katatonik dan tidak terinci. Nilai ketepatan klasifikasi dengan menggunakan kaidah *Gini* adalah 86.9%, sedangkan dengan kaidah *Twoing* adalah 85.6%. Pada pohon regresi diperoleh proporsi varian yang dijelaskan oleh model sebesar 89.7%. Ini menunjukkan bahwa model pohon klasifikasi dan pohon regresi yang dibangun adalah cukup baik.

Kata kunci : *Skizofrenia, CART, Gini, Twoing, Validasi silang*

PENDAHULUAN

Skizofrenia merupakan penyakit otak yang timbul akibat ketidakseimbangan pada *dopamine*, yaitu salah satu sel kimia dalam otak. Penyakit ini ditandai dengan adanya disorganisasi kepribadian yang cukup parah, distorsi realita dan ketidakmampuan berinteraksi dalam kehidupan sehari-hari[1].

Penderita skizofrenia memiliki frasa-frasa kata yang hanya dapat dimengerti oleh dirinya sendiri dan seringkali merasa "aneh" dengan bagian tubuh mereka sendiri. Pada umumnya penderita tidak merasakan emosi apa-apa serta tidak mampu merespon stimulus emosi dengan benar. Mereka seringkali menunjukkan aktivitas motorik dan ekspresi wajah yang aneh. Selain itu penderita juga melakukan gerakan yang tak lazim tanpa berhenti atau mempertahankannya dalam periode waktu yang lama[1].

Penderita skizofrenia khususnya di Kota Bengkulu seringkali ditelantarkan oleh keluarganya sehingga hidup sendiri atau berkeliaran di jalan tanpa ada perhatian dan penanganan khusus. Bahkan tidak sedikit yang menitipkan anggota keluarga mereka yang mengidap skizofrenia tersebut ke rumah sakit jiwa. Satu-satunya rumah sakit yang

menangani para penderita skizofrenia di Provinsi Bengkulu adalah *Rumah Sakit Jiwa dan Ketergantungan Obat (RSJKO) Soeprapto Daerah Bengkulu.*

Penderita skizofrenia yang dirawat di RSJKO Soeprapto Daerah Bengkulu memiliki latar belakang dan karakteristik yang berbeda-beda. Perbedaan ini meliputi jenis kelamin, umur, tingkat pendidikan, status perkawinan, dan lain-lain. Demikian pula halnya dengan jenis skizofrenia yang mereka derita. Untuk mengetahui deskripsi penderita skizofrenia yang dirawat di RSJKO Soeprapto Daerah Bengkulu, lamanya mereka menjalani perawatan, serta jenis-jenis skizofrenia yang paling banyak ditemukan pada penderita tersebut, maka dalam tulisan ini akan dibuat pohon klasifikasi dan pohon regresi menggunakan Analisis CART (*Classification and Regression Tree*) guna mendeskripsikan hal tersebut di atas.

Tujuan penelitian ini adalah untuk mengetahui deskripsi pasien skizofrenia, mengetahui model pohon klasifikasi (*Classification Tree*) diagnosa skizofrenia yang diderita, serta model pohon regresi (*Regression Tree*) lamanya penderita skizofrenia dirawat inap di RSJKO Soeprapto Daerah Bengkulu.

TINJAUAN PUSTAKA

Classification and Regression Trees adalah metode klasifikasi menggunakan data historis untuk membangun suatu pohon keputusan. Metodologi CART mulai dikembangkan pada tahun 80-an oleh Breiman, Friedman, Olshen, dan Stone dalam makalah mereka yang berjudul "*Classification and Regression Trees*" (1984).

CART adalah suatu analisis diskriminan non-parametrik yang dirancang untuk menyajikan kaidah keputusan berbentuk pohon biner yang membagi data pada learning sampel dalam batasan linier univariat. Analisis ini menghasilkan kelompok data hirarkis yang dimulai dari *node root* untuk keseluruhan learning sampel dan berakhir pada kelompok kecil pengamatan yang homogen. Pada setiap *terminal node* diberikan label kelas atau nilai yang diramalkan, sehingga menghasilkan struktur pohon yang dapat ditafsirkan sebagai pohon keputusan[11].

Keuntungan dari penggunaan analisis CART[11] adalah sebagai berikut :

1. Merupakan bentuk statistika non-parametrik, sehingga tidak memerlukan asumsi sebaran dan uji hipotesis.
2. Tidak memerlukan variabel untuk dipilih sebelumnya.
3. Sangat efisien dalam terminologi perhitungan.
4. Dapat menangani dataset dengan struktur yang kompleks.
5. Sangat tangguh dalam menangani outlier, umumnya algoritma pemisahan akan mengisolasikan outlier pada individu node atau beberapa node.
6. Dapat menggunakan sembarang kombinasi data kontinu/numerik dan kategorik.
7. Hasilnya invarian dengan transformasi monoton dari variabel respon, artinya penggantian sembarang variabel dengan algoritmanya atau nilai akar kuadrat, tidak akan menyebabkan struktur pohon berubah.

Pohon Klasifikasi

Classifier[12] atau kaidah klasifikasi adalah suatu cara sistematik dalam memprediksi suatu kasus masuk dalam kelas tertentu. Untuk memberikan formulasi yang lebih tepat, maka disusun suatu himpunan pengukuran $\{x_1, x_2, \ldots, x_n\}$ sebagai faktor pengukuran (*measurement*

vector). Semua vektor pengukuran yang mungkin berada di dalamnya didefinisikan sebagai ruang pengukuran X.

Andaikan suatu kasus atau objek mempunyai *J* kelas yaitu $1, 2, \ldots, j$ dan misalkan C adalah himpunan kelas tersebut dengan $C = (1, 2, \ldots, j)$. Suatu cara sistematik dalam memprediksi anggota kelas tersebut adalah dengan menggunakan suatu aturan yang menempatkan anggota kelas dalam C tersebut pada setiap vektor pengukuran x dalam X.

Definisi 1.

"Suatu *classifier* atau aturan klasifikasi adalah suatu fungsi $d(x)$ pada X sehingga untuk setiap x, $d(x)$ adalah sama dengan salah satu dari $(1, 2, \ldots, j)$."

Cara lain untuk melihat *classifier* adalah dengan mendefinisikan A_j sebagai subset dari X dimana $d(x)$ sama dengan *j* sehingga $A_j = (x ; d(x) = j)$.
Himpunan $A_1, \ldots, A_j$ adalah disjoint dan

$$X = \bigcup_j A_j$$

sehingga A_j adalah bentuk partisi dari X.

Definisi 2.

"*Classifier* adalah suatu partisi pada X dalam J yang memisahkan himpunan bagian $A_1, \ldots, A_j \ni X = \cup A_j$. Sehingga untuk setiap $x \in A_j$ kelas prediksinya adalah *j*."

Dalam konstruksi klasifikasi sistematik, semua data historis dirangkum dalam suatu *learning sample*. Learning sampel merupakan sampel data yang digunakan untuk membangun pohon klasifikasi.

Sehingga diperoleh definisi rumus-rumus sebagai berikut (gambar 2) :

1. Proporsi pengamatan pada kelas ke-*j* terhadap jumlah keseluruhan pengamatan.
$$\pi(j) = \frac{N_j}{N} \quad (1)$$
2. Jumlah pengamatan pada kelas ke-*j*.
$$N_j = \sum_{s=1}^{t} N_j(s) \quad (2)$$
3. Jumlah pengamatan pada node *s*.
$$N(s) = \sum_{j=1}^{k} N_j(s) \quad (3)$$
4. Peluang pengamatan pada node *s*.
$$p(s) = \frac{N(s)}{N} \quad (4)$$
5. Peluang bersama pengamatan pada node *s* kelas ke-*j*.
$$p(j,s) = \frac{N_j(s)}{N} \quad (5)$$

6. Peluang bersyarat pengamatan pada node *s* kelas ke-*j*.

$$p(j|s) = \frac{p(j,s)}{p(s)} = \frac{\frac{N_j(s)}{N}}{\frac{N(s)}{N}} = \frac{N_j(s)}{N(s)} \quad (6)$$

Dari persamaan (3) dan (6) diperoleh rumus berikut :

$$p(1|s) + p(2|s) + p(3|s) + \ldots + p(k|s) = \sum_{j=1}^{k} p(j|s) = \sum_{j=1}^{k} \frac{N_j(s)}{N(s)}$$

$$= \frac{1}{N(s)} . N(s) = 1$$

Suatu ukuran impurity pada node *t* disimbolkan dengan $i(t)$, dimana $i(t)$ merupakan suatu fungsi peluang kelas $p(1|t), p(2|t), \ldots, p(k|t)$. Sehingga secara matematis dapat dituliskan dengan :

$$i(t) = f\,[p(1|t), p(2|t), \ldots, p(k|t)] \quad (7)$$

Definisi 3.

"*Impurity function* adalah suatu fungsi *f* yang didefinisikan pada himpunan $(p_1, p_2, \ldots, p_k)$ yang memenuhi $p_j \geq 0,\ j = 1, \ldots, k,\ \sum_j p_j = 1$ dengan kriteria sebagai berikut :

1. *f* akan *maksimum unik* pada titik $\left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$. Dengan kata lain, masing-masing kelas dalam populasi memiliki peluang yang sama.
2. *f* akan *minimum unik* pada titik $(1,0,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,0,0,\ldots,1)$.
3. *f* adalah fungsi simetrik dari $p_1, p_2, \ldots, p_k$."

Misalkan jumlah terminal node pada pohon klasifikasi adalah $\tilde{T}$ dan diketahui himpunan $I(t) = i(t).p(t)$. Rumus impurity pohon (*tree impurity*) didefinisikan dengan :

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t).p(t) \quad (8)$$

Suatu pohon klasifikasi dibangun berdasarkan aturan pemisahan (*splitting rule*), yaitu aturan yang memisahkan learning sampel ke dalam bagian yang lebih kecil. Setiap kali data yang ada harus dibagi menjadi dua bagian dengan homogenitas maksimum.

Anggap t adalah node ayah yang dipisahkan oleh pembagian x menjadi dua node anak, yaitu node anak kiri t_L dan node anak kanan t_R. Masing-masing node anak tersebut mempunyai peluang, p_L dan p_R dengan $p_R = 1 - p_L$ (gambar 3). Pada sembarang terminal node, akan dipilih pembagian yang paling mengurangi nilai $I(t)$ dengan kata lain akan akuivalen dengan memaksimalkan perubahan fungsi impurity node *t* pada *x* sebagai berikut :

$$\Delta I(x,t) = I(t) - I(t_L) - I(t_R)$$

atau

$$\Delta i(x,t) = i(t) - p_L . i(t_L) - p_R . i(t_R) \quad (9)$$

Nilai pemisahan terbaik $\Delta i(x,t)$ menunjukkan perubahan impurity node *t* pada *x* dengan $t_L \cup t_R = 1$. Oleh karena itu, pemisahan terbaik dari *t* adalah :

$$\arg\max_x \{\Delta i(x,t)\} = \arg\max_x \{i(t) - p_L\, i(t_L) - p_R\, i(t_R)\} \tag{10}$$

Nilai optimal x^* dapat ditentukan dengan cara memaksimalkan $\Delta i(x,t)$ dengan *x* yang berbeda pada masing-masing node *t*. Prosedur semacam ini memungkinkan untuk membangun pohon keputusan dari sembarang bentuk pohon maksimum.

Karena nilai $i(t)$ pada kenyataannya adalah konstan, sehingga hasilnya akan ekuivalen dengan :

$$\begin{aligned} x^* = \arg\max_x \Delta i(x,t) &= \arg\max_x \{-p_L\, i(t_L) - p_R\, i(t_R)\} \\ &= \arg\min_x \{p_L\, i(t_L) + p_R\, i(t_R)\} \end{aligned} \tag{11}$$

dimana t_L dan t_R adalah fungsi eksplisit dari *x*.

Di dalam teori ada beberapa fungsi impurity, tetapi yang secara luas digunakan dalam prakteknya, yaitu Kaidah Pemisahan Gini (*Gini Splitting Rule*) dan Kaidah Pemisahan Twoing (*Twoing Splitting Rule*).

Gini Splitting Rule

Gini Splitting Rule atau disebut juga indeks Gini (*Gini Index*) adalah kaidah yang paling umum digunakan dalam memecahkan permasalahan pohon klasifikasi. Data impurity didefinisikan dengan menggunakan ukuran varian (*variance measure*). Misalkan 1 adalah semua pengamatan pada node *t* kelas ke-*j* dan 0 untuk yang lainnya. Kemudian estimasi varian contoh untuk node *t* pengamatan sebagai berikut :

$$p(j|t)\big(1 - p(j|t)\big)$$

Indeks Gini pada node *t* kelas ke-*j* didefinisikan dengan rumus sebagai berikut :

$$\begin{aligned} i(t) &= \sum_{j=1}^{k} p(j|t)\,(1 - p(j|t)) = \sum_{j=1}^{k} p(j|t) - p^2(j|t) \\ &= \sum_{j=1}^{k} p(j|t) - \sum_{j=1}^{k} p^2(j|t) \\ &= 1 - \sum_{j=1}^{k} p^2(j|t) \end{aligned} \tag{12}$$

Sehingga diperoleh perubahan fungsi impurity node *t* oleh pemisahan *x* sebagai berikut :

$$\Delta i(x,t) = -\sum_{j=1}^{k} p^2(j|t) + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R) \tag{13}$$

Pemisahan terbaik dari *t* dirumuskan dengan :

$$\arg\max_x \Delta i(x,t) = \arg\max_x \left\{ -\sum_{j=1}^{k} p^2(j|t) + p_L \cdot \sum_{j=1}^{k} p^2(j|t_L) + p_R \cdot \sum_{j=1}^{k} p^2(j|t_R) \right\} \tag{14}$$

Indeks Gini akan mencari learning sampel untuk kelas paling besar dan mengisolasinya dari sisa data tersebut. Kaidah ini bekerja dengan baik pada data berukuran besar.

Twoing Splitting Rule

Tidak seperti kaidah *Gini*, kaidah *Twoing* tidak mencari nilai maksimal dari ukuran impurity. Sebagai gantinya kaidah ini mencoba untuk menyeimbangkan konstruksi pohon dengan cara seolah-olah membagi learning sampel menjadi dua kelas. Sehingga pengamatan dapat dibedakan antara faktor umum yang berada pada tingkat teratas dan karakteristik khusus yang berada pada tingkat yang lebih rendah.

Misalkan terdapat himpunan kelas learning sampel $C = \{1, 2, \ldots, k\}$.Himpunan tersebut dibagi menjadi dua bagian yaitu: $C_1 = (c_1, c_2, \ldots, c_n)$ dan $C_2 = C \setminus C_1$ sedemikian sehingga semua pengamatan yang berada pada C_1 mempunyai kelas dummy 1, sedangkan sisanya mempunyai kelas dummy 2[11].

Kemudian akan dihitung nilai $\Delta i(x,t)$ untuk *x* yang berbeda "jika hanya ada dua kelas dummy". Karena nilai $\Delta i(x,t)$ bergantung pada C_1, maka nilai $\Delta i(x,t,C_1)$ adalah maksimal. Dengan kata lain, kaidah Twoing adalah suatu aturan yang digunakan untuk menemukan kombinasi superkelas pada setiap node seolah-olah kenaikan impurity telah dimaksimalkan hanya oleh dua kelas $C = \{1, 2\}$.

Walaupun kaidah Twoing dapat diterapkan terutama untuk data dengan jumlah kelas yang besar, kelemahannya terdapat pada kecepatan perhitungan. Asumsikan bahwa learning sampel mempunyai *J* kelas, kemudian himpunan C dipisahkan menjadi C_1 Dan C_2 dengan 2^{J-1} cara. Pada kasus dimana terdapat 11 data kelas pada learning sampel, maka akan terbentuk 1024 kombinasi.

Pada kaidah twoing, tidak ada ukuran impurity yang spesifik[25]. Sehingga untuk sembarang node, pemisahan yang terbaik ditentukan dengan cara memaksimalkan perubahan impurity pada node anak kanan t_R dan node anak kiri t_L. Ini mengakibatkan timbulnya perbedaan definisi kaidah twoing oleh para peneliti, antara lain :

1. Chee Jen Chang (2002)

$$i(t) = \frac{P_L \cdot P_R}{4}\left(\sum_j \left(|p(j|t_L) - p(j|t_R)|\right)\right)^2 \quad (15)$$

2. David Feldman (2003)

$$i_T(t) = \frac{P_L \cdot P_R}{4}\sum_{j=1}^{J} |p(j|t_L) - p(j|t_R)| \quad (16)$$

3. Roman Timofeev (2004)

$$\Delta i(t) = \frac{P_L \cdot P_R}{4}\left[\sum_{k=1}^{k} |p(k|t_L) - p(k|t_R)|\right]^2 \quad (17)$$

4. Anton Andriyashin (2005)

$$S_L(t) = \{j : p(j|t_L) \geq p(j|t_R)\}$$

$$\max_{s} \Delta i(s,t,S_t) = \frac{P_L \cdot P_R}{4}\left[\sum_{j=1}^{J}\left|p(j|t_L) - p(j|t_R)\right|\right]^2 \quad (18)$$

5. Ozge Sezgin (2006)

$$\Delta i(s) = \frac{P_L \cdot P_R}{4}\left[\sum_{j=1}^{J} p(j|t_L)\, p(j|t_R)\right]^2 \quad (19)$$

6. Zambon, et al (2006)

$$Twoing(s) = \frac{P_L \cdot P_R}{4}\left(\sum_{j}\left|p(j|t_L) - p(j|t_R)\right|\right)^2 \quad (20)$$

Pohon Regresi

Konstruksi pohon pada pohon klasifikasi dan pohon regresi tidak jauh berbeda, yang membedakannya adalah jenis variabel responnya. Pohon regresi adalah jenis pohon keputusan dengan variabel respon kontinu. Tujuan utama pohon regresi adalah untuk menghasilkan struktur pohon prediktor atau kaidah prediksi. Prediktor ini menjalankan dua tujuan utama, yaitu : (1) untuk memprediksi ketelitian variabel respon atau nilai baru dari variabel prediktor, (2) untuk menjelaskan hubungan antara variabel respon dan variabel prediktor[27].

Prediktor pada pohon regresi dibangun dengan mendeteksi heterogenitas (dalam kaitan dengan varian pada variabel prediktor) yang ada di dalam data tersebut. Pohon regresi melakukannya dengan penyekatan berulang data ke dalam kelompok atau terminal node yang secara internal lebih homogen dibandingkan node di atasnya. Pada masing-masing terminal node, nilai rata-rata dari variabel respon dianggap sebagai nilai prediksi.

Terdapat dua kaidah pemisahan atau fungsi impurity pada pohon regresi, yaitu *Least Square (LS) function dan the Least Absolute Deviation (LAD) function*. Karena mekanisme kedua aturan tersebut sama, maka disini hanya akan dibahas ukuran impurity LS. Menggunakan kriteria ini, impurity node diukur dengan jumlah kuadrat node, $SS(t)$ yang didefinisikan sebagai berikut :

$$SS(t) = \sum_{i=1}^{N(t)} (y_{i(t)} - \bar{y}_{(t)})^2 \quad (21)$$

Misalkan diberikan fungsi impurity, $SS(t)$ yang dipisahkan oleh *x* pada node kiri t_L dan node kanan t_R. Sehingga pemisahan terbaik (*goodness of split*) dapat ditunjukkan dengan rumus sebagai berikut :

$$\underset{s}{\operatorname{argmax}}\,(f(s,t)) = \underset{s}{\operatorname{argmax}}\,(SS(t) - SS(t_R) - SS(t_L)) \quad (22)$$

Alternatif lain dari fungsi impurity pada pohon regresi adalah dengan menggunakan nilai varian dari node anak kiri dan node anak kanan. Nilai varian pada suatu node *t* didefinisikan dengan rumus berikut :

$$s^2(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} [y_{i(t)} - \bar{y}_{(t)}]^2 \quad (23)$$

Sehingga diperoleh perubahan fungsi impurity node *t* oleh pemisahan *x* berikut :

$$\Delta x^2(t) = x^2(t) - p_L.x^2(t_L) - p_R.x^2(t_R) \quad (24)$$

Nilai perubahan impurity $\Delta x^2(t)$ menunjukkan nilai pemisahan terbaik node *t* oleh pemisahan *x*, yang dirumuskan dengan :

$$\underset{x}{\operatorname{argmax}} \left(\Delta x^2(t)\right) = \underset{x}{\operatorname{argmax}} \left(x^2(t) - p_L.x^2(t_L) - p_R.x^2(t_R)\right) \quad (25)$$

Sama halnya dengan pohon klasifikasi, nilai optimal x^* dapat ditentukan dengan cara memaksimalkan $\Delta x^2(t)$ dengan *x* yang berbeda pada masing-masing node *t*. Karena nilai $x^2(t)$ adalah konstan, sehingga hasilnya akan ekuivalen dengan :

$$\begin{aligned} x^* &= \underset{x}{\operatorname{argmax}} \left(\Delta x^2(t)\right) = \underset{x}{\operatorname{argmax}} \left(-p_L.x^2(t_L) - p_R.x^2(t_R)\right) \\ &= \underset{x}{\operatorname{argmin}} \left(p_L.x^2(t_L) + p_R.x^2(t_R)\right) \end{aligned} \quad (26)$$

Sehingga diperoleh kesimpulan bahwa pemisahan yang terbaik adalah nilai $\Delta x^2(t)$ yang paling tinggi dengan jumlah varian $\left(p_L.x^2(t_L) + p_R.x^2(t_R)\right)$ yang paling rendah. Prosedur ini menghasilkan node anak kiri dan node anak kanan yang lebih homogen dibandingkan node ayahnya. Masing-masing bagian membagi pengamatan pada node anak kiri dan node anak kanan sehingga nilai rata-rata dari variabel prediktor pada suatu terminal node adalah lebih rendah dibandingkan nilai rata-rata node ayah.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dengan metode dokumentasi, yaitu pengambilan data yang bersumber arsip registrasi pasien skizofrenia yang dirawat di masing-masing ruangan/bangsal di RSJKO Soeprapto Daerah Bengkulu periode 01 Januari 2007 sampai dengan 31 Oktober 2007. Adapun yang menjadi populasi dalam penelitian ini adalah seluruh pasien skizofrenia yang menjalani rawat inap di RSJKO Soeprapto Daerah Bengkulu yang dibagi dalam tiga ruangan, yaitu : ruang Anggrek, Murai, dan Rajawali.

Untuk membangun pohon klasifikasi dan pohon regresi, maka dibuat langkah-langkah pengerjaan sebagai berikut :

1. Data yang diperoleh dibuat dalam bentuk tabel pada program *Microsoft Excel*.
2. Untuk menentukan proporsi pada setiap variabel, digunakan software program *SPSS* versi 16.0. Dalam hal ini variabel berbentuk kategorik harus dibuat menjadi variabel *dummy* untuk memudahkan perhitungan.
3. Langkah selanjutnya adalah membuat model pohon klasifikasi dan pohon regresi.

Pohon Klasifikasi	Pohon Regresi
Respon (Y) = diagnosa skizofrenia	Respon (Y) = lamanya pasien dirawat
Prediktor $(X_i), i = 1,2,\ldots,8$	Prediktor $(X_i), i = 1,2,\ldots,8$
X_1 = jenis kelamin	X_1 = jenis kelamin
X_2 = finansial/ cara pembayaran	X_2 = finansial/ cara pembayaran
X_3 = umur	X_3 = umur
X_4 = pendidikan	X_4 = pendidikan
X_5 = pekerjaan	X_5 = pekerjaan
X_6 = status pernikahan	X_6 = status pernikahan

X_7 = daerah asal	X_7 = daerah asal
X_8 = lama dirawat	X_8 = diagnosa skizofrenia

HASIL DAN PEMBAHASAN

Deskripsi pasien skizofrenia yang dirawat inap di RSJKO Soeprapto Bengkulu meliputi antara lain : jenis kelamin, finansial, diagnosa, tingkat pendidikan, status pekerjaan, status pernikahan, dan daerah asal. Hasil penelitian menunjukkan bahwa mayoritas pasien skizofrenia berjenis kelamin laki-laki sebanyak 80.4% dengan latar belakang kalangan tidak mampu. Ini dibuktikan dengan finansial menggunakan Surat Keterangan Tidak Mampu dengan persentase sebanyak 68.6%. Kemudian disusul dengan Askin, Askes dan Umum. Diagnosa jenis skizofrenia yang diderita adalah 81.7% skizofrenia paranoid, 9.2% skizofrenia residual, 7.8% skizofrenia hebefrenik, dan sisanya masing-masing 0.7% skizofrenia katatonik dan tidak terinci.

Sebaran tingkat pendidikan pasien skizofrenia cukup merata, antara lain : lulusan SD sebanyak 34.6%, lulusan SLTA sebanyak 26.1%, lulusan SLTP sebanyak 25.5%, dan lulusan S.1 sebanyak 2.6%. Sisanya sebanyak 10.5% tidak menempuh pendidikan formal. Ini berarti skizofrenia dapat diderita seseorang tanpa memandang latar belakang pendidikannya. Status pekerjaan dapat dilihat dari mayoritas pasien skizofrenia Tidak Bekerja dengan persentase sejumlah 60.1%. Status Tidak Bekerja disini dapat berupa ibu rumah tangga, mahasiswa, pelajar, atau pengangguran. Status pernikahan cukup beragam, yaitu Belum Kawin sebanyak 56%, Kawin sebanyak 35.9%, sisanya masing-masing 1.3% dan 2.6% memiliki status Janda dan Duda.

Daerah asal pasien skizofrenia tersebar merata di seluruh wilayah Provinsi Bengkulu. Jumlah terbanyak berasal dari Kota Bengkulu sebanyak 20.9%, Rejang Lebong 17%, Bengkulu Selatan 14.4% dan Bengkulu Utara 13.7%. Kabupaten-kabupaten pemekaran, daerah di luar propinsi Bengkulu, serta gepeng (gelandangan dan pengemis) memiliki jumlah yang tidak terlalu banyak.

Jumlah minimum *parent node* dan *child node* serta jumlah maksimum kedalaman pada pohon klasifikasi dan pohon regresi yang dibangun dapat ditentukan sebelumnya dengan menggunakan software SPSS versi 16.0. Pohon klasifikasi yang dibangun dengan menggunakan kaidah Gini memiliki 5 kedalaman maksimum, jumlah minimum *parent node* sebanyak 5 dan *child node* sebanyak 1. Hasil yang diperoleh adalah 23 node, yaitu 1 node root, 10 internal node dan 12 terminal node. Sedangkan pohon klasifikasi yang dibangun dengan menggunakan Kaidah Twoing memiliki 5 kedalaman maksimum, jumlah minimum *parent node* sebanyak 5 dan *child node* sebanyak 2. Hasil yang diperoleh adalah 19 node, yaitu 1 node root, 8 internal node dan 10 terminal node. Variabel yang paling berpengaruh pada pohon klasifikasi baik menggunakan kaidah Gini maupun Twoing adalah variabel "Daerah Asal".

Untuk menangani pohon dengan kompleksitas yang tinggi, maka dilakukan prosedur pembabatan pohon (*pruning tree*). Pada pohon klasifikasi yang dibangun dengan menggunakan kaidah Gini, diperoleh nilai ketepatan klasifikasi untuk keseluruhan data penelitian adalah 86.9 %. Setelah dilakukan *pruning tree*, nilainya adalah tetap. Pada pohon klasifikasi yang dibangun dengan menggunakan kaidah Twoing, diperoleh nilai ketepatan

klasifikasi untuk keseluruhan data penelitian adalah 86.9 %. Setelah dilakukan *pruning tree*, nilainya turun menjadi 85.6 % .

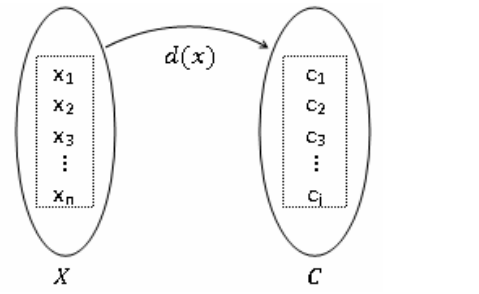
Pohon regresi yang dibangun dengan menggunakan metode *cross-validation* memiliki 5 kedalaman maksimum dengan jumlah minimum *parent node* sebanyak 2 dan *child node* sebanyak 1. Hasil yang diperoleh adalah 23 node, yaitu 1 node root, 10 internal node dan 12 terminal node. Variabel yang paling berpengaruh pada pohon regresi adalah variabel "Umur". Pada pohon regresi diperoleh proporsi varian yang dijelaskan oleh model sebesar 89.7%. Setelah dilakukan *pruning tree*, nilainya turun menjadi 86.9%.

Penurunan nilai ketepatan klasifikasi dan proporsi varian yang dijelaskan oleh model setelah dilakukan *pruning tree* disebabkan karena ketelitian dan kekuratan pohon lebih tinggi sebelum dilakukan *pruning tree*. Nilai sebelum dilakukan *pruning tree* menunjukkan bahwa model pohon klasifikasi dan pohon regresi yang dibangun adalah cukup baik.

KESIMPULAN

Dari analisis hasil penelitian ini dapat diberikan kesimpulan sebagai berikut :

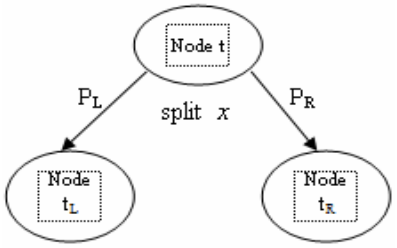
1. Mayoritas penderita skizofrenia di RSJKO adalah sebagai berikut :
   a. Jenis kelamin laki-laki
   b. Diagnosa Skizofrenia Paranoid
   c. Menggunakan Surat Keterangan Tidak Mampu
   d. Tingkat Pendidikan lulusan SD
   e. Tidak memiliki Pekerjaan
   f. Status Belum Kawin
   g. Daerah Asal Kota Bengkulu
2. Variabel yang paling berpengaruh pada pohon klasifikasi adalah "Daerah Asal".
3. Variabel yang paling berpengaruh pada pohon regresi adalah "Umur".
4. Pohon klasifikasi dengan menggunakan kaidah Gini setelah dilakukan *pruning tree* memiliki ketepatan klasifikasi 86.9%, sedangkan kaidah Twoing 85.6%.
5. Proporsi varian yang dijelaskan oleh model pohon regresi adalah 89.7%, setelah *pruning tree* menjadi 86.9%.
6. *Pruning tree* dilakukan untuk menangani pohon dengan kompleksitas yang tinggi.



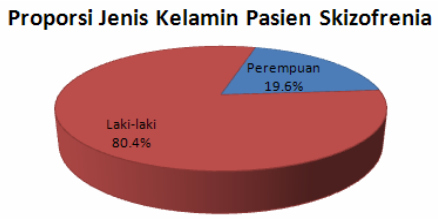
Gambar 1. Definisi *Classifier* pada suatu fungsi $d(x)$

		Node						
		1	2	⋯	s	⋯	t	
Kelas	1				$N_1(s)$			
	2				$N_2(s)$			
	⋮				⋮			
	j	$N_j(1)$	$N_j(2)$	⋯	$N_j(s)$	⋯	$N_j(t)$	N_j
	⋮				⋮			
	k				$N_k(s)$			
					$N(s)$			N

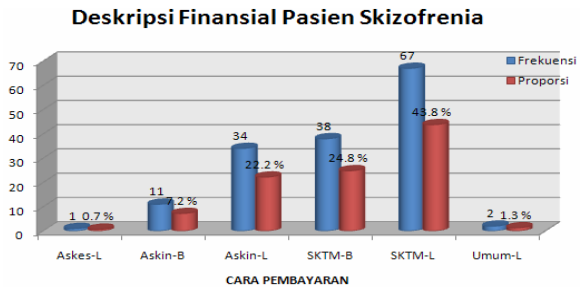
Gambar 2. Ilustrasi Perhitungan Rumus Pohon Klasifikasi



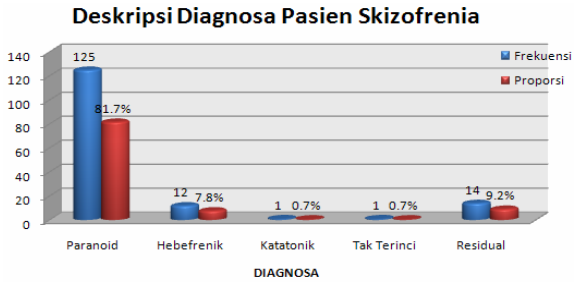
Gambar 3. Algoritma pemisahan pada CART



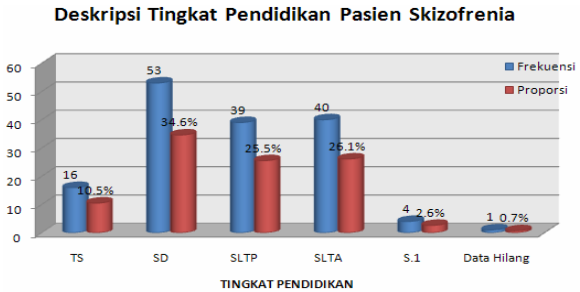
Gambar 4. Grafik Deskripsi Jenis Kelamin Pasien Skizofrenia



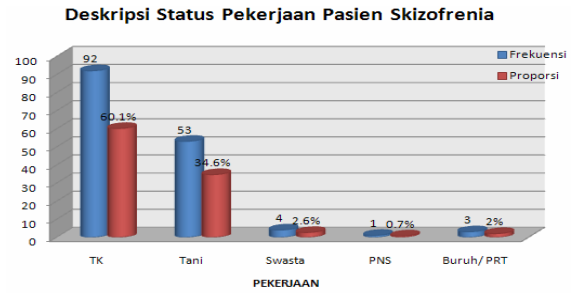
Gambar 5. Grafik Deskripsi Finansial Pasien Skizofrenia



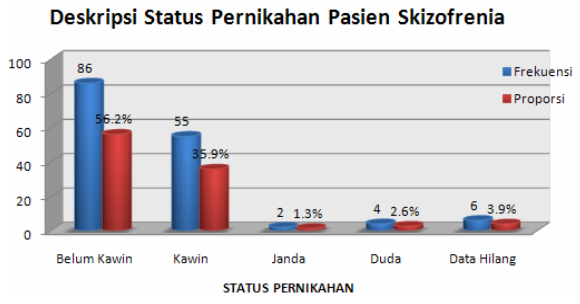
Gambar 6. Grafik Deskripsi Diagnosa Pasien Skizofrenia



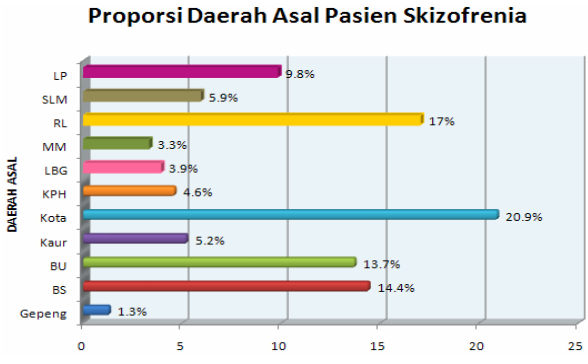
Gambar 7. Grafik Deskripsi Tingkat Pendidikan Pasien Skizofrenia



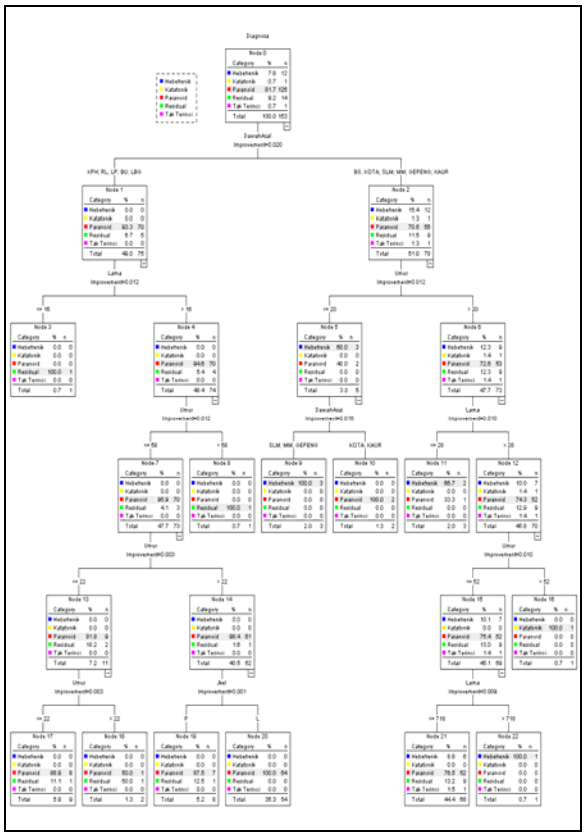
Gambar 8. Grafik Deskripsi Status Pekerjaan Pasien Skizofrenia



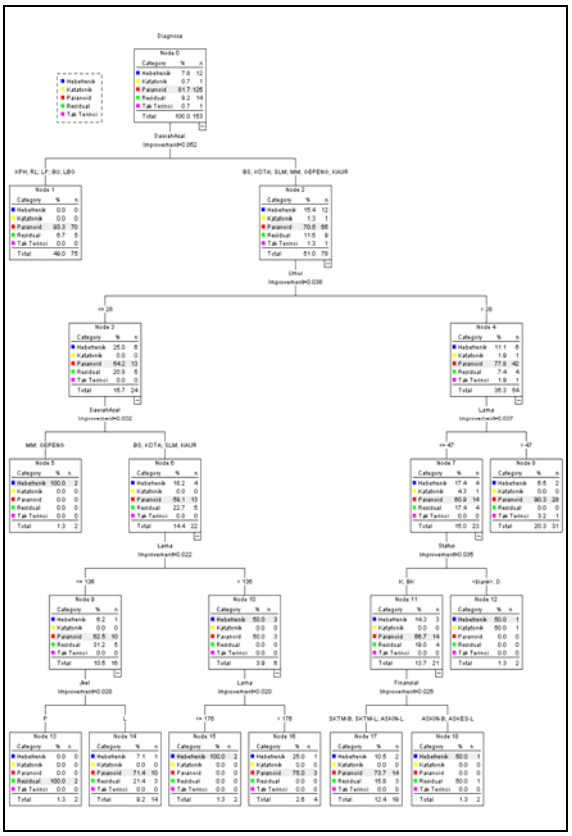
Gambar 9. Grafik Deskripsi Status Pernikahan Pasien Skizofrenia



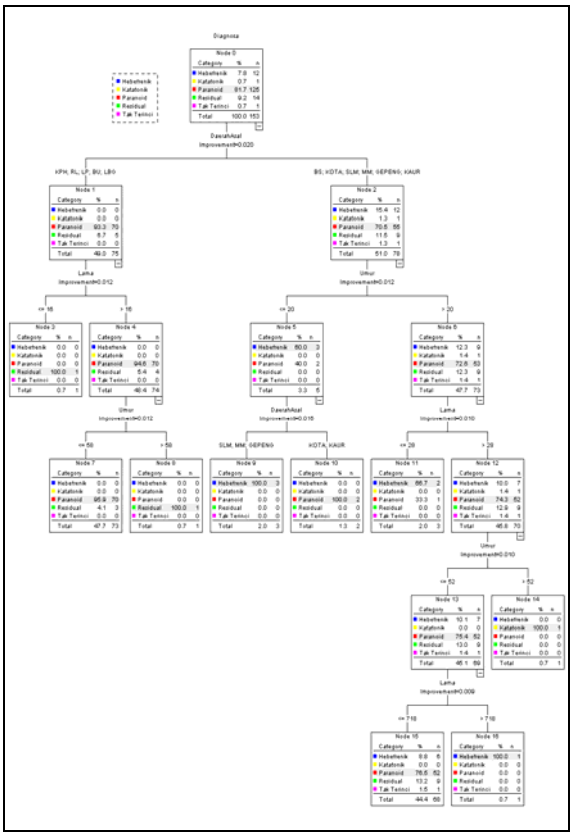
Gambar 10. Grafik Deskripsi Daerah Asal Pasien Skizofrenia



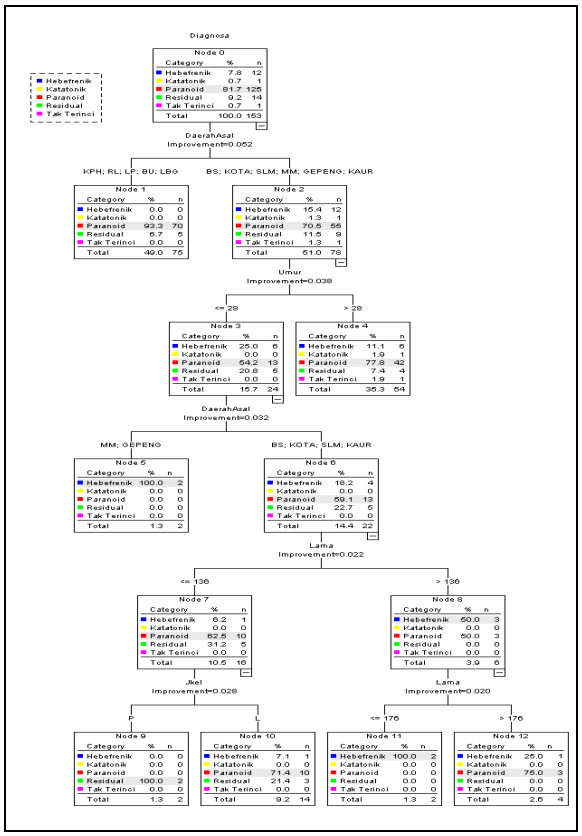
Gambar 11. Pohon Klasifikasi dengan menggunakan kaidah Gini



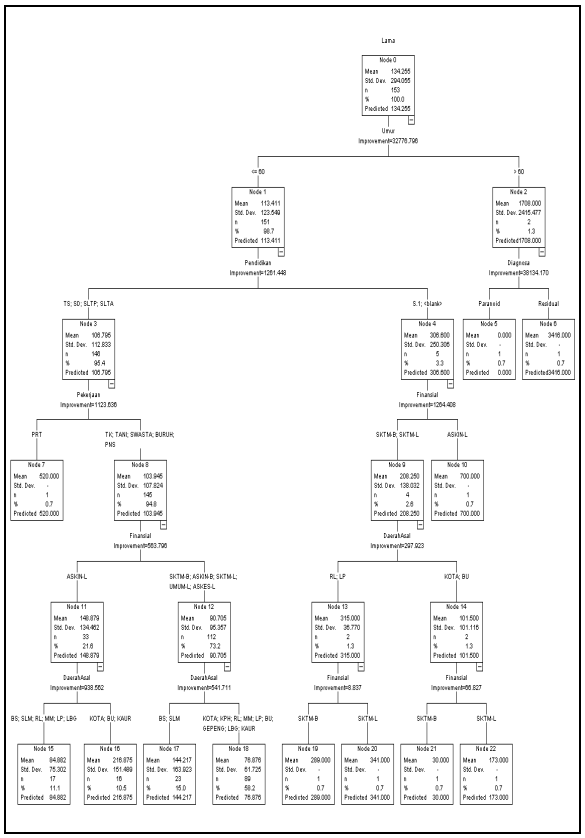
Gambar 12. Pohon Klasifikasi dengan menggunakan kaidah Twoing



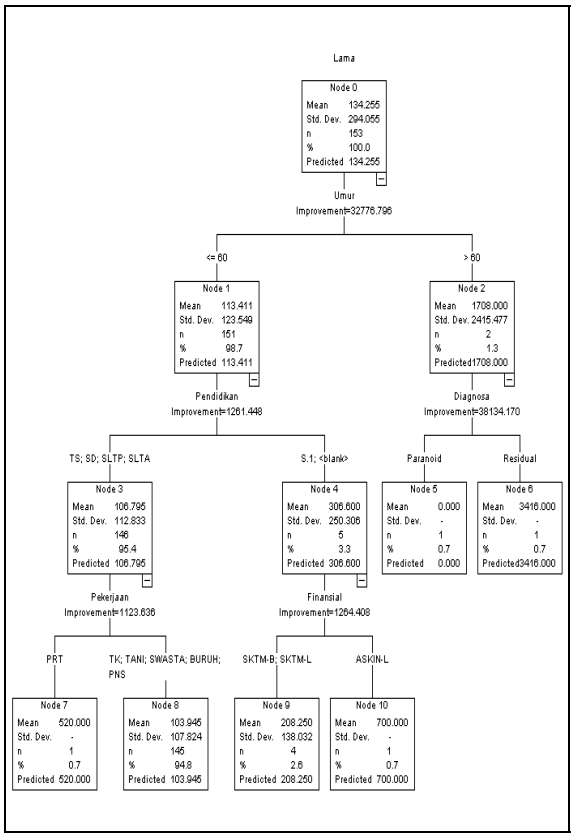
Gambar 13. Pruning tree pada pohon klasifikasi dengan menggunakan kaidah Gini



Gambar 14. Pruning tree pada pohon klasifikasi dengan menggunakan kaidah Twoing



Gambar 15. Pohon Regresi dengan menggunakan cross-validation



Gambar 16. Pruning Tree pada Pohon Regresi

DAFTAR PUSTAKA

[1] Admin. 2004. *Skizofrenia.* Fakultas Psikologi Universitas Muhammadiyah Surakarta. **www.ums.ac.id/fakultas/psikologi/ modules.php?name=News&file=article&sid=14 – 24**

[2] Anonim. 2004a. *Preventing Schizophrenia*. www.schizophrenia.com/prev1.htm

[3] Anonim. 2004b. *Symptoms and Diagnosis of Schizophrenia*. www.schizophrenia.com/ami/diagnosis/organic.html

[4] Anonim. 2005. *Faktor-faktor Penyebab Gangguan Jiwa.* www.balipost.co.id/balipostcetaK/2005/8/3/k4.htm

[5] Anonim. 2007a. *Classification and Regression Trees*. **www.cems.uwe.ac.uk/~rblawton/classification%20and%20regression%20trees.ppt**

[6] Anonim. 2007b. *Decision Tree Learning*. www.wikipedia.com

[7] Anonim. 2007c. *Kesehatan Jiwa*. **faperta.ugm.ac.id/articles/kesehatan_jiwa.pdf**

[8] Anonim. 2007d. *Skizofrenia*. **id.wikipedia.org/wiki/Skizofrenia**

[9] Anonim. 2007e. *Tree Structured Classifier*. **www.stat.psu.edu/~jiali/ course/stat597e/notes2/trees.pdf**

[10] Anonim. 2007f. *SPSS Classification Trees™ 13.0*. **https://www.washington.edu/uware/ spss/docs/ClassificationTrees13.0.pdf**

[11] Andriyashin, A. 2005. *Financial Applications of Classification and Regression Trees*. Center of Applied Statistics and Economics Humboldt University. Berlin. **edoc.hu-berlin.de/master/andriyashin-anton-2005-03-24/PDF/andriyashin.pdf**

[12] Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone. 1984. *Classification and Regression*

Trees. Monterey, California, U.S.A.: Wadsworth, Inc.

[13] Chee, J. C. 2002. *Partitioning Groups using Classification and Regression Tree in BiomedicalResearch.* **binfo.ym.edu.tw/edu/seminars/pdf/CART_YMUBC.pdf**

[14] Devore, J. L. 2004. *Probability and Statistics for Engineering and The Sciences*. Sixth Edition. Thomson Brooks/Cole : Canada.

[15] Feldman, D. & S. Gross. 2003. *Mortgage Default : Classification Trees Analysis.* The Pinhas Sapir Center for Development. Tel-Aviv University. **sapir.tau.ac.il/papers/sapir-wp/3-03.pdf**

[16] Hadi, S. 1977. *Metodologi Research*. Jilid II. Yayasan Penerbitan Fakultas Psikologi Universitas Gajah Mada. Yogyakarta.

[17] Hens, N., L. Bruckers, M. Arbyn, M. Aerts & G. Molenberghs. 2002. *Classification Tree Analysis of Cervix Cancer Screening in the Belgian Health Interview Survey 1997*. Arch Public Health. 60 : 275-294. www.iph.fgov.be/aph/pdf/aphfull60_275_294.pdf

[18] Johnson, R. A. 2000. *Probability and Statistics for Engineering*. Sixth Edition. Prentice Hall International, Inc. New Jersey : USA.

[19] Kandouw, A., JES Kandouw, S. D. Elvira & I. Ariawan. 2007. *Proporsi Gangguan Depresi pada Penyalahguna Zat yang Menjalani Rehabilitasi di RS Marzoeki Mahdi*. Cermin Dunia Kedokteran No. 156. Departemen Psikiatri Fakultas Kedokteran Universitas Indonesia. Jakarta. **www.kalbe.co.id/.../ 156_08ProporsiGangguanDepresipenyalahguna.html**

[20] Kucukkocaoglu, G. & O. Sezgin. 2007. *IPO Mechanism Selection by Using Classification and Regression Trees (CART)*. Başkent University, Faculty of Economics and Administrative Sciences, Management Department, Bağlıca, Ankara. Turkey.

[21] Kuntjoro, Z. S. 2002. *Mengenal Gangguan Jiwa Pada Lansia*. www.e-psikologi.com/usia/140502.htm

[22] Kutner, M.H., C.J. Nachtsheim, J. Neter & W. Li. 2005. *Applied Linear Statistical Models*. Fifth Edition. Mc Graw-Hill International Edition.

[23] Lazarusli, C. A. 2005. *Dia Sanggup Melakukannya Berjam-jam*. Katarsis Edisi I Maret 2005. www.unika.ac.id/fakultas/psikologi/katarsis/katarsis.pdf

[24] Rachmat, A. 2007. *Manipulasi Tree*. Handout Struktur Data Prodi Teknik Informatika UKDW.

[25] Sezgin, O. 2006. *Statistical Methods In Credit Rating*. Department of Financial Mathematics. The Middle East Technical University. Turkey **www3.iam.metu.edu.tr/iam/ images/2/21/Özgesezginthesis.pdf**

[26] Timofeev, R. 2004. *Classification and Regression Trees (CART) Theory and Applications*. Center of Applied Statistics and Economics Humboldt University. Berlin. **edoc.hu-berlin.de/master/ timofeev-roman-2004-12-20/PDF/timofeev.pdf**

[27] Yohannes, Y. & P. Webb. 1999. *Classification and Regression Trees, CART™: A User Manual For Identifying Indicators of Vulnerability to famine and Chronic Food Insecurity*. International Food Policy Research Institute. Washington, U.S.A. **www.ifpri.org/pubs/microcom/micro3.pdf**

[28] Zambon, M., R. Lawrence, A. Bunn & S. Powell. 2006. *Effect of Alternative Splitting Rules on Image Processing Using Classification Tree Analysis*. Photogrammetric Engineering & Remote Sensing Vol. 72, No. 1 : 25–30.

[29] Zhou, Z. H. 2007. *Data Mining Chapter 5 : Classification and Regression*. Department of Computer Science and Technology Nanjing University. China

Analisis Kapabilitas Proses Dengan Pendekatan Bagan Kendali

Rita Trijayanti[1], Sigit Nugroho[2], Jose Rizal[2]

[1]Alumni Jurusan Matematika Fakultas MIPA Universitas Bengkulu
[2]Staf Pengajar Jurusan Matematika Fakultas MIPA Universitas Bengkulu

ABSTRAK

Kualitas suatu produk terjamin bukan dari output yang dihasilkan tetapi pada saat produk tersebut sedang diproses. Oleh sebab itu, pengendalian proses (*process control*) merupakan aspek yang sangat penting dalam menghasilkan produk atau jasa.

Analisis kapabilitas proses merupakan suatu analisis untuk memprediksi seberapa konsisten proses memenuhi spesifikasi yang telah ditentukan konsumen. Analisis kapabilitas proses yang baik apabila proses produksi berada di dalam batas spesifikasi mutu yang telah ditentukan. Beberapa teknik yang dapat digunakan dalam analisis kapabilitas proses yaitu bagan kendali dan perencanaan eksperimen.

Penelitian ini bertujuan untuk menentukan apakah suatu produksi telah memenuhi spesifikasi yang telah ditentukan. Data yang digunakan dalam skripsi ini adalah data hasil pengukuran panjang, lebar dan tinggi tahu yang diproduksi oleh *home* industri X. Teknik yang digunakan dalam menganalisis kapabilitas proses yaitu teknik bagan kendali. Hasil penelitian menunjukkan bahwa untuk data panjang tahu, lebar tahu, dan tinggi tahu setelah dilakukan Analisis Kapabilitas Proses diperoleh hasil bahwa proses produksi pada *home* industri X tersebut tidak *capable*.

Kata Kunci : Analisis Kapabilitas Proses, Bagan Kendali Rata-rata, dan Bagan Kendali Range.

PENDAHULUAN

Latar Belakang

Mutu dari suatu produk yang dihasilkan dari suatu proses produksi sangatlah penting bagi kemajuan/kelangsungan suatu perusahaan atau industri. Kualitas suatu produk terjamin bukan dari output yang dihasilkan tetapi pada saat produk tersebut sedang diproses. Oleh sebab itu, pengendalian proses (*process control*) merupakan aspek yang sangat penting dalam menghasilkan produk atau jasa. Proses Statistik Kontrol (*Statistical Process Control*) merupakan alat bantu dalam mengendalikan proses produksi secara statistik. Dengan menggunakan proses statistik kontrol ini, diharapkan proses-proses yang dijalankan oleh suatu perusahaan dalam kondisi terkendali. Proses statistik kontrol diperlukan karena produk atau jasa yang dihasilkan bervariasi.

Variasi atau variabilitas adalah ketidakseragaman hasil dari suatu produk/jasa yang tidak memenuhi spesifikasi. Variasi proses terdiri dari dua macam penyebab, yaitu penyebab umum (*common cause*) dan penyebab khusus (*assignable cause* atau *special*

cause). Pelanggan menuntut produk/jasa dengan variabilitas yang sekecil-kecilnya. Oleh karena itu, perusahaan harus melakukan *improvement* dan memastikan bahwa variasi/varabilitas karakteristik mutu produk/jasanya masih dalam batas-batas yang masih bisa ditoleransi pelanggan (berada dalam spesifikasi). Untuk menguji variabiltas dalam karakteristik proses dan apakah proses mampu menghasilkan produk/jasa yang sesuai dengan spesifikasi, maka digunakanlah Analisis Kapabilitas Proses (Ariani, 2005).

Analisis kapabilitas proses merupakan suatu analisis untuk memprediksi seberapa konsisten proses memenuhi spesifikasi yang telah ditentukan konsumen. Analisis kapabilitas proses yang baik apabila proses produksi berada di dalam batas spesifikasi mutu yang telah ditentukan. Beberapa teknik yang dapat digunakan dalam analisis kapabilitas proses yaitu bagan kendali dan perencanaan eksperimen.

Bagan kendali merupakan suatu analisis multivariat yang digunakan untuk mendeskripsikan keadaan dari suatu proses (pengukuran, produksi) dengan tampilan berupa grafik dua dimensi yang memuat garis-garis kendali, Batas Kendali Atas (BKA) dan Batas Kendali Bawah (BKB) sebagai dasar pengukuran dalam menentukan proses dalam keadaan terkendali atau tidak. Bagan kendali adalah bagan yang digunakan untuk mengendalikan proses produksi yang terbentuk dari data variabel.

Data dalam penelitian ini merupakan data primer yang diambil dari pengukuran tahu yang diproduksi oleh *home* industri X. Pengolahan data menggunakan bantuan Software Minitab versi 15.

Tujuan

Tujuan dari penelitian ini adalah untuk menentukan apakah produksi tahu yang diproduksi oleh *home* industri X tersebut telah memenuhi keinginan/spesifikasi dari konsumen (*capable*).

TINJAUAN PUSTAKA

Bagan Kendali

Bagan kendali merupakan suatu analisis statistik multivariat yang digunakan untuk mendeskripsikan keadaan dari suatu proses (pengukuran, produksi) dengan tampilan berupa grafik dua dimensi yang memuat garis-garis kendali Batas Kendali Atas (BKA) dan Batas Kendali Bawah (BKB) sebagai dasar pengukuran dalam menentukan proses dalam keadaan terkendali atau tidak. Shewhart, (1924) pertama kali memperkenalkan bagan kendali untuk mendeteksi penyebab-penyebab khusus dari suatu proses.

Tujuan dari bagan kendali antara lain: mendiagnosis, mengoreksi, dan menyingkirkan variabilitas dalam suatu proses dengan harapan kualitas hasil akhir proses dapat memenuhi suatu standar yang ditentukan. Sedangkan manfaat dari bagan kendali, antara lain :

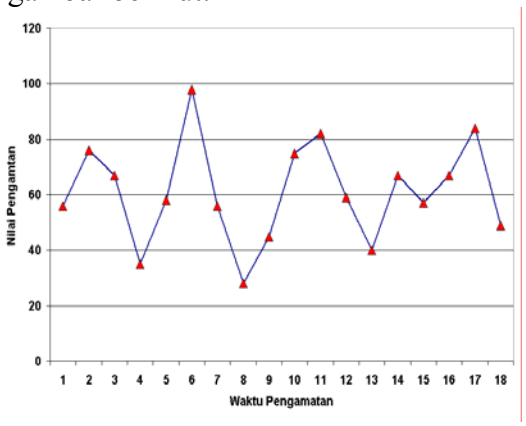
1. Untuk memonitor variabilitas hasil pengukuran parameter proses.
2. Mengidentifikasi penyimpangan dini dan mengambilan tindakan sebelum *process out of control*.

Variabilitas dalam karakteristik kualitas produksi terdiri dari dua macam penyebab yaitu penyebab umum (*random cause* atau *chance cause*) dan penyebab khusus (*assignable cause* atau *special cause*). Penyebab umum timbul dari penyimpangan dalam bahan baku,

kondisi emosional karyawan, penurunan kinerja mesin, penurunan suhu udara, naik turunnya kelembaban udara. Penyebab umum ini sudah melekat pada proses. Sedangkan penyebab khusus dapat ditimbulkan dari tiga sumber, yaitu: mesin yang dipasang dengan tidak baik, kesalahan operator (*human error*), dan bahan baku yang cacat.
Komponen-komponen yang terdapat dalam sebuah bagan kendali adalah:

1. Waktu pengamatan yang menjadi sumbu horizontal.
2. Ukuran statistik yang menjadi sumbu vertikal.
3. Batas kendali atas *(Upper Control Limit).*
4. Batas kendali bawah *(Lower Control Limit).*
5. Garis tengah *(Center Line).*
6. Titik-titik sampel yang menggambarkan keadaan proses tiap pengamatan.

Seperti yang tampak pada gambar berikut:



Dari bagan kendali di atas, dapat dilihat apakah proses terkendali atau tidak terkendali. Proses dikatakan terkendali bila titik-titik terdistribusi acak di sekitar garis tengah dan semua titik berada di dalam batas kendali (BKA dan BKB). Sedangkan proses dikatakan tidak terkendali pada suatu interval waktu tertentu bila terdapat titik-titik yang berada di luar batas kendali pada saat interval tersebut.

Asumsi Pada Bagan Kendali

Asumsi yang mendasari bagan kendali adalah asumsi kenormalan (data yang diukur harus berdistribusi Normal). Dalam statistika peranan Distribusi Normal sangat penting, karena dalam hampir semua prinsip dan teknik analisis statistika dikembangkan dari konsep distribusi Normal. Jadi sebelum dilakukan pengolahan analisis statistik, perlu untuk melakukan pengujian bahwa data sampel yang di ambil berdistribusi normal atau mendekati normal (menguji asumsi kenormalan data).

Definisi 1

X peubah acak kontinu berdistribusi normal dengan rataan μ dan variansi $\sigma^2 < \infty$, disingkat $X \sim N(\mu, \sigma^2)$, jika fungsi kepekatan peluangnya adalah:

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad -\infty<x,\mu<\infty \quad \sigma>0.$$

Pengamatan terhadap fungsi kepekatan peluang dari $X\sim N\left(\mu,\sigma^2\right)$ menunjukkan bahwa grafik fungsi fungsi $f(x)$ bersifat:

1. Simetris terhadap sumbu $x=\mu$.
2. $f(x)$ maksimum di $x=\mu$ dan $f(\mu)=\frac{1}{\left(\sigma\sqrt{2\pi}\right)}$.
3. Jika $x\to-\infty$ *atau* $x\to\infty$, maka $f(x)\to 0$. Jadi sumbu x merupakan sumbu asimtot.
4. $x=\mu-\sigma$ dan $x=\mu+\sigma$ merupakan absis-absis titik belok.

Untuk mengetahui suatu variabel acak X berdistribusi normal atau tidak yaitu dengan melakukan pengujian hipotesis kenormalan. Prosedur umum dalam melakukan pengujian hipotesis antara lain:

1. Tulis hipotesis nol (H_0).
2. Pilih hipotesis tandingan (H_1) yang sesuai.
3. Pilih taraf kebeartian berukuran α.
4. Pilih uji statistik yang sesuai dan tentukan daerah kritisnya. (bila keputusan didasarkan pada nilai-p, maka tidaklah perlu menyatakan daerah kritisnya).
5. Hitung nilai uji statistik dari data sampel.

Kesimpulan: tolak H_0 bila nilai statistik uji berada di dalam daerah kritis (atau, bila nilai-p hitung lebih kecil atau sama dengan taraf keberartian yang ditentukan). Sebaliknya, terima H_0.

Definisi 2

Nilai-p adalah taraf keberartian terkecil sehingga nilai uji statistik yang diamati masih berarti.

Pendekatan nilai-p sebagai alat bantu dalam pengambilan keputusan cukuplah wajar, karena hampir semua uji hipotesis yang dilakukan komputer menampilkan nilai statistik uji beserta nilai-p.

Apabila data sampel tidak berdistribusi normal, maka salah satu teknik untuk mengatasinya yaitu dengan mentransformasikan data. Transformasi yang paling sering digunakan adalah transformasi *Tukey* atau menggunakan transformasi *Box-Cox* dalam aplikasi Minitab dan diharapkan data hasil transformasi berdistribusi normal (mendekati normal).

Analisis Pola Pada Bagan Kendali

Beberapa kriteria yang dapat diterapkan untuk mengetahui apakah proses tidak terkendali, yaitu:

- Satu atau beberapa titik di luar batas kendali atas dan batas kendali bawah.
- Ada tujuh titik berturut-turut di atas atau di bawah garis tengah.
- Suatu giliran dengan paling sedikit tujuh atau delapan titik, dengan macam giliran dapat berbentuk giliran naik atau giliran turun, giliran diatas atau dibawah garis tengah.

Dalam membentuk suatu bagan kendali didasarkan pada dua kategori yaitu:

1. Bagan Kendali Variabel (*Variable Control Chart*)
   Bagan kendali variabel adalah bagan kendali dimana data yang dikumpulkan dan yang akan dianalisa adalah data variabel (data yang diperoleh dengan melakukan pengukuran dengan alat ukur). Dalam kasus diagram kendali variabel, pengendalian ditekankan pada kecenderungan rata-rata dan variabilitas.
   Didalam bagan kendali variabel terdapat 3 metode pembuatan bagan kendali berdasarkan pada besarnya pengumpulan data (n) yaitu :
   a. Untuk $n = 1$, menggunakan $\overline{X}$ *moving range chart* .
   b. Untuk $2 \le n \le 9$, menggunakan $\overline{X} - R$ *chart* .
   c. Untuk $n \ge 10$, menggunakan $\overline{X} - S$ *chart* .
   Menurut Besterfield *dalam* Ariani (2005), dalam melakukan pengendali kualitas proses statistik untuk data variabel diperlukan beberapa langkah, yaitu:
      a. Pemilihan karakterisik kualitas.
      b. Pemilihan sub kelompok.
      c. Pengumpulan data.
      d. Penentuan garis pusat.
      e. Penyusunan revisi terhadap garis pusat dan batas-batas pengendali.
      f. Interpretasi terhadap pencapaian tujuan.

2. Bagan Kendali Sifat (*Attribute Control Chart*)
   Bagan kendali sifat adalah bagan kendali dimana data yang dikumpulkan dan yang akan dianalisa adalah data yang diperoleh dengan cara melakukan penghitungan. Dalam kasus bagan kendali sifat, proporsi cacat dari sampel menjadi acuan.
   Didalam bagan kendali sifat terdapat 4 metode pembuatan bagan kendali berdasarkan pengumpulan data didalam pengecekan *defective* (cacat) yaitu :
   a. $p - chart$ adalah bagan kendali sifat dimana data yang dikumpulkan tergolong diterima atau ditolak (mengecek *defective*) dan dalam setiap pengamatan besar subgrupnya berbeda.
   b. $np - chart$ adalah bagan kendali sifat dimana data yang dikumpulkan digolongkan diterima atau ditolak (mengecek *defective/nonconforming*) dan dalam setiap pengamatan besarnya subgrup sama.
   c. $u - chart$ adalah bagan kendali sifat dimana data yang dikumpulkan adalah *defect/nonconformity* dalam subgrup dimana disetiap pengamatan besar subgrupnya berbeda.

d. $c-chart$ adalah bagan kendali sifat dimana data yang dikumpulkan adalah *defect–defect/noncomformity* dalam subgrup dimana dalam setiap pengamatan besar subgrupnya sama.

Bagan Kendali $\overline{X}$ dan *R*

Bagan kendali rata-rata ($\overline{X}$) dan jarak (R) merupakan dua bagan kendali yang saling membantu dalam mengambil keputusan mengenai kualitas proses. Bagan kendali rata-rata digunakan untuk melihat apakah proses masih berada dalam batas kendali atau tidak. Sedangkan bagan kendali jarak (*range*) digunakan untuk mengetahui tingkat keakurasian atau ketepatan proses yang diukur dengan mencari *range* dari sampel yang diambil dalam setiap observasi. Bagan kendali rata-rata dan bagan kendali *range* juga digunakan untuk mengetahui dan menghilangkan penyebab khusus yang membuat terjadinya penyimpangan.

Batas-batas Kontrol Pada Bagan Kendali $\overline{X}$ dan *R*

Parameter untuk bagan kendali $\bar{x}$ adalah:

$$BKA = \bar{\bar{x}} + \frac{3}{d_2\sqrt{n}}\overline{R}$$

$$Garis\ Tengah = \bar{\bar{x}}$$

$$BKB = \bar{\bar{x}} - \frac{3}{d_2\sqrt{n}}\overline{R}$$

misalkan: $A_2 = \frac{3}{d_2\sqrt{n}}$

maka, persamaan dapat disederhanakan menjadi:

$$BKA = \bar{\bar{x}} + A_2\overline{R}$$

$$Garis\ Tengah = \bar{\bar{x}}$$

$$BKB = \bar{\bar{x}} - A_2\overline{R}$$

Parameter bagan kendali R adalah:

$$BKA = \overline{R} + 3\hat{\sigma}_R = \overline{R} + 3\ d_3\frac{\overline{R}}{d_2}$$

$$Garis\ Tengah = \overline{R}$$

$$BKB = \overline{R} - 3\hat{\sigma}_R = \overline{R} - 3\ d_3\frac{\overline{R}}{d_2}$$

misalkan

$$D_3 = 1-3\frac{d_3}{d_2} \text{ dan } D_4 = 1+3\frac{d_3}{d_2}$$

Persamaan dapat disederhanakan menjadi:

$$BKA = D_4 \overline{R}$$

$$Garis\ Tengah = \overline{R}$$

$$BKB = D_3 \overline{R}$$

Bagan Kendali $\overline{X}$ dan *S*

Bagan kendali $\overline{X}$ dan S digunakan apabila ukuran sampel *n* cukup besar, misalnya *n* =10 atau lebih. Bagan kendali *S* (standar deviasi) digunakan untuk mengukur tingkat keakurasian proses suatu produksi.

Batas-batas Kontrol Pada Bagan Kendali $\overline{X}$ dan *S*

Batas kendali atas dan batas kendali bawahnya yaitu:

$$BKA = C_4\sigma + 3\sigma\sqrt{1-C_4^2}$$

$$BKB = C_4\sigma - 3\sigma\sqrt{1-C_4^2}$$

misalkan

$$B_6 = C_4 + 3\sqrt{1-C_4^2} \quad \text{dan} \quad B_5 = C_4 - 3\sqrt{1-C_4^2}$$

Persamaan dapat disederhanakan menjadi:

$$BKA = B_6\sigma$$

$$BKB = \overline{S} - \frac{3\overline{S}\sqrt{(1-C_4)}}{C_4}$$

misalkan

$$B_4 = 1+\frac{3\sqrt{(1-C_4)}}{C_4} \quad \text{dan} \quad B_3 = 1-\frac{3\sqrt{(1-C_4)}}{C_4}$$

persamaan dapat disederhanakan menjadi:

$$BKA = B_4\overline{S}$$

$$Garis\ tengah = \overline{S}$$

$$BKB = B_3\overline{S}$$

nilai B_4 diperoleh dari $B_4 = \frac{B_6}{C_4}$ dan B_3 diperoleh dari $B_3 = \frac{B_5}{C_4}$

untuk bagan kendali $\bar{X}$ adalah Batas-batas kendalinya adalah:

$$BKA = \bar{\bar{x}} + \frac{3\bar{S}}{C_4\sqrt{n}}$$

$$Garis\ tengah = \bar{\bar{x}}$$

$$BKB = \bar{\bar{x}} - \frac{3\bar{S}}{C_4\sqrt{n}}$$

misalkan $A_3 = \frac{3}{C_4\sqrt{n}}$

persamaan dapat disederhanakan menjadi:

$$BKA = \bar{\bar{x}} + A_3\bar{S}$$

$$Garis\ tengah = \bar{\bar{x}}$$

$$BKA = \bar{\bar{x}} - A_3\bar{S}$$

Interpretasi Bagan Kendali

Apabila data sampel dalam observasi tersebut berada dalam kondisi *out of statistical control*, maka langkah selanjutnya adalah mencari penyebab kesalahan dengan menggunakan teknik perbaikan kualitas yaitu bagan kendali. Apabila kesalahan atau penyimpangan disebabkan oleh sebab umum, maka kondisi dianggap *in statistical control.* Namun apabila kesalahan atau penyimpangan disebabkan oleh sebab khusus berarti kondisi dianggap *out of statistical control* dan harus segera diselesaikan dengan cara mengeliminasi data tersebut.

Apabila data telah berada pada kondisi *in statistical control,* maka yang harus dilakukan selanjutnya adalah menguji kapabilitas proses. Pengujian ini bertujuan untuk mengetahui apakah proses telah berada dalam batas spesifikasi yang telah ditentukan.
Berikut contoh bagan kendali *out of statistical control* dan *in statistical control* dengan menggunakan data kekuatan meledak botol minuman ringan (Montgomery, 2001).

ANALISIS KAPABILITAS PROSES

Apabila suatu proses sudah stabil atau terkontrol dilihat dari segi variabilitas maupun rata-ratanya maka dapat dikatakan bahwa proses tersebut berjalan dengan baik. Namun, suatu produk tidak hanya cukup dengan kata baik saja, tetapi harus mampu memenuhi keinginan atau spesifikasi dari konsumen. Untuk itu perlu dilakukan analisis lebih lanjut tentang kapabilitas suatu produk yang dilihat dari proses produksinya yang disebut sebagai analisis kapabilitas proses.

Pengertian dan Asumsi

Analisis kapabilitas proses merupakan suatu tahapan yang harus dilakukan dalam mengadakan pengendalian proses statistik (*statistical process control)* dan merupakan suatu studi guna menaksir kapabilitas proses dalam bentuk distribusi probabilitas yang mempunyai bentuk, rata-rata (*mean*), dan penyebaran (*standard deviation*). Analisis kapabilitas proses dilakukan apabila variabilitas dan rata-rata sudah stabil (Pyzdek *dalam* Ariani, 2005).

Analisis kapabilitas proses adalah suatu analisa untuk memprediksi seberapa konsisten proses memenuhi spesifikasi yang ditentukan oleh konsumen. Proses disebut *capable* jika mampu menghasilkan hampir 100% output sesuai spesifikasi. Kapabilitas adalah kemampuan proses untuk menghasilkan output sesuai spesifikasi.

Dalam analisis kapabilitas proses ada dua asumsi yang harus dipenuhi, yaitu:

1. Asumsi kenormalan
   Asumsi kenormalan pada Analisis Kapabilitas Proses mengikuti asumsi kenormalan pada bagan kendali.
2. Data yang diukur terkendali statistik
   Data yang diukur mencerminkan terkendali secara statistik saat diplotkan pada sebuah bagan kendali. Artinya tidak ada data yang keluar dari Batas Kendali Atas (BKA) maupun Batas Kendali Bawah (BKB) atau dengan kata lain data terkontrol.

Rasio Kapabilitas Proses

Salah satu komponen terpenting dalam analisis kapabilitas proses adalah rasio kapabilitas proses atau *Process Capability Ratio* (*PCR*). Rasio kapabilitas proses digunakan untuk mengindikasikan *capable* atau tidaknya suatu proses dengan batasan *capable* apabila *PCR* > 1,33. Rasio kapabilitas proses dibagi menjadi tiga yaitu:

1. *Potential Capability Index* (C_p)
2. *Real Capability Index (*C_{pk})
3. *Mean Capability Index* **(**C_{pm}**)**

***Potential Capability Index* (**C_p**)**

C_p digunakan apabila proses berada dalam batas pengendali statistik dengan bagan kendali proses statistik berdistribusi normal dan mean proses (μ) dianggap sama (terpusat) dengan target (T). Karena μ tidak pernah diketahui maka μ ditaksir oleh $\bar{x}$, dan target merupakan titik tengah dari *BSB* dan *BSA*. Sehingga C_p dapat dihitung dengan rumus:

$$C_p = \frac{BSA - BSB}{6\sigma} \qquad (29)$$

Keterangan:

BSA	:	Batas Spesifikasi Atas
BSB	:	Batas Spesifikasi Bawah

Real Capability Index **(C_{pk})**

Rasio kapabilitas proses C_{pk} dibangun karena C_p tidak cukup memadai untuk kasus $\mu \neq T$. Kasus dimana μ tidak berada ditengah batas spesifikasi sehingga perlu dilakukan proses *centering* (proses pemusatan) yang akan membuat μ berada di tengah batas spesifikasi. Untuk mengkarakteristikkan proses *centering* maka C_{pk} harus dibandingkan dengan C_p yaitu dengan menggunakan spesifikasi satu sisi

$$C_{pa} = \frac{BSA - \mu}{3\sigma}$$

$$C_{pb} = \frac{\mu - BSB}{3\sigma}$$

Keterangan :

C_{pa} adalah rasio kemampuan proses atas

C_{pb} adalah rasio kemampuan proses bawah

sehingga C_{pk} diformulasikan dengan:

$$C_{pk} = \min\left(C_{pa}, C_{pb}\right)$$

$$= \min\left(\frac{BSA - \mu}{3\sigma}, \frac{\mu - BSB}{3\sigma}\right)$$

apabila $\mu < T$

$$\frac{BSA - \mu}{3\sigma} > \frac{\mu - BSB}{3\sigma} \qquad \text{Maka } C_{pk} = \frac{\mu - BSB}{3\sigma}$$

untuk $\mu > T$

$$\frac{BSA - \mu}{3\sigma} < \frac{\mu - BSB}{3\sigma} \qquad \text{Maka } C_{pk} = \frac{BSA - \mu}{3\sigma}$$

Mean Capability Index **(C_{pm})**

Untuk sembarang nilai μ yang berada diantara *BSB* dan *BSA*, C_{pk} berbanding terbalik dengan σ sehingga apabila σ mendekati 0 maka C_{pk} akan semakin besar. Nilai C_{pk} yang besar tidak memberikan semua informasi tentang lokasi mean (μ) di interval *BSB* dan *BSA*. Untuk memperbaiki kekurangan C_{pk} perlu dilakukan rasio kapabilitas proses yang lebih baik yaitu dengan menggunakan C_{pm}.

Formulasi C_{pm} diberikan sebagai berikut:

$$C_{pm} = \frac{BSA - BSB}{6\tau}$$

HASIL DAN PEMBAHASAN

Analisis Kapabilitas Proses untuk Panjang Tahu

Pengujian Kenormalan Panjang Tahu

Prosedur pengujian hipotesis secara statistik terdiri dari beberapa langkah. Langkah-langkah tersebut adalah:

1. Merumuskan hipotesis

H_0 : Data pengamatan rata-rata panjang tahu mengikuti distribusi normal.

H_1 : Data pengamatan rata-rata panjang tahu tidak mengikuti distribusi normal.

2. Menentukan taraf signifikan

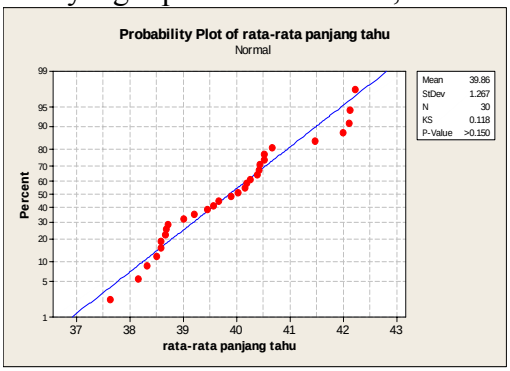
Taraf signifikan yang digunakan dalam penelitian ini adalah $\alpha = 0,05$.

3. Menentukan statistik uji

Statistik uji yang digunakan yaitu statistik uji Kolmogoro-Smirnov (KS). Daerah penolakan untuk statistik uji Kolmogoro-Smirnov (KS) yaitu apabila $KS_{hitung} < KS_{tabel}$ atau jika nilai p-value < α.

4. Menentukan nilai statistik uji dari data sampel

Normalitas univariat terhadap data panjang tahu dalam analisis ini diuji dengan menggunakan Software Minitab versi 15. Hasilnya adalah seperti yang disajikan gambar 4 yaitu gambar uji normalitas panjang tahu. Dengan menguji rata-rata dari panjang tahu nilai statistik Kolmogorov-Smirnov yang diperoleh sebesar 0,118.



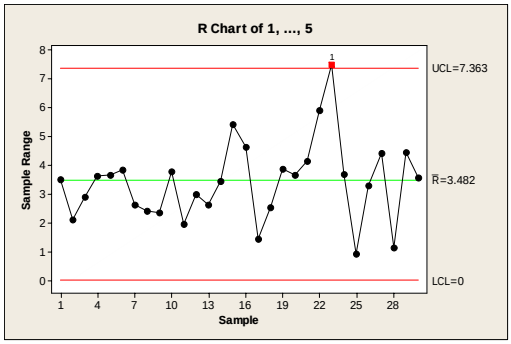
Gambar 4. Uji normalitas panjang tahu

5. Mengambil keputusan atau kesimpulan

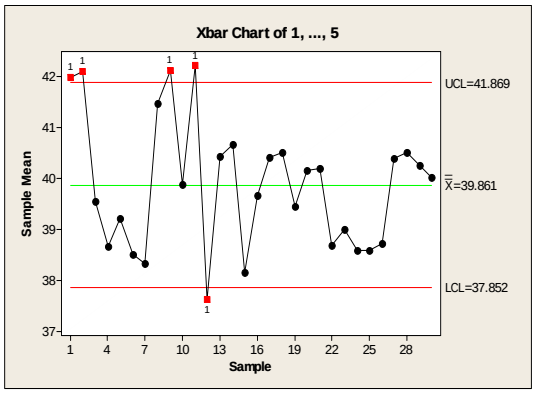
Nilai statistik $KS_{hitung} = 0{,}118$ sedangkan nilai statistik $KS_{tabel} = 0,242$ (uji dua arah) pada lampiran 3 yang artinya $KS_{hitung} < KS_{tabel}$. Oleh karena itu, dapat disimpulkan data pengamatan rata-rata panjang tahu mengikuti distribusi normal.

Hasil Analisis Bagan Kendali Panjang Tahu

Dengan menggunakan Software Minitab 15, diperoleh bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ untuk panjang tahu seperti yang nampak pada ganbar berikut



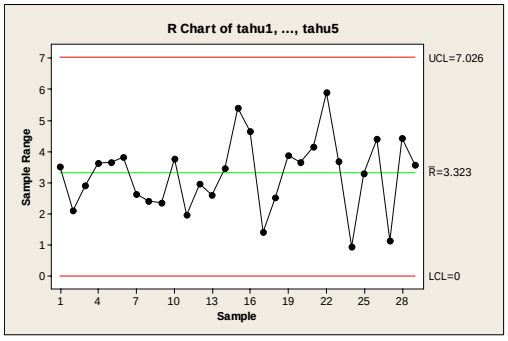
Bagan kendali *Range* (R) panjang tahu



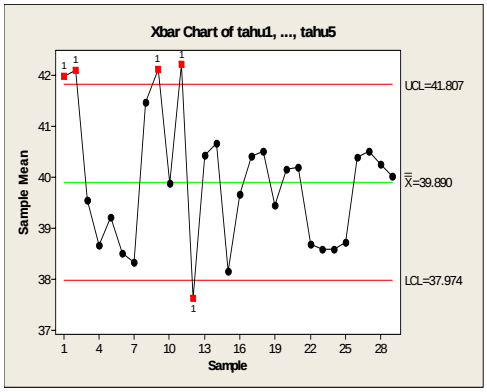
Bagan kendali rata-rata $(\bar{X})$ panjang tahu

Berdasarkan bagan kendali hasil observasi di atas, ternyata untuk bagan kendali *Range* (R) data dari hasil observasi ke-23 (7,84) berada diluar batas kendali dan pada bagan kendali rata-rata $(\bar{X})$ data dari hasil observasi ke-1 (41,89), ke-2 (42,10), ke-9 (42,12), ke-11 (42,21) dan ke-12 (37,62) juga berada diluar batas kendali dan ternyata penyebabnya termasuk dalam sebab khusus (*assignable cause*) sehingga harus dilakukan revisi.

Bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ untuk pengukuran panjang tahu setelah direvisi:



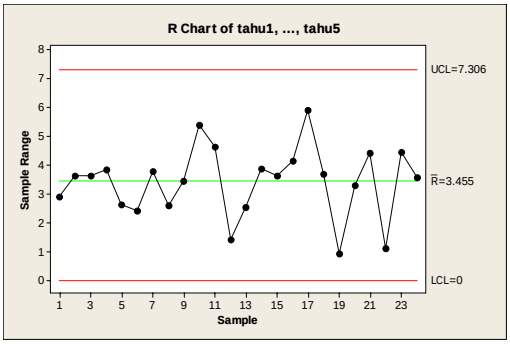
Bagan kendali *Range* (R) panjang tahu setelah revisi pertama



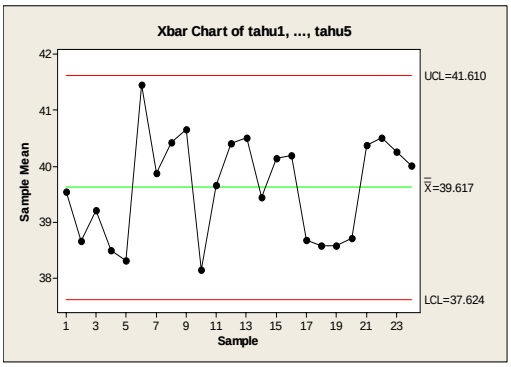
Bagan kendali rata-rata $(\bar{X})$ panjang tahu setelah revisi pertama

Berdasarkan bagan kendali revisi diatas, bagan kendali *Range* (R) sudah terkendali karena tidak ada data yang berada diluar batas kendali, sedangkan untuk bagan kendali rata-rata $(\bar{X})$ proses masih berada di luar kendali karena ada data yang berada diluar batas kendali yaitu data ke-1 (41,98), ke-2 (42,10), ke-9 (42,12), ke-11 (42,21) dan ke-12 (37,62). Untuk itu perlu dilakukan revisi kembali.

Berikut bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ untuk pengukuran oanjang tahu setelah revisi kedua.



Bagan kendali *Range* (R) panjang tahu setelah revisi kedua



Bagan kendali rata-rata $(\bar{X})$ panjang tahu setelah revisi kedua

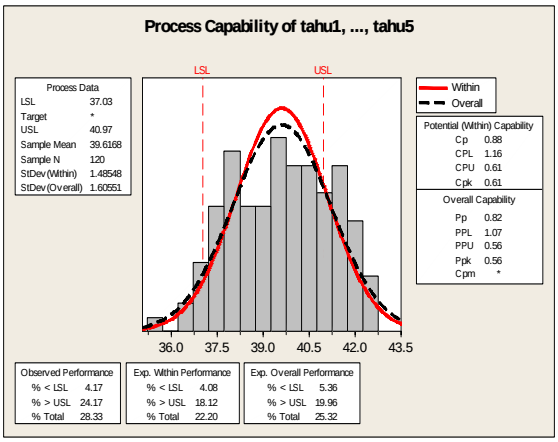
Pada bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ seluruh data hasil observasi berada diantara batas kedali yang menunjukkan bahwa data telah terkendali secara statistik (*in statistical control*). Dengan batas-batas bagan kendali *Range* (R) yaitu untuk batas kendali atas sebesar 7,306, batas kendali bawah sebesar 0 dan garis tengah sebesar 3,455. Sedangkan untuk bagan kendali rata-rata $(\bar{X})$ batas kendali atas sebesar 41,610, batas kendali bawah sebesar 37,624 dan garis tengah sebesar 39,617.

Karena variabilitas dan rata-rata sudah stabil untuk mencapai tujuan penelitian ini maka akan dilakukan analisis lebih lanjut yaitu analisis kapabilitas proses. Analisis ini dilakukan untuk melihat apakah data panjang tahu sudah *capable*. Data yang tersisa untuk dilakukan analisis lebih lanjut ini sebanyak 24 data panjang tahu.

Analisis Kapabilitas Proses Panjang Tahu

Dengan menggunakan Software Minitab versi 15, Analisis kapabilitas ini dilakukan untuk melihat apakah panjang tahu yang di produksi oleh *home* industri X *capable*.

Kasus ini merupakan kasus $\mu \neq T$. Oleh karena itu harus dibandingkan antara C_{pk} dan C_p. Terlihat bahwa $C_{pk} < C_p$ dan C_p =0,88, nilai C_{pk} =0,61. Nilai C_p, C_{pk} semuanya < 1,33. Dapat disimpulkan bahwa proses tersebut tidak *capable*.



Gambar Proses *capability* panjang tahu

Pada gambar diatas terlihat bahwa $\mu \neq T$ dan terdapat titik-titik proses yang berada di luar batas spesifikasi. Seberapa banyak titik-titik proses yang berada di luar batas spesifikasi secara keseluruhan dapat dilihat pada Expected Overall Performance.
Baik secara Within maupun Overall ternyata terdapat titik-titik proses yang berada di luar batas spesifikasi. Ini semakin memperjelas bahwa proses ini tidak *capable*. Untuk Pengujian data Lebar dan data Tinggi tahu langkah-langkah yang digunakan sama seperti langkah-langkah pengujian pada data Panjang tahu.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan analisis dan pembahasan pada bab IV, dapat di tarik kesimpulan sebagai berikut:

1. Pada pengukuran panjang tahu analisis terhadap bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ dilakukan sebanyak 2 kali revisi guna mendapatkan data yang terkendali secara statistk. Batas-batas yang diperoleh dari bagan kendali *Range* (R) panjang tahu yaitu untuk batas kendali atas sebesar 7,306, batas kendali bawah sebesar 0 dan garis tengah sebesar 3,455. Sedangkan nilai untuk batas-batas bagan kendali rata-rata $(\bar{X})$ panjang tahu yaitu untuk batas kendali atas sebesar 41,610, batas kendali bawah sebesar 37,624 dan garis tengah sebesar 39,617. Data yang tersisa dan yang akan dianalisis lebih lanjut dengan menggunakan analisis kapabilitas proses sebanyak 24 data pengamatan panjang tahu.
2. Untuk pengukuran lebar tahu analisis terhadap bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ dilakukan sebanyak satu kali revisi guna mendapatkan data yang

terkendali statistik. Batas-batas yang diperoleh dari bagan kendali *Range* (R) yaitu untuk batas kendali atas sebesar 35,87, batas kendali bawah sebesar 0 dan garis tengah sebesar 33,634. Sedangkan untuk bagan kendali rata-rata $(\bar{X})$ batas kendali atas sebesar 41,610, batas kendali bawah sebesar 37,624 dan garis tengah sebesar 39,617. Data yang tersisa dan yang akan dianalisis lebih lanjut dengan menggunakan analisis kapabilitas proses sebanyak 28 data pengamatan lebar tahu.

3. Untuk pengukuran tinggi tahu analisis terhadap bagan kendali *Range* (R) dan bagan kendali rata-rata $(\bar{X})$ dilakukan sebanyak satu kali revisi guna mendapatkan data yang terkendali secara statistik. Batas-batas yang diperoleh yaitu untuk bagan kendali *Range* (R) batas kendali atas sebesar 7,963, batas kendali bawah sebesar 0 dan garis tengah sebesar 3,767. Sedangkan untuk bagan kendali rata-rata $(\bar{X})$ batas kendali atas sebesar 32,131, batas kendali bawah sebesar 27,787 dan garis tengah sebesar 29,959. Data yang tersisa dan yang akan dianalisis lebih lanjut dengan menggunakan analisis kapabilitas proses sebanyak 28 data pengamatan tinggi tahu.
4. Hasil analisis kapabilitas proses adalah sebagai berikut untuk panjang tahu nilai C_p =0,88 dan C_{pk} =0,61 untuk lebar tahu nilai C_p =0,89 dan C_{pk} =0,57 dan untuk tinggi tahu nilai C_p =0,88 dan C_{pk} =0,5. Nilai C_p dan C_{pk} lebih kecil dari 1,33. untuk itu data panjang, lebar dan tinggi tidak *capable.*

Saran

Beberapa saran yang dapat diberikan penulis sebagai bahan penelitian lebih lanjut adalah sebagai berikut:

1. Karena Kemampuan Proses yang diperoleh tidak *capable*, maka disarankan kepada Produsen tahu agar kualitas tahu terutama dari panjang, lebar, dan tinggi tahu ditingkatkan supaya memenuhi spesifikasi yang telah ditentukan konsumen.
2. Bagi penelitian di masa yang akan datang, disarankan sebaiknya sebelum sampel data diolah, sampel yang terbesar dan yang terkecil dalam setiap pengamatan dihilangkan guna mengurangi terjadinya variabilitas dalam proses kemudian dianalisis sesuai dengan prosedur analisis kapabilitas proses.
3. Teknik yang digunakan dalam menyelesaikan Analisis Kapabilitas Proses adalah teknik bagan kendali, teknik rancangan percobaan bisa digunakan untuk menentukan *capable* atau tidaknya suatu produk untuk bahan skripsi mendatang.
4. Karena panjang, lebar dan tinggi tahu merupakan karakteristik kualitas yang saling berhubungan maka sebaiknya menggunakan bagan kendali multivariat untuk menyelesaikan analisis kemampuan proses.

DAFTAR PUSTAKA

Anonim. 2006. SPC (Statistical Process Control) http://www.unido.org/en/doc/4268. 18 Mei 2007. 08:50 WIB.

Ariani, D.W. 2005, *Pengendalian Kualitas Statistik (Pendekatan Kuantitatif dalam Manajemen Kualitas)*. Yogyakarta: C.V. Andi Offset.

Grant, E.L dan R.S, Leavenworth. 1989, *Pengendalian Mutu Statistik*. Edisi ke-1. Jakarta:Erlangga.

Haibaho, C dan Nawalo, W. 1988, *Metoda Statistik Untuk Peningkatan Mutu*. Jakarta: P.T Mediyatama Asrana Perkasa.

Herrhyanto, N. 2003, *Statistik Matematika Lanjutan*. Bandung: Pustaka Setia.

Iriawan, Nur, 2006, *Mengolah Data Statistik dengan Mudah Menggunakan Minitab 14*. Yogyakarta: Andi.

Montgomery, D.C. 2001, *Introduction To Statistical Quality Control*. 4th edition. New York: John Wiley and Sons.

Rohatgi, V.K. 1975, *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Wiley and Sons.

Soejoeti, Z. 1996, *Pengantar Pengendalian Kualitas Statistik*. Yogyakarta: Gajah Mada University Press.

Sumanto. 2005, *Workshop On Process Capability Analysis Using Minitab*. Bandung: Modul KBK Statistik Departemen Matematika-FMIPA ITB: Tidak Dipublikasikan.

Walpole, R.E dan Raymond H.M. 1995. *Ilmu Peluang dan Statistik untuk Insinyur dan Ilmuwan*. Bandung: ITB.

Analisis Korespondensi Jumlah Penderita Penyakit Menular di Kota Bengkulu

Priska Julianti[1], Sigit Nugroho[2] dan Jose Rizal[2]
[1] Mahasiswa Jurusan Matematika FMIPA Universitas Bengkulu
[2] Staf Pengajar Jurusan Matematika FMIPA Universitas Bengkulu

ABSTRAK

Analisis korespondensi digunakan untuk menganalisis baris dan kolom secara bersamaan dari suatu tabel kontingensi dua arah dalam ruang vektor berdimensi dua. Keunggulan analisis ini adalah menampilkan hasil analisisnya dalam bentuk grafik yang mudah untuk dipahami dan diinterpretasikan. Data yang digunakan berbentuk kategori dan diskrit.

Tujuan penelitian ini adalah menerapkan teori Analisis Korespondensi guna mendapatkan gambaran beberapa penyakit menular seperti TB paru, PNEUMONIA, rabies, DBD, diare, dan malaria di Kecamatan yang ada di Kota Bengkulu, dengan pendekatan jarak Kai Kuadrat. Hasil Analisis Korespondensi ini dipetakan dalam ruang vektor berdimensi dua dengan mereduksi dimensi data berdasarkan nilai *inersia*nya. Kecamatan yang memiliki kecenderungan terhadap jenis penyakit TB Paru adalah Sungai Serut dengan jarak Kai Kuadrat 0.028, Kecamatan yang cenderung terhadap jenis penyakit PNEUMONIA adalah Gading Cempaka dengan jarak Kai Kuadrat sebasar 0.003, dan jenis penyakit DBD cenderung banyak ditemukan di Kecamatan Ratu Agung dengan jarak Kai Kuadratnya 0.042. Dalam pereduksian dimensi informasi yang hilang sebesar L =7.03%.

Kata kunci : Analisis Korespondensi, Tabel kontingensi dua arah, nilai *inersia,* penyakit menular.

PENDAHULUAN

Penyakit menular merupakan suatu wabah penyakit yang dapat menulari banyak orang pada kurun waktu yang relatif singkat atau cepat. Dengan banyaknya jenis penyakit menular saat ini, tentunya banyak faktor yang mempengaruhi penyebarannya. Sehingga diduga kondisi suatu daerah atau kawasan pemukiman mempengaruhi tingkat penyebaran penyakit menular.

Kota Bengkulu memiliki 8 kecamatan yaitu kecamatan Muara Bangkahulu, Gading Cempaka, Teluk Segara, Selebar, Kampung Melayu, Ratu Agung, Ratu Samban dan Sungai Serut. Informasi mengenai penyakit menular yaitu DBD, diare,malaria, PNEUMONIA, rabies, TB paru dan jumlah penderita penyakit menular untuk masing-masing kecamatan telah tersedia di Dinas Kesehatan Kota Bengkulu. Tulisan ini berisi penelitian mengenai analisis statistik dalam melihat hubungan antara lingkungan dan jenis penyakit. Khususnya akan dilihat kecamatan mana yang paling dominan terhadap penyakit menular dengan menggunakan Analisis Korespondensi.

TINJAUAN PUSTAKA

Menurut Greenacre *dalam* Sudarsono dan Latra (2005) *Analyses des Correspondances* atau Analisis Korespondensi adalah teknik analisis data yang memperagakan baris dan kolom secara bersamaan dari suatu tabel kontingensi dua arah dalam ruang vektor berdimensi dua. Dalam Analisis Korespondensi ada beberapa asumsi yang harus dipenuhi, yaitu

1. Ukuran jarak Kai Kuadrat antar titik-titik (nilai kategori) analogi dengan konsep korelasi antar variabel.
2. Variabel kolom yang tepat di variabel kategori baris diasumsikan homogen.
3. Analisis Korespondensi hanya digunakan untuk dua atau tiga variabel.
4. Analisis Korespondensi adalah sebuah teknik nonparametrik yang tidak memerlukan pengujian asumsi seperti kenormalan, autokorelasi, multikolinearitas, heteroskedastisitas, linieritas sebelum melakukan analisis selanjutnya.
5. Dimensi yang terbentuk dalam Analisis Korespondensi disebabkan dari kontribusi titik-titik dari dimensi yang terbentuk dan penamaan dari dimensinya subjektif dari kebijakan, pendapat dan *error*.
6. Dalam Analisis Korespondensi variabel yang digunakan yaitu variabel diskrit yang mempunyai banyak kategori.

Tabel Kontingensi Dua Arah

Jika X dan Y adalah dua peubah yang masing-masing mempunyai sebanyak a dan b kategori, maka dapat dibentuk suatu matriks data pengamatan **P** dengan ukuran $a \times b$, Dengan $p_{ij} \geq 0$ menyatakan frekuensi dari sel ke (i, j).

Tabel Kontingensi Dua Arah

	Y_1	…	Y_j	…	Y_b	Total
X_1	p_{11}	…	p_{1j}	…	p_{1b}	$p_{1.}$
⋮		⋮		⋮		
X_i	p_{i1}	…	p_{ij}	…	p_{ib}	$p_{i.}$
⋮		⋮		⋮		
X_a	p_{a1}	…	p_{aj}	…	p_{ab}	$p_{a.}$
Total	$p_{.1}$	…	$p_{.j}$	…	$p_{.b}$	$p_{..}$

keterangan :

$p_{i.} = \sum_{j=1}^{b} p_{ij}$ $\quad i = 1, 2, \ldots, a$ peluang marjinal X

$p_{.j} = \sum_{i=1}^{a} p_{ij}$ $\quad j = 1, 2, \ldots, b$ peluang marjinal Y

$p_{..} = \sum_{i} \sum_{j} p_{ij}$ Jumlah total frekuensi dari matriks **P**

p_{ij} adalah frekuensi pengamatan ke i baris pada j kolom

Profil Baris dan Profil Kolom

Profil adalah proporsi dari setiap baris atau kolom Matriks Korespondensi yaitu setiap frekuensi pengamatan baris ke-i dan kolom ke-j dibagi dengan jumlah setiap total baris dan kolomnya masing-masing. $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{K}$ disebut profil baris (*row profile*). $\mathbf{C} = \mathbf{D_c}^{-1}\mathbf{K}'$ disebut sebagai profil kolom (*column profile*). **K** disebut matriks korespondensi dengan $k_{ij} = \frac{p_{ij}}{p_{..}}$ merupakan elemen-elemen setiap matriks korespondensi.

Untuk menampilkan profil-profil baris dan profil-profil kolom tersebut kedalam ruang dimensi *Euclid* yang berdimensi dua digunakan pendekatan jarak Kai Kuadrat yaitu :

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(p_{ij} - m_{ij}\right)^2}{m_{ij}}$$

Penguraian Nilai Singular (*Singular Value Decomposition*)

Untuk mereduksi dimensi data berdasarkan keragaman data (nilai *eigen*/*inersia*) terbesar dengan mempertahankan informasi optimum, diperlukan penguraian nilai singular. Teorema Dekomposisian Nilai Singular, yaitu misalkan A matriks berukuran $m \times n$ maka ada matriks diagonal Σ berukuran $r \times r$ dan $r \leq \min\{m,n\}$, matriks orthogonal U berukuran $m \times m$, matriks orthogonal V berukuran $m \times n$, sehingga $A = U\Sigma V^t$ dengan Σ adalah matriks berukuran $m \times n$ yang mempunyai bentuk $\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$, $\lambda_1^2 \geq \ldots \geq \lambda_m^2$ adalah nilai inersia dari $U'U$. Berdasarkan Teorema Dekomposisian Nilai Singular tersebut, Matriks yang akan di *singular value decomposition* adalah matriks $U = \mathbf{D_r}^{-1/2}(\mathbf{K} - \mathbf{rc}')\mathbf{D_c}^{-1/2}$ yang hasilnya adalah :

A adalah matriks berukuran $(a \times m)$

B adalah matriks berukuran $(b \times m)$

dan Λ merupakan suatu matriks yang elemen-elemennya adalah nilai singular, dimana nilai singular adalah akar dari nilai *inersia*nya.

Penguraian Nilai Singular Umum

Koordinat dari baris dan kolomnya ditentukan dengan menggunakan GSVD dari matriks $(\mathbf{K}-\mathbf{rc}')$ hasilnya $\mathbf{A}\Lambda\mathbf{B}'$,dengan $\mathbf{A}$ adalah matriks berukuran $a\times \mathrm{m}$, $\mathbf{B}$ adalah matriks berukuran $b\times \mathrm{m}$, Λ adalah matriks diagonal yang mempunyai unsur-unsur diagonalnya nilai singular dari matriks $\mathbf{K}-\mathbf{rc}'$, dimana berlaku $\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A}=\mathbf{I}_m$ dan $\mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B}=\mathbf{I}_m$.

Tiap himpunan titik dapat dihubungkan dengan sumbu utama dari himpunan titik lainnya yaitu:

	Rumusan dari koordinat baris	Rumusan untuk koordinat kolom
Analisis profil baris	$F=\mathbf{D}_\mathbf{r}^{-1}\mathbf{A}\Lambda$	$G = \mathbf{D}_\mathbf{c}^{-1}\mathbf{B}$
Analisis profil kolom	$F=\mathbf{D}_\mathbf{r}^{-1}\mathbf{A}$	$G=\mathbf{D}_\mathbf{c}^{-1}\mathbf{B}\Lambda$
Analisis keduanya (baris dan kolom)	$F=\mathbf{D}_\mathbf{r}^{-1}\mathbf{A}\Lambda$	$G=\mathbf{D}_\mathbf{c}^{-1}\mathbf{B}\Lambda$

Sumber : Bee- Leng Lee

Nilai *Inersia*

Nilai *inersia* menunjukkan kontribusi dari baris ke-i pada *inersia* total. Sedangkan dimaksud *inersia* total adalah jumlah bobot kuadrat jarak titik-titik ke pusat, massa dan jarak. Jumlah bobot kuadrat koordinat titik-titik dalam sumbu utama ke-d pada tiap-tiap himpunan yaitu λ_d^2 yang dinotasikan dengan λ_d. Nilai ini disebut sebagai *inersia* utama ke-d. Persamaan *inersia* utama baris dan kolom serta pusatnya dapat dinyatakan sebagai berikut:

inersia utama baris adalah $F'\mathbf{D}_\mathbf{r}F=\Lambda$

bukti $F'\mathbf{D}_\mathbf{r}F=\Lambda$, akan ditunjukkan sebagai berikut :

$$F'\mathbf{D}_\mathbf{r}F=\left(\mathbf{D}_\mathbf{r}^{-1}\mathbf{A}\Lambda\right)'\mathbf{D}_\mathbf{r}(\mathbf{D}_\mathbf{r}^{-1}\mathbf{A})$$

$$=\Lambda'\mathbf{A}'\left(\mathbf{D}_\mathbf{r}^{-1}\right)^{-1}\mathrm{I}\mathbf{A}$$

$=\Lambda'\mathbf{A}'\mathbf{D}_\mathbf{r}^{-1}\mathbf{A}$, dengan menggunakan persamaan $\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A}=\mathbf{I}_m$, didapatkan $\Lambda'\mathbf{I}_\mathbf{m}=\Lambda'$.

Karena matriks Λ' adalah simetris sehinggga $\Lambda'=\Lambda$ jadi $F'\mathbf{D}_\mathbf{r}F=\Lambda$.

inersia utama kolom adalah $G'\mathbf{D}_\mathbf{c}G=\Lambda$

bukti $G'\mathbf{D}_\mathbf{c}G=\Lambda$, akan ditunjukkan sebagai berikut :

$$G'\mathbf{D}_\mathbf{c}G=\left(\mathbf{D}_\mathbf{c}^{-1}\mathrm{B}\Lambda\right)'\mathbf{D}_\mathrm{c}\left(\mathbf{D}_\mathrm{c}^{-1}\mathbf{B}\right)$$

$$=\Lambda'\mathbf{B}'\left(\mathbf{D}_\mathbf{c}^{-1}\right)^{-1}\mathbf{I}\mathbf{B}$$

$= \mathbf{\Lambda' B' D_r^{-1} B}$, dengan menggunakan $\mathbf{B' D_c^{-1} B} = \mathbf{I}_m$, didapatkan $\mathbf{\Lambda' I_m} = \mathbf{\Lambda'}$. Karena matriks $\mathbf{\Lambda'}$ adalah simetris sehingga $\mathbf{\Lambda'} = \mathbf{\Lambda}$ jadi $G' \mathbf{D_c} G = \mathbf{\Lambda}$.

Besaran relatif untuk mengukur besarnya kehilangan informasi dapat dirumuskan sebagai berikut:

$$L = 1 - \frac{\sum_{i=1}^{d} \lambda_i^2}{\sum_{i=1}^{m} \lambda_i^2}$$

Uji Kesesuaian Kai Kuadrat (*Test of Goodness of Fit*)

Uji yang sesuai untuk mengetahui ada tidaknya hubungan antara dua variabel kategori yang berupa tabel kontingensi, adalah *Pearson Chi-Square test*.

Koefisien Kontingensi

Untuk melihat keeratan hubungan atau kecenderungan antara variabel satu dengan yang lainnya. Dengan menggunakan rumusan koefisien kontingensi sebagai berikut :

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

keterangan:

$\chi^2 =$ Statistik uji Kai Kuadrat

$N =$ banyaknya populasi sampel

Nilainya $0 \le C < 1$

METODE PENELITIAN

Jenis data yang digunakan adalah data sekunder berupa jumlah penderita peyakit menular terbayak, hasil pencatatan setiap bulan pada tahun 2006 di Kota Bengkulu.

PEMBAHASAN

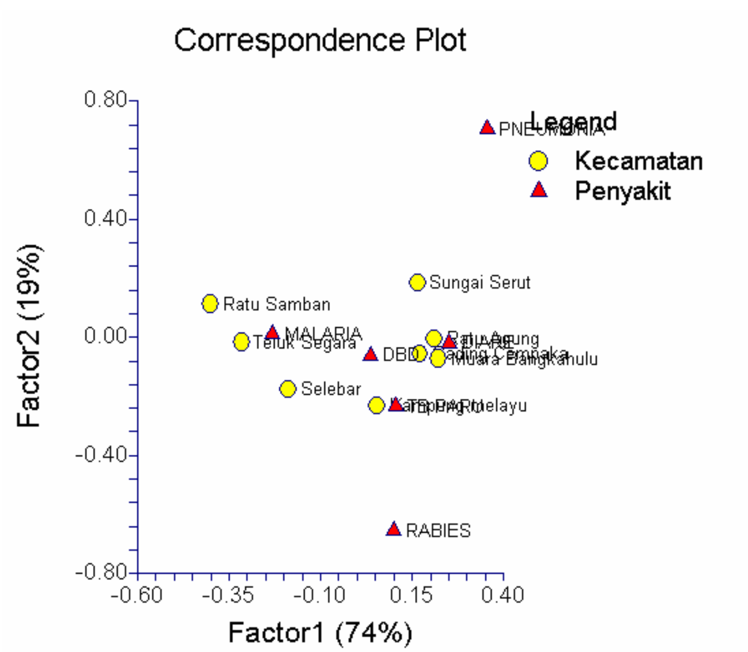
Analisis Korespondensi ini digunakan untuk memperoleh informasi mengenai posisi variabel penyakit menular dengan kecamatan. Analisis ini merupakan analisis *multivariate* yang mereduksi matriks data dengan ruang dimensi tinggi menjadi ruang dimensi dua. Untuk memperoleh informasi yang lengkap (100%) seperti informasi awal, dibutuhkan lima dimensi. Total keragaman yang dapat disumbang oleh dimensi pertama 73,63%, bila ditambah dengan dimensi kedua total keragaman yang dapat dijelaskan sebesar 92,97% dan 98,06% bila menggunakan tiga dimensi secara bersamaan, dan dimensi keempat menjelaskan keragaman sebesar 99,39%. Penambahan variasi yang dapat dijelaskan dengan ditambahkannya dimensi kedua adalah sebesar 19,33% sedangkan penambahan dimensi ketiga menambah sebesar 5,09% atau lebih kecil daripada penambahan dimensi kedua.

Oleh karena itu pemetaan yang disajikan menggunakan dua dimensi. Besarnya informasi yang hilang dalam pereduksian dimensi data dalam analisis ini sebesar L=7.03%. Nilai singular dari setiap dimensi diatas adalah akar dari setiap *inersia* atau *eigenvalue* dari setiap dimensinya dapat dilihat dari hasil nilai *singular value decomposition.*

Analisis Tabel Kontingensi

Factor	Singular value	inertia	Individual percent	Cumulative percent	Bart Chart
1	**0.24062**	**0.057914**	**73.63**	**73.63**	**|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIIII**
2	**0.1233**	**0.015207**	**19.33**	**92.97**	**|IIIIIIIIIIIII**
3	**0.0632**	**0.004003**	**5.09**	**98.06**	**|III**
4	**0.0316**	**0.001048**	**1.33**	**99.39**	**|I**
5	**0.0224**	**0.000482**	**0.61**	**100.00**	|
Total		**0.078654**			

Sumber: Analisis dengan NCSS 2000



Plot di atas dapat dilihat bahwa penyakit menular malaria cenderung berada di Kecamatan Teluk Segara dan mempunyai kedekatan pada Kecamatan Ratu Samban. Sedangkan penyakit menular TB paru cenderung pada Kecamatan Kampung melayu yang mempunyai kedakatan atau kemiripan dengan Kecamatan Selebar. Dan diare cenderung berada di Kecamatan Muara Bangkahulu yang mempunyai kedekatan dan kemiripan

dengan Kecamatan Ratu Agung dan Gading Cempaka. Dan penyakit menular DBD dan diare mempunyai kedekatan ataupun kemiripan. Jika semakin kecil jarak Kai_Kuadrat antar dua variabel Kecamatan dan kategori penyakit menular itu , maka kedua variabel itu memiliki suatu kemiripan dan kedekatan.

PENUTUP

Kesimpulan

Dari hasil penelitian yang dilakukan dengan judul Analisis Korespondensi jumlah Penderita Penyakit menular di Kota Bengkulu, ada beberapa kesimpulan sebagai berikut :

1. Berdasarkan grafik data dari jumlah penderita penyakit menular di Kota Bengkulu pada tahun 2006, tercatat bahwa ada sebanyak 16283 penderita penyakit menular yaitu DBD, diare, malaria, PNEUMONIA, rabies, dan TB Paru
2. Kecamatan yang mempunyai kemiripan dan kedekatan jenis penyakit menular DBD tersebar banyak di kecamatan Ratu Agung dengan jarak Kai Kuadrat 0.042, Ratu Samban jarak kai Kuadratnya 0.049, Teluk Segara jarak Kai Kuadratnya adalah 0.382 dan Muara Bangkahulu dengan jarak Kai Kuadrat 0.069. Sehingga solusi yang tepat untuk masing-masing Kecamatan itu yaitu kebijakan yang tepat dalam pemberatasan mengenai penyakit menular DBD di Kecamatan tersebut.
3. Kecamatan yang cenderung terhadap jenis penyakit TB Paru adalah Sungai Serut dengan jarak Kai Kuadratnya 0.028. Sehingga solusi yang tepat untuk Kecamatan Sungai Serut yaitu kebijakan yang mengarah pada penanggulangan mengenai penyakit menular TB paru.
4. Kecamatan yang cenderung terhadap jenis penyakit PNEUMONIA adalah Gading Cempaka dengan jarak Kai Kuadratnya 0.003. Sehingga solusinya yaitu suatu kebijakan dalam penanggulangan penyakit menular PNEUMONIA.

SARAN

Beberapa saran yang dapat diberikan oleh penulis sebagai bahan penelitian lanjutan adalah sebagai berikut :

1. Variabel yang digunakan dalam penelitian ini hanya dua variabel diskrit yang berkategori nominal serta data tercatat untuk tiap bulannya dalam satu tahun, oleh karena itu bagi peneliti yang ingin menggunakan Analisis Korespondensi sebaiknya dicoba untuk variabel yang lebih dari dua variabel dan berkategori banyak.
2. Dalam penulisan ini pembahasan mengenai data hilang belum banyak dibicarakan oleh peneliti, karenanya ini dapat menjadi salah satu bahan pertimbangan bagi pengembangan Analisis Korespondensi.

Kajian Hubungan Koefisien Korelasi Pearson (ρ), Spearman-*Rho* (r), Kendall-*Tau* (τ), Gamma (G) , dan Somers (d_{yx})

Resi Vusvitasari[1], Sigit Nugroho[2], dan Syahrul Akbar[2]
[1]Alumni Jurusan Matematika Fakultas MIPA Universitas Bengkulu
[2]Staf Pengajar Jurusan Matematika Fakultas MIPA Universitas Bengkulu

Abstrak

Penelitian ini bertujuan mengkaji tentang hubungan koefisien korelasi Pearson (r), Spearman-*rho* (ρ), Kendall-*tau* (τ), Gamma (G) dan Somers (d_{yx}) serta mempelajari penggunaan dari masing-masing koefisien korelasi untuk skala ordinal. Metode yang digunakan dalam penulisan ini adalah studi literatur dan data yang digunakan adalah data simulasi yang dibuat menggunakan program komputer Microsoft EXCEL. Data simulasi terdiri dari dua, yaitu data tidak normal (seragam) dan data normal yang dibangkitkan dari data seragam. Hasil penelitian menunjukkan bahwa untuk data seragam, koefisien korelasi yang diberikan oleh koefisien korelasi Spearman-*rho* (ρ) dan Kendall-*tau* (τ) lebih besar dibandingkan dengan Koefisien korelasi Pearson (r). Dan untuk data normal, koefisien korelasi yang diberikan oleh koefisien korelasi Pearson (r) lebih besar dibandingkan koefisien korelasi Spearman-*rho* (ρ) dan Kendall-*tau* (τ). Ini membuktikan bahwa koefisien korelasi Pearson sesuai digunakan untuk data yang berdistribusi normal, sedangkan koefisien korelasi Spearman-*rho* (ρ) dan Kendall-*tau* (τ) digunakan untuk data yang tidak normal. Penelitian juga menunjukan adanya hubungan yang linier antara koefisien korelasi Gamma dan Somers.

Kata Kunci : Korelasi, Ukuran Asosiasi , Tabel Kontingensi, Statistika Nonparametrik.

Pendahuluan

Kekuatan dan sifat ketergantungan antar variabel merupakan masalah sentral yang ingin diketahui pada suatu penelitian. Kadang peneliti ingin mengetahui apakah terdapat hubungan antara dua variabel dan seberapa kuat hubungan kedua variabel tersebut. Uji statistika yang mengukur keeratan hubungan antara dua variabel ini disebut analisis korelasi (*correlation*). Ukuran untuk menentukan kuatnya atau derajat keeratan hubungan antar dua variabel dinamakan koefisien korelasi (*the correlation coefficient*).

Dalam statistika parametrik, koefisien korelasi antara dua variabel (*bivariat*e) yang biasa digunakan adalah koefisien korelasi momen hasil kali Pearson, yang dinotasikan dengan r. Dimana skala data pengamatan serendah-rendahnya adalah interval atau rasio. Jika data pengamatan adalah berupa skala ordinal, dalam hal ini untuk uji korelasi statistika nonparametrik, maka ada beberapa koefisien korelasi yang dapat digunakan, yaitu koefisien korelasi peringkat Spearman-*rho* (ρ), Kendall-*tau* (τ), Gamma (G), dan Somers (d_{yx}). Dari keempat koefisien korelasi ini banyak peneliti yang mungkin ingin tahu kapan harus menggunakan koefisien korelasi peringkat Spearman-*rho* (ρ), Kendall-*tau* (τ), Gamma (G), dan Somers (d_{yx}) dalam suatu penelitian. Untuk mengetahui penggunaan dari masing-

masing koefisien korelasi untuk data berskala ordinal di atas, maka perlu dipelajari tentang sifat-sifat dari keempat koefisien korelasi ini. Sehingga nantinya para peneliti dapat mengetahui dan benar-benar tepat dalam memilih koefisien korelasi mana yang akan digunakan dalam penelitian.

Tujuan dari penulisan skripsi ini adalah untuk mengkaji tentang hubungan koefisien korelasi momen hasil kali Pearson (r), Spearman-*rho* (ρ), Kendall-*tau* (τ), Gamma (G) dan Somers (d_{yx}) serta mempelajari penggunaan dari koefisien korelasi peringkat Spearman-*rho* (ρ), Kendall-*tau* (τ), koefisien korelasi Gamma (G), dan Somers (d_{yx}) dalam suatu penelitian. Kemudian akan diberikan contoh kasus yang terkait dengan pembahasan.

Pengantar Teori Statistika Nonparametrik

Peneliti sering kali mengalami kesulitan untuk memperoleh data kontinu pada penelitian yang mengikuti distribusi normal. Hal ini salah satunya karena jumlah data sampel yang didapat tidak cukup banyak sehingga tidak memenuhi distribusi normal. Selain itu, banyak pengukuran data dilakukan secara kualitatif dan data dalam penelitian yang diperoleh sering berupa kategori yang hanya dapat dihitung frekuensinya atau berupa data yang hanya dapat dibedakan berdasarkan tingkatan atau rankingnya.

Menghadapi kasus data kategorikal (nominal) atau data ordinal seperti itu, jelas peneliti tidak mungkin mempergunakan metode statistika parametrik. Karena apabila asumsi-asumsi tidak dapat terpenuhi, akan menghasilkan suatu kesimpulan yang tidak valid. Kesulitan-kesulitan dalam data tetap harus diatasi supaya analisis data bisa dilakukan dan menghasilkan suatu kesimpulan yang valid. Sebagai gantinya diciptakan oleh pakar metode statistika alternatif yang sesuai yaitu metode statistika nonparametrik sebagai pelengkap statistika parametrik.

Metode statistika nonparametrik merupakan suatu metode analisis data tanpa memperhatikan bentuk distribusinya sehingga statistika ini sering juga disebut metode bebas sebaran (*distribution free methods*), karena model uji statistikanya tidak menetapkan syarat-syarat tertentu tentang bentuk distribusi parameter populasinya. Artinya bahwa metode statistika nonparametrik ini tidak menetapkan syarat bahwa observasi-observasinya harus ditarik dari populasi yang berdistribusi normal dan tidak menetapkan syarat homoskedastisitas (*homoscedasticity*).

Selain tidak menetapkan syarat mengenai distribusi populasinya, statistika nonparametrik juga tidak menetapkan syarat-syarat mengenai parameter-parameter populasi yang merupakan induk sampel penelitiannya.

Suatu metode statistika dapat dikatakan nonparametrik apabila memenuhi paling sedikit satu kriteria dibawah ini :

1. Metode ini digunakan untuk data pengamatan dengan skala nominal
2. Metode ini digunakan untuk data pengamatan dengan skala ordinal
3. Metode ini digunakan untuk data pengamatan dengan skala interval atau rasio, dimana distribusi populasinya tidak diketahui.

Pemilihan macam uji statistika nonparametrik mana yang paling sesuai didasarkan pada beberapa kriteria. Pertama didasarkan pada skala pengukuran variabel penelitiannya, baik itu skala nominal, ordinal, atau skala interval/rasio. Pada dasarnya uji yang sesuai bagi variabel dengan skala nominal atau ordinal adalah uji statistika

nonparametrik, namun terdapat juga uji statistika nonparametrik yang berlaku pada variabel yang berskala interval. Kedua, pemilihan uji statistika nonparametrik didasarkan pada banyaknya sampel penelitian, apakah berupa sampel tunggal, dua sampel berpasangan atau dua sampel independen, atau sampelnya lebih dari dua buah yang berpasangan atau yang independen. Ketiga, pemilihan uji nonparametrik didasarkan pada jenis penelitiannya, apakah berupa uji kesesuaian (*goodness-of-fit*), uji banding, uji independensi, atau apakah berupa uji keterikatan (korelasi) antara dua kumpulan atribut atau dua kumpulan skor. Contoh uji statistika nonparametrik diantaranya adalah uji binomial, uji median, uji tanda, uji *Kolmogorov-Smirnov*, uji korelasi peringkat, uji *Wilcoxon-Mann-Whitney*, uji *Friedman* dan lain sebagainya.

Analisis Korelasi Statistika Nonparametrik

Analisis korelasi merupakan uji statistika yang mengukur keeratan hubungan antara dua variabel. Keeratan hubungan antara dua variabel dapat diukur kekuatannya. Indeks yang mengukur keeratan hubungan dua variabel disebut koefisien korelasi. Nilai koefisien korelasi paling (r) dapat dinyatakan sebagai berikut :

$$-1 \leq r \leq 1 \qquad [1]$$

r = 1, hubungan X dan Y sempurna dan positif (mendekati 1, yaitu hubungan sangat kuat dan positif).

r = -1, hubungan X dan Y sempurna dan negatif (mendekati -1, yaitu hubungan sangat kuat dan negatif).

r = 0, hubungan X dan Y lemah sekali atau tidak ada hubungan.

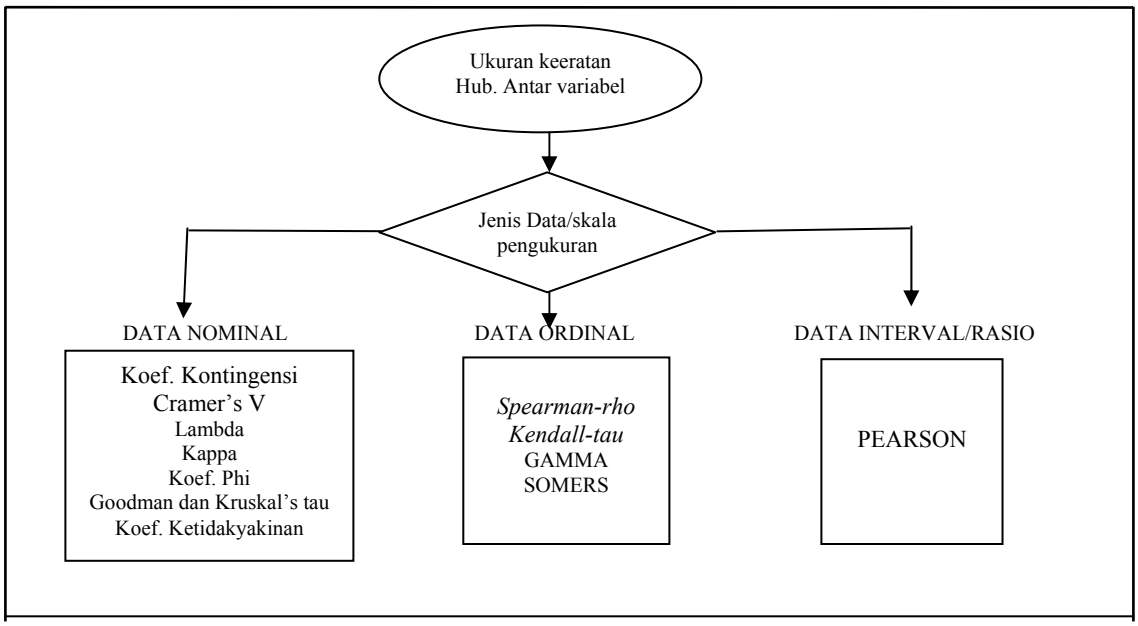
Sama halnya dengan statistika parametrik, analisis korelasi pada statistika nonparametrik juga mempelajari apakah ada hubungan antar dua variabel. Hanya pada korelasi nonparametrik, data atau variabel yang akan diuji dan diukur korelasinya adalah data nominal atau ordinal. Sebagai contoh, apakah motivasi seseorang mempengaruhi kepuasan bekerja orang tersebut. Di sini variabel motivasi ataupun kepuasan kerja adalah data ordinal, karena tidak mungkin motivasi dan kepuasan diukur seperti pengukuran tinggi badan atau berat badan yang secara riil dapat dilihat.

Dalam statistika parametrik, Koefisien korelasi yang dikenal luas dan paling sering digunakan adalah koefisien korelasi momen hasil kali Pearson yang dinotasikan dengan r, dimana rumus r adalah sebagai berikut:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\left[\left(\sum (X - \bar{X})^2\right)\left(\sum (Y - \bar{Y})^2\right)\right]^{\frac{1}{2}}} \qquad [2]$$

dengan X dan Y adalah variabel-variabel yang diamati dan banyaknya sampel pengamatan. Perhitungan dalam teknik korelasi ini mensyaratkan bahwa populasi asal sampel mempunyai dua varian (bivariat) dan berdistribusi normal. Selain itu teknik korelasi ini dalam aplikasinya digunakan untuk mengukur korelasi data dengan skala pengukuran interval atau rasio. Sedangkan dalam statistika nonparametrik, untuk kasus pengukuran

analisis korelasi antara dua kumpulan skor dapat digolongkan berdasarkan pada skala pengukuran variabel penelitiannya (skala nominal, ordinal, atau interval-rasio). Adapun penggolongan uji korelasi itu adalah sebagai berikut:



Gambar 1. *Pembagian Korelasi*

Dari bagan di atas, dapat dilihat bahwa koefisien korelasi nonparametrik untuk jenis data dengan skala pengukuran ordinal, terdapat empat macam koefisien korelasi yang dapat digunakan, yaitu koefisien korelasi peringkat *Spearman-rho* yang dinotasikan dengan ρ, koefisien korelasi *Kendall-tau* yang dinotasikan dengan τ, koefisien korelasi Somers yang dinotasikan dengan d_{yx}, dan koefisien korelasi Gamma yang dinotasikan dengan G. Keempat koefisien korelasi ini didasarkan pada ranking. Hanya saja antara koefisien korelasi peringkat *Spearman-rho* dengan koefisien korelasi *Kendall-tau*, Gamma, dan Somers ada sedikit perbedaan. Dimana untuk koefisien korelasi peringkat *Spearman-rho* memperhitungkan besarnya perbedaan rank pasangan nilai pengamatan (X_i, Y_i), sedangkan untuk koefisien korelasi *Kendall-tau*, koefisien korelasi Somers, dan koefisien korelasi Gamma hanya memperhitungkan kekuatan asosiasi berdasarkan arah pasangan nilai pengamatan (X_i, Y_i) dan tidak memperhitungkan besarnya perbedaan pasangan nilai pengamatan (X_i, Y_i) seperti pada koefisien korelasi peringkat *Spearman-rho* atau dengan kata lain koefisien korelasi *Kendall-tau*, koefisien korelasi Somers, dan koefisien korelasi Gamma didasarkan pada konsep pasangan *concordant* (C) dan *discordant* (D).

Untuk mengetahui keeratan hubungan antara dua variabel, tidak hanya dilihat dari besarnya nilai koefisien korelasi antara dua variabel yang diberikan. Akan tetapi perlu juga dilakukan uji signifikansi, dalam hal ini pengujian hipotesis dari kedua variabel tersebut.

Hipotesis-hipotesis

Uji dua arah

H_0 : X_i dan Y_i saling bebas
H_1 : X_i dan Y_i dependen (hubungan positif atau negatif) .
Uji satu arah (Positif)
H_0 : X_i dan Y_i saling bebas
H_1 : X_i dan Y_i dependen (hubungan positif) .
Uji satu arah (Negatif)
H_0 : X_i dan Y_i saling bebas
H_1 : X_i dan Y_i dependen (hubungan negatif) .

Koefisien Korelasi Nonparametrik Untuk Skala Ordinal

Koefisien korelasi merupakan ukuran yang menyatakan keeratan hubungan antara dua variabel. Koefisien korelasi bivariat yang paling lama dan banyak digunakan adalah korelasi yang dikembangkan oleh *Karl Pearson*. Perhitungan dalam korelasi ini didasarkan pada data sebenarnya (variabel asli). Secara statistik, koefisien korelasi momen hasil kali Pearson atau sering disingkat dengan koefisien korelasi Pearson yang dinotasikan dengan *r* dirumuskan sebagai berikut:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)\right]^{\frac{1}{2}}} \qquad [3]$$

Dalam aplikasinya koefisien korelasi ini digunakan untuk mengukur keeratan hubungan di antara hasil-hasil pengamatan dari populasi yang mempunyai dua varian (bivariat). Perhitungan dalam tekhnik korelasi ini mensyaratkan bahwa populasi asal sampel mempunyai dua varian dan berdistribusi normal. Selain itu teknik korelasi ini dalam aplikasinya digunakan untuk mengukur korelasi data interval atau rasio.

1. **Koefisien korelasi peringkat Spearman - *rho* (ρ)**

Ukuran korelasi nonparametrik yang analog dengan koefisien korelasi Pearson (*r*) adalah koefisien korelasi yang dikembangkan oleh *Charles Spearman* (1908) yaitu koefisien korelasi peringkat Spearman. Statistik ini kadang disebut dengan Spearman- *rho*, dan dinotasikan dengan ρ. Jika pada koefisien korelasi Pearson (*r*) digunakan untuk mengetahui korelasi data kuantitatif (skala interval dan rasio), maka pada koefisien korelasi peringkat Spearman-*rho* digunakan untuk pengukuran korelasi pada statistik nonparametrik (skala ordinal). Ini merupakan ukuran korelasi yang menuntut kedua variabel diukur sekurang-kurangnya dalam skala ordinal sehingga obyek-obyek penelitiannya dapat diranking dalam dua rangkaian berurut.

Misal data terdiri dari sampel acak bivariat berukuran *n*, yaitu $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$. Misalkan $R(X_i)$ adalah rank dari X_i dibandingkan dengan nilai X lainnya, untuk $i = 1,2,\ldots,n$. $R(X_i) = 1$ jika X_i adalah nilai X terkecil dari $X_1, X_2, \ldots, X_n$, $R(X_i) = 2$ jika X_i adalah nilai X terkecil kedua, dan seterusnya dengan rank *n* ditandai sebagai nilai X_i terbesar. Begitu juga untuk $R(Y_i)$. Jika di antara nilai X_i atau di antara nilai Y_i terdapat angka sama, maka masing-masing nilai yang sama diberi peringkat rata-rata dari posisi-posisi yang seharusnya.

Rumus koefisien korelasi peringkat Spearman- *rho* merupakan turunan rumus koefisien korelasi Pearson, yaitu

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2\right]^{\frac{1}{2}}} \qquad [4]$$

dimana untuk koefisien korelasi peringkat Spearman-*rho* (ρ), variabel asli diganti dengan rank-ranknya, maka X_i diganti dengan $R(X_i)$ dan Y_i diganti dengan $R(Y_i)$. Sehingga rumus koefisien korelasi peringkat Spearman-*rho* (ρ) adalah

$$\begin{aligned}\rho &= \frac{\sum_{i=1}^{n}\left[R(X_i) - \overline{R(X)}\right]\left[R(Y_i) - \overline{R(Y)}\right]}{\left[\sum_{i=1}^{n}\left(R(X_i) - \overline{R(X)}\right)^2 \sum_{i=1}^{n}\left(R(Y_i) - \overline{R(Y)}\right)^2\right]^{\frac{1}{2}}} \qquad [5]\\ &= \frac{\sum_{i=1}^{n}\left[R(X_i) - \frac{n+1}{2}\right]\left[R(Y_i) - \frac{n+1}{2}\right]}{\left(\frac{n(n^2-1)}{12} \cdot \frac{n(n^2-1)}{12}\right)^{\frac{1}{2}}}\\ &= \frac{\sum_{i=1}^{n}\left[R(X_i) - \frac{n+1}{2}\right]\left[R(Y_i) - \frac{n+1}{2}\right]}{\frac{n(n^2-1)}{12}}\end{aligned}$$

untuk mempermudah perhitungan, maka persamaan diatas dapat disederhanakan sebagai berikut

$$\rho = 1 - \frac{6\sum_{i=1}^{n}\left[R(X_i) - R(Y_i)\right]^2}{n(n^2-1)} = 1 - \frac{6T}{n(n^2-1)} \qquad [6]$$

dimana $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left[R(X_i) - R(Y_i)\right]^2$, yaitu jumlah kuadrat dari selisih-selisih antara rank-rank X_i *dan* Y_i untuk masing-masing pengamatan.

Langkah-langkah untuk menghitung koefisien korelasi Spearman-*rho* (ρ) adalah sebagai berikut :

- Berilah peringkat untuk masing-masing pengamatan *X* mulai dari 1 hingga *n*, juga untuk pengamatan *Y* beri peringkat mulai dari 1 sampai *n*.
- Tentukan harga $\sum_{i=1}^{n} d_i^2$, yaitu jumlah kuadrat dari selisih-selisih antara rank-rank X_i *dan* Y_i untuk masing-masing pengamatan.
- Gunakan persamaan [6] untuk menghitung ρ.

2. Koefisien korelasi Kendall - *Tau* (τ)

Koefisien korelasi yang kedua yang biasa digunakan untuk mengukur kekuatan korelasi untuk data penelitian dengan skala pengukuran ordinal adalah koefisien korelasi yang dikenalkan oleh *M.G. Kendall* (1938) yaitu koefisien korelasi Kendall-*tau* yang dinotasikan dengan τ. Koefisien korelasi ini memiliki sifat yang sama dengan koefisien korelasi peringkat Spearman-*rho*, tetapi berbeda dasar logikanya. Jika untuk koefisien korelasi peringkat Sperman-*rho* didasarkan pada peringkat (*rank*), dimana baik variabel *X* dan variabel *Y* masing-masing kita ranking. Sedangkan untuk koefisien korelasi Kendall-

tau salah satu variabelnya yang diberi peringkat (diurutkan), yaitu variabel *X* saja atau variabel *Y* saja dalam hal ini biasanya adalah variabel *X*. Sedangkan variabel *Y* akan dilihat apakah nilai variabel *Y* itu searah (konkordan) atau berlawanan arah (diskordan) dengan variabel X yang sudah diurutkan.

Jika ada data bivariat $(X_i, Y_i), i = 1,2,...,n$ dimana *X* dan *Y* sekurang-kurangnya berskala ordinal. Maka untuk setiap pasangan nilai observasi $(X_i, Y_i)\, dan\, (X_j, Y_j)\, untuk\, i \neq j$ dapat didefinisikan pasangan nilai sebagai berikut :

i. Pasangan $(X_i, Y_i)\, dan\, (X_j, Y_j)$ konkordan, jika $(X_i - X_j)(Y_i - Y_j) > 0$
   artinya adalah jika $X_i > X_j\ maka\ Y_i > Y_j$ atau jika $X_i < X_j\ maka\ Y_i < Y_j$ sehingga $(X - X)\ dan\ (Y - Y)$ memiliki tanda yang sama, yaitu sama-sama positif atau sama-sama negatif dengan hasil kali yang selalu positif.

ii. Pasangan $(X_i, Y_i)\, dan\, (X_j, Y_j)$ diskordan, jika $(X_i - X_j)(Y_i - Y_j) < 0$
   artinya adalah jika $X_i > X_j\ maka\ Y_i < Y_j$ atau jika $X_i < X_j\ maka\ Y_i > Y_j$ sehingga $(X - X)\ dan\ (Y - Y)$ memiliki tanda yang berlawanan dengan hasil kali yang selalu negatif.

Secara keseluruhan, untuk *n* pengamatan ada sebanyak $\binom{n}{2} = \frac{n\,(n\,-\,1)}{2}$ pasangan yang mungkin. Jika ada sebanyak *C* pasangan yang searah (konkordan) dan *D* pasangan yang berlawanan arah (diskordan), maka Kendall-*tau* dapat dihitung sebagai berikut:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} \qquad [7]$$

Langkah-langkah untuk menghitung koefisien korelasi Kendall-*tau* (τ) adalah sebagai berikut :

- Susunlah pasangan-pasangan (X_i,Y_i) dalam sebuah kolom menurut besarnya nilai-nilai pengamatan *X*, dari nilai pengamatan *X* yang paling kecil. Disini dapat dikatakan bahwa nilai-nilai *X* berada dalam urutan yang wajar (*natural order*).
- Perbandingkan setiap nilai pengamatan *Y* satu demi satu dengan setiap nilai *Y* yang ada di sebelah bawahnya. Jika nilai *Y* yang di bawah lebih besar dari *Y* yang di atasnya, maka arah nilai pengamatannya sama (konkordan). Dan jika nilai *Y* yang di bawah lebih kecil dari *Y* yang di atasnya, maka arah nilai pengamatannya berlawanan (diskordan).
- Tetapkan *C* sebagai banyaknya pasangan konkordan dan *D* banyaknya pasangan diskordan.
- gunakan persamaan [7] untuk menghitung τ.

Koefisien korelasi Gamma (*G*)

Sebelumnya sudah dibahas dua koefisien korelasi untuk dua variabel dengan skala pengukuran ordinal, yaitu Spearman-*rho* dan Kendall-*tau*. Akan tetapi, jika data pasangan pengamatan banyak mengandung angka sama atau ada situasi dimana data pengamatan

ditampilkan dalam bentuk tabel kontingensi, maka penggunaan koefisien korelasi Spearman-*rho* dan Kendall-*tau* akan kurang efektif. Dengan demikian untuk data pasangan pengamatan yang keduanya bertipe ordinal dan ditampilkan dalam bentuk tabel kontingensi, koefisien korelasi yang dapat digunakan adalah koefisien korelasi Gamma (*G*) dan koefisien korelasi Somers (d_{yx}).

Koefisien korelasi yang ketiga yang dapat digunakan untuk mengukur korelasi untuk data penelitian dengan skala pengukuran ordinal adalah koefisien korelasi Gamma, yang dinotasikan dengan *G*. Koefisien korelasi ini dikenalkan oleh *Goodman* dan *Kruskal* (1954). Koefisien korelasi ini memiliki dasar logika yang sama dengan koefisien korelasi Kendall-*tau*, yaitu didasarkan pada banyaknya pasangan konkordan (*C*) dan pasangan diskordan (*D*).

Misalkan ada dua pengamatan bivariat *X* dan *Y*, dimana keduanya merupakan variabel terurut. Pengamatan X_i terdiri dari $X_1, X_2, \ldots, X_k$, $i = 1,2,\ldots,k$ dimana $X_1 < X_2 < \cdots < X_k$. Begitu juga dengan pengamatan Y_j terdiri dari $Y_1, Y_2, \ldots, Y_r$, $j = 1,2,\ldots,r$ dimana $Y_1 < Y_2 < \cdots < Y_r$.

Untuk menghitung statistik *G* dari dua pasangan pengamatan untuk data ordinal, $X_1, X_2, \ldots, X_k$ dan $Y_1, Y_2, \ldots, Y_r$ yang disusun dalam tabel kontingensi seperti dibawah ini,

Tabel 1. Tabel Kontingensi Data Kategorik Peringkat.

	X_1	X_2	...	X_k	Total
Y_1	n_{11}	n_{12}	...	n_{1k}	R_1
Y_2	n_{21}	n_{22}	...	n_{2k}	R_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
Y_r	n_{r1}	n_{r2}	...	n_{rk}	R_r
Total	C_1	C_2	...	C_k	N

maka statistik *G* didefinisikan sebagai berikut,

$$G = \frac{C - D}{C + D} \qquad [8]$$

dimana $C = \sum_{i.j} n_{ij} N_{ij}^{+}$ dengan $i = 1,2,\ldots,r-1$ dan $j = 1,2,\ldots,k-1$ [9]

$$D = \sum_{i.j} n_{ij} N_{ij}^{-} \quad \text{dengan} \quad i = 1,2,\ldots,r-1 \text{ dan } \quad j = 1,2,\ldots,k \qquad [10]$$

N_{ij}^{+} *dan* N_{ij}^{-} didefinisikan sebagai berikut:

$$N_{ij}^{+} = \sum_{p=i+1}^{j} \sum_{q=j+1}^{k} n_{pq} \text{ dan } N_{ij}^{-} = \sum_{p=i+1}^{j} \sum_{q=1}^{k-1} n_{pq}$$

Langkah-langkah menghitung koefisien korelasi Gamma (*G*) adalah sebagai berikut:

- Hitung banyaknya pasangan konkordan dan diskordan dari tabel kontingensi yang diberikan. Dimana untuk menghitung banyaknya pasangan konkordan dapat digunakan persamaan [9] dan untuk menghitung banyaknya pasangan diskordan gunakan persamaan [10].
- Setelah banyaknya pasangan konkondan (C) dan diskordan (D) sudah diketahui, substitusikan C dan D ke persamaan [8].

3. **Koefisien korelasi Somers** (d_{yx})

Koefisien korelasi yang dapat digunakan untuk mengukur kekuatan korelasi untuk data penelitian dimana kedua variabel berskala ordinal dan data ditampilkan dalam bentuk tabel kontingensi selain koefisien korelasi Gamma (G) adalah koefisien korelasi Somers, yang dinotasikan dengan (d_{yx}). Koefisien korelasi ini dikenalkan oleh *Somers* (1962). Koefisien korelasi ini juga memiliki dasar logika yang sama dengan koefisien korelasi *Kendall-tau* dan Gamma, yaitu didasarkan pada banyaknya pasangan konkordan (C) dan pasangan diskordan (D).

Untuk menghitung statistik (d_{yx}) dari dua buah pengamatan terurut X dan Y, yaitu $X_1, X_2, \ldots, X_k$ dan $Y_1, Y_2, \ldots, Y_r$. Diberikan dalam tabel kontingensi dibawah ini,

Tabel 1. Tabel Kontingensi Data Kategorik Peringkat.

	X_1	X_2	…	X_k	Total
Y_1	n_{11}	n_{12}	…	n_{1k}	R_1
Y_2	n_{21}	n_{22}	…	n_{2k}	R_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
Y_r	n_{r1}	n_{r2}	…	n_{rk}	R_r
Total	C_1	C_2	…	C_k	N

Sehingga statistik (d_{yx}) didefinisikan sebagai berikut :

$$d_{yx} = \frac{2(C - D)}{N^2 - \sum_{i=1}^{k} C_i^2} \qquad [11]$$

N adalah banyaknya pengamatan dan C_i merupakan frekuensi marginal dari nilai pengamatan X. Statistik (d_{yx}) menyatakan selisih proporsi pasangan konkordan dan diskordan diantara pasangan dengan nilai pasangan pengamatan yang berangka sama untuk variabel X.

Langkah-langkah menghitung koefisien korelasi Somers (d_{yx}) adalah sebagai berikut:

- Dengan cara yang sama seperti koefisien korelasi Gamma (G), hitung banyaknya pasangan konkordan dan diskordan dari tabel kontingensi yang diberikan menggunakan persamaan [9] dan [10].
- Selanjutnya hitung jumlah kuadrat dari banyaknya frekuensi dalam tiap baris di setiap kolomnya.

- Kemudian substitusikan ke persamaan [11].

Hubungan antara Gamma (G) dan Somers (d_{yx})

Dari informasi yang diberikan, diketahui bahwa rumus koefisien korelasi Gamma (G) adalah sebagai berikut : $G = \dfrac{C - D}{C + D}$

dan koefisien korelasi Somers (d_{yx}) adalah sebagai berikut : $d_{yx} = \dfrac{2(C - D)}{N^2 - \sum_{i=1}^{k} C_i^2}$

sehingga hubungan antara Gamma (G) dan Somers (d_{yx}) adalah

$$\frac{G}{d_{yx}} = \frac{\dfrac{C - D}{C + D}}{\dfrac{2(C - D)}{N^2 - \sum_{i=1}^{k} C_i^2}} = \frac{N^2 - \sum_{i=1}^{k} C_i^2}{2(C + D)}$$

dengan demikian

$$G = \left(\frac{N^2 - \sum_{i=1}^{k} C_i^2}{2(C + D)} \right) d_{yx} \qquad [12]$$

atau

$$d_{yx} = \left(\frac{2(C + D)}{N^2 - \sum_{i=1}^{k} C_i^2} \right) G \qquad [13]$$

untuk data yang sama, dalam perhitungan G dan (d_{yx}) nilai $\dfrac{2(C + D)}{N^2 - \sum_{i=1}^{k} C_i^2}$ konstan.

Teladan Penerapan

Sebagai ilustrasi, data yang digunakan untuk perhitungan koefisien korelasi nonparametrik untuk skala ordinal adalah data simulasi. Simulasi data ini merupakan dua buah data berpasangan (X,Y). Data simulasi terdiri dari dua jenis, yaitu data tidak normal (seragam) dan data normal. Dimana simulasi data dibuat menggunakan program komputer Microsoft EXCEL.

1. Perhitungan Koefisien Korelasi Pearson, Spearman -*rho*,dan *Kendall-tau*

Perhitungan koefisien korelasi Pearson, koefisien korelasi Spearman-*rho*, dan koefisien korelasi Kendall-*tau* (τ), data simulasi yang digunakan itu sama. Data sebanyak 100 sampel, dimana setiap sampel terdiri dari 12 pengamatan. Data simulasi di sini ada dua macam, yang pertama data simulasi dengan sebaran seragam. Kemudian data tadi dibangkitkan sehingga datanya berdistribusi normal. Salahsatu contoh sampelnya adalah sebagai berikut:

Tabel 2. Simulasi Data Seragam

	Simulation Data	
	X	Y
1	48	1
2	55	2
3	48	3
4	7	4
5	17	5
6	84	6
7	87	7
8	22	8
9	45	9
10	56	10
11	74	11
12	34	12

Tabel 3. Simulasi Data Normal

	Simulation Data	
	X	Y
1	-10	11
2	-2	-6
3	4	-17
4	3	-16
5	3	0
6	-4	7
7	-2	-2
8	-13	9
9	2	-2
10	-4	-3
11	0	8
12	-5	2

Setelah simulasi data untuk masing-masing sampel dibuat, maka langkah selanjutnya adalah menghitung masing-masing koefisien korelasi Pearson, Spearman-*rho*, dan Kendall-*tau* (τ) untuk masing-masing sampel baik itu data seragam maupun data normal. Dengan demikian dapat diperoleh secara keseluruhan nilai koefisien korelasi Pearson, Spearman-*rho*, dan Kendall-*tau* untuk 100 sampel baik untuk data seragam maupun data normal yang ditampilkan pada tabel D dan tabel E (di lampiran).

Tabel D menunjukkan nilai masing-masing koefisien korelasi Pearson, Spearman-*rho*, dan Kendall-*tau* dari 100 sampel untuk data seragam. Untuk data seragam, diharapkan bahwa nilai koefisien korelasi Spearman-*rho* dan Kendall-*tau* lebih baik (lebih besar) dibandingkan dengan koefisien korelasi Pearson. Dari Tabel D dapat dilihat bahwa dari 100 sampel kebanyakan nilai-nilai koefisien korelasi yang diberikan oleh koefisien korelasi Spearman-*rho* dan Kendall-*tau* cenderung lebih besar dibandingkan dengan koefisien korelasi Pearson. Ini sesuai dengan yang diharapkan, karena data pengamatan merupakan sampel acak dengan distribusi seragam dan koefisien korelasi yang baik digunakan untuk menghitung data semacam itu adalah koefisien korelasi Spearman-*rho* dan Kendall-*tau* dimana data pengamatannya tidak berdistribusi normal. Sedangkan koefisien korelasi Pearson digunakan untuk menghitung koefisien korelasi dimana data pengamatannya berdistribusi normal, sehingga untuk data seragam nilai koefisien korelasi Pearson yang diberikan lebih kecil.

Tabel E menunjukkan nilai masing-masing koefisien korelasi Pearson, Spearman-*rho*, dan Kendall-*tau* dari 100 sampel untuk data normal. Dan untuk data normal, diharapkan bahwa nilai koefisien korelasi Pearson lebih baik (lebih besar) dibandingkan dengan koefisien korelasi Spearman-*rho* dan Kendall-*tau*. Dari Tabel E untuk ketiga nilai koefisien korelasi, dapat dilihat bahwa dari 100 sampel kebanyakan nilai-nilai koefisien korelasi Pearson lebih besar dibandingkan dengan koefisien korelasi Spearman-*rho* dan Kendall-*tau*. Ini sesuai dengan yang diharapkan, karena koefisien korelasi Pearson memang baik digunakan untuk data pengamatan yang berdistribusi normal. Sedangkan koefisien korelasi Spearman-*rho* dan Kendall-*tau* digunakan untuk data yang tidak berdistribusi normal sehingga nilai koefsien yang diberikan lebih kecil dibandingkan dengan koefisien korelasi Pearson.

Perhitungan Koefisien Korelasi Gamma (*G*) dan Somers (d_{yx})

Untuk perhitungan koefisien korelasi Gamma (G) dan Somers (d_{yx}), data simulasi yang digunakan juga sama. Data sebanyak 100 sampel, dimana setiap sampel terdiri dari 1000 pengamatan. Data simulasi di sini juga dua macam yaitu data seragam dan data normal. Karena koefisien korelasi Gamma (G) dan Somers (d_{yx}) merupakan suatu ukuran asosiasi dimana data pengamatannya berupa data kategori peringkat dan ditampilkan dalam bentuk tabel kontingensi sehingga data dibangkitkan ke bentuk tabel kontingensi. Disini masing-masing pengamatan X dan Y dibagi menjadi beberapa kelas, dalam hal ini dibagi ke dalam beberapa kategori. Berikut contoh simulasi data untuk salah satu sampel dengan 1000 pengamatan untuk data simulasi dari data seragam dan data normal.

Tabel 4. Data Simulasi Kategorik Seragam

	X	Y			
1	0.84352	0.46344	4	2	42
2	0.45049	0.51671	2	2	22
3	0.66206	0.33482	3	1	31
4	0.65512	0.17671	3	0	30
5	0.55677	0.29441	2	1	21
6	0.03931	0.25021	0	1	1
7	0.55919	0.99774	2	4	24
8	0.54403	0.13505	2	0	20
9	0.17538	0.63256	0	3	3
10	0.20846	0.51695	1	2	12
. . .					
1000	0.68083	0.61992	3	3	33

Tabel 5. Data Simulasi Kategorik Normal

	X	Y			
1	17.45115	14.72188	1	1	11
2	31.07570	26.42992	3	2	32
3	32.10446	21.43789	3	2	32
4	35.45599	42.42763	3	4	34
5	38.26679	34.41158	3	3	33
6	29.08129	23.76656	2	2	22
7	29.15713	40.98426	2	4	24
8	42.04156	8.83054	4	0	40
9	33.67225	38.77020	3	3	33
10	27.38279	41.45445	2	4	24
. . .					
1000	22.02840	44.96036	2	4	24

Dari tabel 4.9 dan 4.10 di atas untuk tiga kolom terakhir merupakan alat bantu untuk membangkitkan data dari data pengamatan X dan Y yang diberikan ke bentuk data kategorik, dalam hal ini dibuat ke dalam bentuk tabel kontingensi. Setelah melalui beberapa proses, maka data pengamatan X dan Y di atas menghasilkan tabel kontingensi sebagai berikut:

Tabel 6.Tabel Kontingensi Data Seragam

	0	1	2	3	4	
0	51	39	41	48	38	217
1	31	28	34	35	34	162
2	38	45	52	53	41	229
3	48	41	31	40	38	198
4	43	43	37	28	43	194
	211	196	195	204	194	1000

Tabel 7.Tabel Kontingensi Data Normal

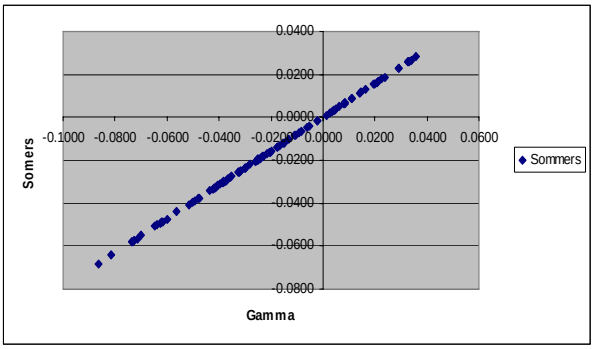
	0	1	2	3	4	
0	0	1	10	6	3	20
1	2	18	53	52	21	146
2	6	47	115	110	48	326
3	11	45	125	115	41	337
4	2	15	47	50	13	127
	21	126	350	333	126	956

Setelah tabel kontingensi dibuat, dapat dihitung nilai koefisien korelasi Gamma (G) dan Somers (d_{yx}).

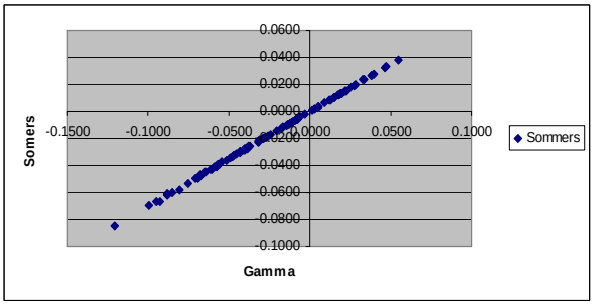
Tabel F dan tabel G (pada lampiran) menunjukan nilai-nilai koefisien korelasi Gamma (G) dan Somers (d_{yx}) untuk data seragam dan data normal. Baik untuk data normal ataupun data seragam, kebanyakan nilai koefisien korelasi Gamma (G) yang diberikan cenderung lebih besar dibandingkan dengan koefisien korelasi Somers (d_{yx}). Ini disebabkan koefisien korelasi Gamma tidak memperhatikan banyaknya data kembar, sedangkan untuk koefisien korelasi Somers (d_{yx}) banyaknya data kembar untuk pengamatan X diperhatikan.

Baik untuk data seragam maupun data normal dari tabel nilai-nilai koefisien korelasi Gamma (G) dan Somers (d_{yx}) untuk 100 sampel, nilai *slope* (kemiringan garis) yang diberikan berturut-turut sebesar 0.7882 dan 0.7021. Nilai *slope* menunjukkan dua hal, yaitu arah hubungan dan besarnya perubahan pada nilai-nilai koefisien korelasi Gamma (G) yang terjadi sehubungan dengan perubahan pada nilai-nilai koefisien korelasi Somers (d_{yx}).

Arah hubungan dapat dilihat dari tanda aljabar (positif atau negatif) pada nilai *slope*. Karena nilai *slope* yang diberikan oleh Tabel F dan tabel G bernilai positif, ini menyatakan bahwa arah hubungan antara nilai-nilai koefisien korelasi Gamma (G) dan Somers (d_{yx}) adalah positif baik untuk data seragam ataupun data normal. Dimana hubungan yang positif menunjukkan bahwa kenaikan nilai koefisien korelasi Gamma (G) diikuti oleh kenaikan pada nilai koefisien korelasi Somers (d_{yx}) dan sebaliknya penurunan nilai koefisien korelasi Gamma (G) diikuti oleh penurunan pada nilai koefisien korelasi Somers (d_{yx}). Untuk melihat pola hubungan antara nilai-nilai koefisien korelasi Gamma (G) dan Somers (d_{yx}) baik untuk data seragam maupun data normal dapat dilihat pada gambar grafik di bawah berikut ini:



Gambar 2. *Hubungan linier antara Gamma dan Somers untuk data seragam*



Gambar 3. *Hubungan linier antara Gamma dan Somers untuk data normal*

Pola hubungan antara nilai-nilai koefisien korelasi Gamma (G) dan Somers (d_{yx}) baik untuk data seragam maupun data normal yang terlihat pada gambar 2 dan gambar 3 di atas adalah hubungan yang bersifat *linier* karena dapat dihampiri oleh sebuah garis lurus. Sehingga hubungan antara nilai-nilai koefisien korelasi Gamma (G) dan nilai-nilai koefisien korelasi Somers (d_{yx}) adalah *linier*.

Kesimpulan

Koefisien korelasi yang dapat digunakan untuk skala data ordinal adalah koefisien korelasi Spearman-*rho* (ρ), Kendall-*tau* (τ), Gamma (G), dan Somers (d_{yx}). Untuk data yang tidak normal (data seragam), nilai koefisien korelasi yang diberikan oleh koefisien korelasi Spearman- *rho* dan Kendall- *tau* lebih besar dibandingkan dengan koefisien korelasi Pearson. Sedangkan untuk data normal nilai koefisien korelasi Pearson lebih besar dibandingkan dengan koefisien korelasi Spearman- *rho* dan Kendall- *tau*. Hal ini menunjukkan bahwa :

- Koefisien korelasi Pearson (r) baik digunakan jika data pengamatan berdistribusi normal dan skala data serendah-rendahnya adalah interval atau rasio.
- Koefisien korelasi Spearman-*rho* (ρ) dan Kendall-*tau* (τ) baik digunakan untuk pasangan pengamatan yang tidak berdistribusi normal.

Koefisien korelasi Gamma (G) dan Somers (d_{yx}) digunakan untuk pasangan pengamatan dengan skala data ordinal dalam bentuk kategorik peringkat (data ditampilkan dalam bentuk tabel kontingensi) dan Koefisien korelasi Gamma (G) dan Somers (d_{yx}) menunjukkan hubungan yang linier.

DAFTAR PUSTAKA

[1] Agresti, 1984. *Analysis of Ordinal Categorical Data.* John Wiley and Sons. New York.

[2] Anonim. 2003. *Introduction to Exact Nonparametric Inference*. http://www.cytel.com/Products/StatXact/Intro_Nonparametric_Inference.pdf

[3] Anonim. 2000. *Correlation*. http://www.blackwellpublishing.com/content/BPL_Images/Content_store/Sample_chapter/9781405127806/Petrie%20sample%20Ch26.pdf

[4] Aryee, M. 2002. *Measures of Association*. http://academic.shu.edu/eop/worksheets/exac2126/PRE--Measures%20of%20Association--1203.doc

[5] Azizi.2005. *Analisis Berstatistik Lanjutan*. http://www.geocities.com/kheru2006/vii.htm

[6] Conover, W.J. 1971. *Practical Nonparametric Statistics*. Wiley International Edition. John Wiley and Sons. New York, NY.

[7] Daniel, W. 1989. *Statistika Nonparametrik Terapan*. Penerbit PT. Gramedia. Jakarta.

[8] Djarwanto, 1997. *Statistika Nonparametrik*. BPFE - Yogyakarta. Yogyakarta.

[9] Elifson, K.W, and R. Runyon. 1990. *Fundamental of Social Statistics*. Second Edition. McGraw-Hill International Edition. Singapore.

[10] Gibbon, J.D. 1985. *Nonparametric Statistical Inference*. Marcel Dekker. New York, NY.

[11] Loether, H. and D.G. McTavish.1988. *Descriptive and Inferential Statistics*: *An Introduction*, *Third Edition*. Allyn and Bacon. Needham Heights, MA.

[12] Lohninger, H. 2006. *Ordinal Association*. http://www.statisticssolutions.com/ordinal-association.htm

[13] SAS Institute, 1999. *Measures of Association*. http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm

[14] Scheaffer R.L. 1999. *Categorical Data Analysis*. http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf

[15] Siegel, S., and J. Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill International Edition. Singapore.

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.

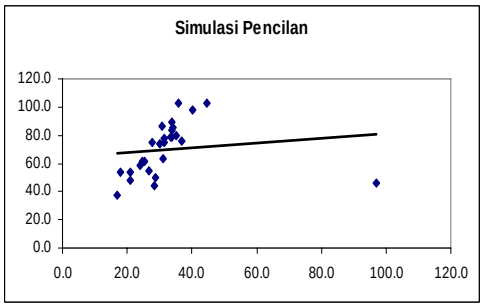
Dengan memperlihatkan pola pada tiap kasus penurunan koefisien korelasi diatas, dapat dilihat bahwa :

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.

Analisis Korelasi Kanonik pada Data Normal Multivariat yang Saling Berkorelasi dalam Pembentukan Model Regresi Linier Sederhana

Rosa Ayu Oktarina[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 berguna untuk memeriksa dan memodelkan hubungan diantara variabel-variabel secara fungsional atau berbentuk fungsi. Jika dalam suatu penelitian atau kasus menggunakan peubah yang relatif banyak, maka analisis yang digunakan adalah analisis peubah ganda. Dalam penelitian ini akan dibahas suatu pendekatan dalam mencari hubungan antara variabel bebas dengan variabel tak bebas pada data normal multivariat yang saling berkorelasi dengan menggunakan analisis korelasi kanonik (*Canonical Correlation Analysis*). Dengan melihat pasangan-pasangn kombinasi linier dari masing-masing himpunan variabel. Dari hasil penelitian dengan menggunakan dua contoh kasus yang datanya dibangkitkan terlihat bahwa besar atau kecilnya sampel sangat mempengaruhi oleh nilai korelasi kanonik dan besar atau kecilnya nilai korelasi kanonik juga dipengaruhi oleh korelasi antara variabel X_i dan Y_i.

Kata Kunci :Analisis Regresi, Analisis Korelasi kanonik.

PENDAHULUAN

Analisis regresi adalah teknik statistik yang berguna untuk memeriksa dan memodelkan hubungan diantara variabel-variabel secara fungsional atau berbentuk fungsi. Penerapan analisis regresi dapat dijumpai secara luas di banyak bidang seperti teknik, ekonomi, manajemen, ilmu-ilmu biologi, ilmu-ilmu sosial, dan ilimu-ilmu pertanian.

Analisis regresi linier dikelompokkan menjadi dua yaitu: Analisis Regresi Linier Sederhana dan Analisis Regresi Berganda. Analisis Regresi Linier Sederhana bertujuan mempelajari hubungan linier antara dua variabel, variabel ini dibedakan menjadi variabel bebas (X) dan variabel tak bebas (Y). Sedangkan Analisis Regresi Berganda adalah analisis regresi yang seringkali digunakan untuk mengatasi permasalahan analisis yang melibatkan hubungan dari dua atau lebih variabel bebas.

Jika dalam suatu penelitian atau kasus menggunakan peubah yang relatif banyak, maka analisis yang akan digunakan adalah analisis peubah ganda. Beberapa hal yang mendasari penggunaan analisis ini adalah antara peubah satu dengan peubah lain ada korelasi. Salah satu jenis analisis yang masuk katagori analisis peubah ganda adalah Analisis Korelasi Kanonik *(Canonical Correlation Analysis)*. Dalam penelitian ini akan dibahas suatu pendekatan dalam mencari hubungan antara variabel bebas dan variabel tak bebas pada data normal multivariat yang saling berkorelasi, yaitu dengan menggunakan analisis korelasi kanonik (*Canonical Correlation Analysis*).

Misalkan saja terdapat *n* pengamatan bebas pada ($p+q$) variabel yang berdistribusi normal, yaitu $Y_1, Y_2, \ldots, Y_q$ yang merupakan himpunan dari variabel-variabel tak bebas *(dependent variable)* dan $X_1, X_2, \ldots, X_p$ yang merupakan himpunan dari variabel bebas *(independent variable)*. Selanjutnya dari variabel-variabel tersebut akan dilihat pasangan-pasangan kombinasi linier dari masing-masing himpunan.

Kombinasi linier dari himpunan variabel ini disebut juga sebagai variabel kanonik, selanjutnya variabel kanonik ini akan dibentuk menjadi model regresi linier sederhana dengan menggunakan metode kuadrat terkecil.

Analisis Regresi Berganda

Hubungan fungsional antar variabel dapat ditulis sebagai
Bila variabel x lebih dari satu maka hubungan tersebut dapat ditulis sebagai

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} \quad i \neq k \text{ dengan } i, k = 1, 2, \ldots, n. \tag{1}$$

persamaan (1) merupakan penduga dari

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \tag{2}$$

dimana $\beta_0, \beta_1, \beta_2,$ *dan* β_k merupakan parameter.
Apabila dinyatakan dalam bentuk matriks persamaan (2-4) dapat dinyatakan sebagai

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \tag{3}$$

Menurut Johnson dan Wichern (1998), syarat galat diasumsikan memiliki sifat :

1. $E(\varepsilon_j) = 0$
2. $Var(\varepsilon_j) = \sigma^2$ (konstan)
3. $Cov(\varepsilon_j, \varepsilon_k) = E(\varepsilon_j, \varepsilon_k) = 0,\ j \neq k.$

Metode Kuadrat Terkecil

Untuk menaksir parameter $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$ maka akan digunakan metode kuadrat terkecil, dengan asumsi-asumsi sebagai berikut :

1. $E(\varepsilon_1) = 0$ untuk semua *i*
2. $E(\varepsilon_i \varepsilon_j) = \sigma^2 e$ untuk $i \neq j$; $i, j = 1, 2, \ldots, n$.
3. X matriks yang mempunyai rank $k+1 < n$.

Misalkan $b' = (b_0, b_1, \cdots, b_k)$ adalah vektor baris yang merupakan taksiran untuk $\beta' = (\beta_0, \beta_1, \cdots, \beta_k)$. Dari persamaan (2-2) diperoleh $\underline{e} = \underline{y} - \underline{x}\underline{\beta}$ dimana e merupakan vektor residu. Untuk memperoleh $\underline{\hat{\beta}}$ maka $\underline{e}'\underline{e}$ diturunkan terhadap $\underline{\hat{\beta}}$ dan menyamakannya dengan 0.Sehingga di peroleh

$$\underline{\hat{\beta}} = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y} \tag{4}$$

Untuk membuktikan bahwa $\underline{\hat{\beta}}$ adalah taksiran terbaik untuk $\underline{\beta}$ maka perlu dibuktikan bahwa $\underline{\hat{\beta}}$ mempunyai sifat tak bias dan variansi minimum.

Dapat dilihat bahwa taksiran kuadrat terkecil $\underline{\hat{\beta}}$ adalah tak bias, dengan variansi $Var\,(\underline{\hat{\beta}}) = \sigma^2(\underline{x}'\underline{x})^{-1}$ merupakan variansi terkecil dari semua penaksir linier takbias yang dijamin oleh Teorema Gauss-Markov berikut.

Teorema Gauss-Markov

Penaksir kuadrat terkecil $\underline{\hat{\beta}} = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}$ mempunyai variansi terkecil dalam himpunan semua penaksir linear takbias.

Analisis Korelasi Kanonik

Analisis korelasi kanonik pertama kali diperkenalkan oleh Hotelling (1936), sebagai suatu teknik statistika peubah ganda yang menyelediki keeratan hubungan antara dua gugus peubah. Analisis korelasi kanonik digunakan untuk mencari / mengidentifikasi serta mengukur hubungan antara dua himpunan variabel (Johnson dan Wichern, 1998). Analisis korelasi kanonik memfokuskan pada korelasi antara kombinasi linier variabel pada suatu himpunan dan kombinasi linier variabel dalam himpunan lainnya.

Analisis korelasi kanonik merupakan analisis regresi ganda dengan q buah variabel tak bebas dan p buah variabel bebas. yang modelnya adalah sebagai berikut

$$\underbrace{Y_1, Y_2, \ldots, Y_q}_{\text{Matriks}} \quad = \quad \underbrace{X_1, X_2, \ldots, X_p}_{\text{Matriks}}$$

Ciri data untuk korelasi kanonik adalah semua data untuk analisis korelasi kanonik bertipe matriks, yakni data interval atau data rasio. Di samping itu, analisis kanonik juga mampu menguraikan struktur hubungan di dalam gugus peubah bebas maupun di dalam gugus peubah tak bebas (Sudjana, 2002).

Dalam analisis korelasi kanonik ada beberapa asumsi yang harus dipenuhi yaitu:

1. Ada hubungan yang bersifat linier (linearitas) antar dua variabel.
2. Perlunya *Multivariate Normality* untuk menguji signifikansi setiap fungsi kanonik.
3. Tidak ada Multikolinearitas antar anggota kelompok variabel, baik variabel dependen maupun variabel independen.

Jika terjadi penyimpangan asumsi maka penanganan yang dapat dilakukan yaitu dengan melakukan transformasi data.

Proporsi Keragaman

Besarnya keragaman sampel yang diterangkan oleh peubah kanonik yang dipilih dapat dihitung dengan menggunakan pendekatan berikut :

Proporsi keragaman X yang diterangkan adalah :

$$R^2_{Z^{(1)}|U_1,U_2,\ldots,U_r} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{U_i,z_k^{(1)}}}{p} \qquad (5)$$

Sedangkan proporsi keragaman Y yang juga diterangkan adalah:

$$R^2_{Z^{(2)}|V_1,V_2,..,V_r} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{q} r^2_{V_i,z_k^{(2)}}}{q} \qquad (6)$$

Besar atau kecilnya nilai proporsi keragaman menunjukkan baik atau tidaknya jumlah kanonik yang dipilih. Semakin besar nilai proporsi keragaman ini menggambarkan semakin baik peubah kanonik yang dipilih menerangkan keragaman asal.

Pendugaan Koefisien Kanonik

Misal, ingin di buat hubungan antara gugus peubah tak bebas $Y_1, Y_2, \ldots, Y_q$ yang dinotasikan dengan vektor peubah acak Y, dengan gugus peubah bebas $X_1, X_2, \ldots, X_p$ yang dinotasikan dengan vektor peubah acak X, dimana $p \le q$.

Misalkan, karakteristik dari vektor peubah acak X dan Y adalah sebagai berikut:

$$E(X^{(1)}) = \mu^{(1)}; \qquad Cov(X^{(1)}) = \sum_{11}$$

$$E(X^{(2)}) = \mu^{(2)}; \qquad Cov(X^{(2)}) = \sum_{22}$$

$$Cov(X^{(1)}, X^{(2)}) = \sum_{12} = \sum'_{21}$$

Kombinasi linear dari kedua gugus peubah tersebut dapat ditulis sebagai berikut:

$$U = \underline{a}'X^{(1)} = a_1X_1^{(1)} + a_2X_2^{(1)} + \cdots + a_pX_p^{(1)}$$

$$V = \underline{b}'X^{(2)} = b_1X_1^{(2)} + b_2X_2^{(2)} + \cdots + b_qX_q^{(2)}$$

Sehingga,

$$Var\ (U) = \underline{a}'Cov(X^{(1)})\,a = \underline{a}'\sum_{11} a$$

$$Var\ (V) = \underline{b}'Cov(X^{(2)})\,b = \underline{b}'\sum_{22} b$$

$$Cov\ (U,V) = \underline{a}'Cov\ (X^{(1)}, X^{(2)})\,b = \underline{a}'\sum_{12} b$$

Dari sini dicari koefisien vektor a dan b sehingga,

$Corr\ (U,V) = \dfrac{\underline{a}'\sum_{12} b}{\sqrt{\underline{a}'\sum_{11} a}\,\sqrt{\underline{b}'\sum_{22} b}}$ sebesar mungkin.

Sehingga dapat didefenisikan bahwa pasangan pertama dari peubah kanonik adalah kombinasi linier U_1, V_1 yang memiliki ragam satu dan korelasi terbesar , pasangan kedua dari peubah kanonik adalah kombinasi linier U_2, V_2 yang memiliki ragam satu dan korelasi terbesar kedua serta tidak berkorelasi dengan peubah kanonik pertama, pasangan ke-k dari peubah kanonik adalah kombinasi linier U_k, V_k yang memiliki ragam satu dan korelasi terbesar ke-k serta tidak berkorelasi dengan peubah kanonik *1, 2, ..., k-1* (Johnson dan Wichern,1998).

Dengan demikian dapat dituliskan :
Peubah kanonik pertama :

$$U_1 = \underline{a}'_1 X^{(1)} \quad Var = (U_1) = 1$$

$$V_1 = \underline{b}'_1 X^{(2)} \quad Var = (V_1) = 1$$

Maksimum $Corr(U_1, V_{1)} = \rho_1^*$

Peubah kanonik kedua :

$$U_2 = \underline{a}'_2 X^{(1)} \quad Var = (U_2) = 1$$

$$V_2 = \underline{b}'_2 X^{(2)} \quad Var = (V_2) = 1$$

$$Cov(U_1, U_2) = 0$$

$$Cov(V_1, V_2) = 0$$

Maksimum $Corr(U_2, V_2) = \rho_2^*$

Peubah kanonik *ke – k* :

$$U_k = \underline{a}'_k X^{(1)} \quad Var = (U_k) = 1$$

$$V_k = \underline{b}'_k X^{(2)} \quad Var = (V_k) = 1$$

$$Cov(U_k, U_1) = 0$$

$$Cov(V_1, V_k) = 0$$

Maksimum $Corr(U_k, V_{k)} = \rho_k^*$

Teorema Ketaksamaan Cauchy-Schwarz

Misalkan b dan d adalah vektor $p \times 1$, dengan $(b'd)^2 \le (b'b)(d'd)$ jika dan hanya jika $b = c\,d$ (atau $d = c\,b$) untuk beberapa nilai c konstan.
Bukti :

Dengan menggunakan ketaksamaan Cauchy-Schwarz atau metode langrange maka diperoleh $\rho_1^{*2} \ge \rho_2^{*2} \ge \cdots \ge \rho_p^{*2}$ adalah nilai eigen dari matriks $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$ yang berpadanan dengan vektor eigen $f_1, f_2, \cdots, f_p$. Disamping itu, $\rho_1^{*2} \ge \rho_2^{*2} \ge \cdots \ge \rho_p^{*2}$ juga merupakan nilai eigen dari matriks $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$ yang berpadanan dengan vektor eigen $e_1, e_2, \cdots, e_p$.
Sehingga vektor koefisien a dan b diperoleh sebagai berikut :

$$a_1 = e_1 \Sigma_{11}^{-1/2} \qquad b_1 = f_1 \Sigma_{22}^{-1/2}$$

$$a_2 = e_2 \Sigma_{11}^{-1/2} \qquad b_2 = f_2 \Sigma_{22}^{-1/2}$$

$$\cdots\cdots \qquad\qquad \cdots\cdots$$

$$a_p = e_p \Sigma_{11}^{-1/2} \qquad b_p = f_p \Sigma_{22}^{-1/2}$$

Penerapan Analisis Korelasi Kanonik

Pada skripsi ini contoh kasus yang digunakan dalam penerapan analisis korelasi kanonik, menggunakan data yang telah dibangkitkan dari program Minitab Release 13.20 yang dibangkitkan sebanyak 5 variabel independen dan 5 variabel dependen sebanyak 25 dan 50 sampel. Masing-masing variabel independent diberikan simbol X_1 hingga X_5 dan variabel dependen diberikan simbol Y_1 hingga Y_5 yang telah dipilih secara acak . Kemudian dilakukan analisis dengan menggunakan software SAS 6.12, diperoleh hasil sebagai berikut :

Dari hasil output SAS 6.12, pada teladan 1 dan teladan 2, diperoleh nilai korelasi antar variabel-variabelnya. Variabel dependen maupun variabel independennya menghasilkan nilai korelasi yang cukup tinggi, hal ini menunjukkan terjadi multikolinieritas antar variabel-variabel tersebut, jika dilakukan analisis dengan menggunakan analisis regresi ganda, akan diperoleh model dugaan yang tidak mencerminkan keadaan yang sesungguhnya, sehingga perlu dilakukan analisis korelasi kanonik dalam masalah ini.

Dari hasil output SAS juga akan dihasilnya nilai korelasi dan proporsi keragaman dari masing-masing pasangan kombinasi liniernya. Kemudian akan dipilih pasangan kombinasi yang memiliki nilai proporsi keragaman yang paling tinggi, selanjutnya akan diperoleh data baru yang dapat mewakili data asalnya. Kemudian dari data tersebut akan diperoleh model dugaan sebagai berikut :

Dari data hasil pasangan kombinasi linier antara U_1 dan V_1 di atas maka dapat dibuat model dugaan dengan menggunakan metode kuadrat terkecil.

1. Untuk data simulasi pertama (*n = 25*)

Diperoleh model dugaan sebagai berikut

$$V_1 = -0.0123 + 0.7816U_1$$

dengan kuadrat galatnya adalah 12.87, sedangkan koefisien determinasinya $(R^2) = 0.5326$.

2. Untuk data simulasi kedua (*n = 50)*

Diperoleh model dugaan sebagai berikut

$$V_1 = -0.1506 + 0.7583U_1$$

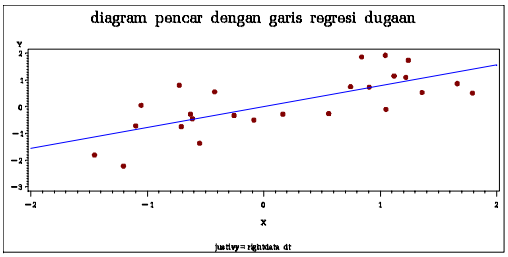
dengan jumlah kuadrat galat yang diperoleh berdasarkan output SAS 6.2 adalah 24.35, dan koefisien determinasinya $(R^2) = 0.5365$

Langkah selanjutnya adalah mengecek asumsi-asumsi yang ada pada model regresi yang diperoleh dari teladan 1 dan teladan 2. Data U_1 dan V_1 adalah merupakan data acak yang dibangkitkan dari software Minitab Release 13.20 dengan menyebar normal, dan nilai V_1 adalah bebas. Kemudian dari output SAS 6.12 (lihat lampiran 2) plot residualnya tidak membentuk model tertentu sehingga asumsi kehomogenan terpenuhi.

Untuk mengecek asumsi garis linier pada data U_1 dan V_1 dari kedua contoh teladan di atas, akan dilihat plot kenormalan dugaan garis regresinya. Apabila titik-titik dugaan yang dihasilkan telah sesuai atau mendekati garis lurus yang ditentukan berdasarkan data asal, maka titik dugaan tersebut dapat dikatakan telah mengikuti distribusi normal.

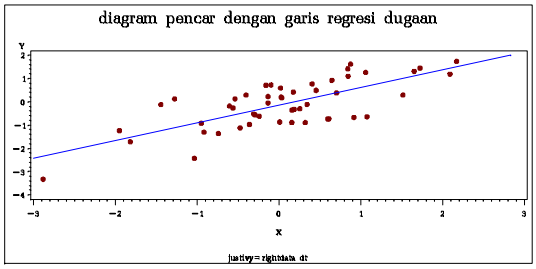
Sebaliknya, apabila titik dugaan tidak mengikuti garis lurus atau banyak titik-titik yang menyimpang, maka ada indikasi bahwa titik dugaan tidak mengikuti data normal.

Untuk mengecek asumsi garis linier pada data U_1 dan V_1 dengan jumlah $n = 25$, dari output SAS 6.12 dapat dilihat pada Gambar 2. bahwa garis dugaan merupakan garis yang lurus.



Gambar 2. Dugaan Garis Linier dengan $n = 25$

Sedangkan asumsi garis linier dengan jumlah $n = 50$, dari hasil output pada gambar 3, terlihat bahwa garis dugaannya merupakan garis yang lurus.



Gambar 3. Dugaan Garis Linier dengan $n = 50$

Dari model dugaan yang telah diperoleh tersebut untuk jumlah $n = 25$ dan jumlah $n = 50$ serta telah dilakukan uji asumsi pada model regresinya. Sehingga, dapat disimpulkan bahwa model dugaan yang dibuat telah memenuhi kaidah yang berlaku.

Kesimpulan Dan Saran

Dari hasil penelitian yang telah dilakukan maka dapat diambil kesimpulan sebagai berikut :

1. Dari data yang telah dibangkitkan dengan menggunakan program MINITAB Release 13.20 diperoleh besarnya nilai korelasi pada teladan 1 adalah

$$Corr(U_1, V_1) = 0.7334$$
$$Corr(U_2, V_2) = 0.5927$$
$$Corr(U_3, V_3) = 0.2583$$
$$Corr(U_4, V_4) = 0.1625$$
$$Corr(U_5, V_5) = 0.0480$$

Besarnya nilai korelasi pada teladan 2 adalah

$Corr\,(U_1, V_1) = 0.7885$

$Corr\,(U_2, V_2) = 0.3374$

$Corr\,(U_3, V_3) = 0.2236$

$Corr\,(U_4, V_4) = 0.1479$

$Corr\,(U_5, V_5) = 0.0370$

2. Besar atau kecilnya sampel mempengaruhi nilai korelasi kanonik. Semakin banyak data yang digunakan sebagai sampel, maka semakin rendah nilai korelasi kanonik yang diperoleh, tergantung dengan baik atau tidaknya data dipilih.
3. Besar atau kecilnya nilai analisis korelasi kanonik sangat dipengaruhi oleh korelasi antara variabel X_i dan Y_i.

Saran

Dalam penulisan skripsi ini, penulis menggunakan data yang telah dibangkitkan dengan menggunakan software MINITAB Release 13.20. Mengingat pentingnya aplikasi analisis korelasi kanonik digunakan untuk mengetahui keeratan hubungan antara variabel, maka untuk penelitian lebih lanjut sebaiknya menggunakan data riil, sehingga penerapannya lebih nyata dan terfokus pada kasus yang dianalisis.

DAFTAR PUSTAKA

Anonim. 2007. *.Analisis Korelasi Kanonik : Analisis Pengolahan Data*. http//www.deptan.go.id/editama/statistik/web_statistik.doc.

Anonim. 2007. *Correlation Analisys With Sas : The Cancorr Procedure.* http//www.okstate.edu/sas/v7/sas pdf/stat/chap 18.pdf.

Hotelling. 1936. *Korelasi Kanonik* : *Contoh Penggunaan Korelasi Kanonik.* http//www. Youngstatistician.com.

Johnson, D. E. 1998. *Applied Multivariate Methods for Analysis*. Kansas State University, USA.

Johnson, W and Wichern, D. 1998. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, New Jarsey.

Santoso, S. 2004. *Buku Latihan SPSS Statistik Multivariat*. Elex Media Komputindo, Jakarta.

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

Sudjana. 2002. *Teknik Analisis Regresi dan Korelasi Bagi Para Peneliti*. Edisi ke-3. Penerbit Tarsito, Bandung.

Supranto, J. 1989. *Statistika : Teori dan Aplikasi, jilid 2*. Edisi ke-5. Penerbit Erlangga.

Model Regresi Linier Berganda dan Cobb-Douglas dalam Analisis Fungsi Produksi
(Kasus Produksi Padi Di Desa Darat Sawah Ulu Kecamatan Seginim Kabupaten Bengkulu Selatan)

Akira Takarada[1], Buyung Keraman[2], dan Syahrul Akbar[2]
[1]Alumni Jurusan Matematika Fakultas MIPA Universitas Bengkulu
[2]Staf Pengajar Jurusan Matematika Fakultas MIPA Universitas Bengkulu

ABSTRAK

Penelitian ini bertujuan mencari model terbaik yang dapat digunakan untuk melihat pengaruh faktor-faktor produksi seperti Luas lahan, Jumlah Benih, Jumlah Jam Kerja, Jumlah Pupuk (Pupuk Urea, pupuk KCL, dan pupuk TSP), serta Jumlah Pestisida terhadap produksi padi.

Model yang digunakan adalah model regresi linier berganda dan model *Cobb-Douglas*. Kedua model tersebut dibandingkan untuk menentukan model terbaik yang dapat digunakan sebagai model pendugaan berdasarkan data produksi padi para Petani di Desa Darat Sawah Ulu Kecamatan Seginim, Kabupaten Bengkulu Selatan. Dari analisis dengan menggunakan bantuan paket program SAS 6.12, SPSS 11.0 dan Minitab 14 diperoleh kesimpulan bahwa model yang digunakan untuk menduga hasil produksi adalah:

$$Y = 266.4X_1 + 0.54X_2 + 3.94X_3 + 0.60X_4 + 81.51X_5$$

Kata kunci : Analisis Regresi Linier Berganda, Analisis Regresi *Cobb-Douglas*

PENDAHULUAN

Latar Belakang

Salah satu model matematika yang dapat mengungkapkan hubungan sepasang variabel atau lebih adalah model regresi. Selain mencerminkan perilaku hubungan antara variabel, analisis regresi juga digunakan untuk melakukan pendugaan parameter dalam model dan kepentingan peramalan. Jika model regresi dibentuk untuk meramalkan variabel tak bebas dari variabel-variabel bebas, maka model regresi yang dibentuk haruslah model regresi terbaik, yaitu model yang mencerminkan pola/perilaku hubungan yang sesungguhnya antara variabel bebas dengan variabel tak bebas. Setiap pembentukan model regresi terbaik dapat dilakukan dengan melihat berbagai kriteria, diantaranya adalah Koefisien Determinasi (R^2), Koefisien Determinasi Terkoreksi $(R^2 - adjusted)$, Uji Analisis Varians, Uji Parsial *(uji-t)*, dan Kuadrat Tengah Galat *(Mean Square Error)*.

Demikian juga dalam pembentukan model fungsi produksi dapat dilihat dengan kriteria tersebut. Dalam fungsi produksi telah banyak dibahas dan digunakan berbagai model pendugaan fungsi produksi, diantaranya: fungsi linier, fungsi kuadratik, fungsi C*obb-Douglas*, dan sebagainya (Sukartawi, 1990). Dari beberapa kajian yang telah dilakukan, tidak ada satupun rekomendasi yang menyatakan bahwa model terbaik dari fungsi produksi adalah mengikuti suatu model/fungsi tertentu, akan tetapi tergantung dari

masing-masing data yang ada. Model fungsi produksi yang biasa digunakan adalah Model *Cobb-Douglas* karena model ini memiliki keunggulan-keunggulan dalam analisis fungsi produksi.

Berdasarkan latar belakang tersebut, pada skripsi ini dibahas perbandingan model untuk analisa fungsi produksi antara Model Regresi *Cobb-Douglas* yang dibandingkan dengan Model Regresi Linier Berganda.

Tujuan

1. Membentuk model regresi yang terbaik bagi Model Regresi *Cobb-Douglas* dan Model Regresi Linier Berganda berdasarkan kriteria, yaitu pengujian Koefisien Regresi, Koefisien Determinasi, pemeriksaan asumsi sisaan, Nilai Elastisitas Produksi, dan Produk Marginal.
2. Membandingkan model regresi terbaik yang dibentuk, yaitu antara Model Regresi *Cobb-Douglas* dengan Model Regresi Linier Berganda.
3. Menentukan variabel bebas (faktor produksi) yang berpengaruh secara nyata terhadap variabel tak bebas (hasil produksi).

TINJAUAN PUSTAKA

Fungsi Produksi

Fungsi produksi merupakan suatu fungsi yang menggambarkan hubungan antara output (nilai produksi) sebagai variabel tak bebas dengan input-inputnya sebagai variabel bebas. Dalam analisis fungsi produksi, faktor produksi sering disebut dengan istilah input. Macam faktor produksi atau input ini, berikut jumlah dan kualitasnya perlu diketahui oleh seorang produsen. Oleh karena itu, untuk menghasilkan suatu produk diperlukan pengetahuan hubungan antara faktor produksi (input) dan produk (output). Dalam bentuk matematika sederhana fungsi produksi dapat dituliskan sebagai berikut:

$$Y = f(X_1, X_2, ..., X_p)$$

Dengan Y merupakan jumlah produksi yang diperoleh sebagai hasil penggunaan faktor-faktor produksi $X_j = (j = 1,2,..., p)$ secara serentak dan p adalah jumlah faktor produksi.

Dalam analisis fungsi produksi dikenal istilah Elastisitas Produksi (E_p), yaitu persentase perubahan output sebagai akibat dari persentase perubahan input. Jumlah elastisitas faktor-faktor produksi menggambarkan fase pergerakan usaha *(return to scale.*

Ada tiga fase pergerakan usaha dalam Analisis Fungsi Produksi yaitu:

1. $\sum_{i=1}^{k} E_{pi} > 1$, artinya produksi berada pada fase kenaikan hasil yang meningkat.
2. $\sum_{i=1}^{k} E_{pi} = 1$, artinya produksi berada pada fase kenaikan hasil dengan laju yang tetap.

3. $0 < \sum_{i=1}^{k} E_{pi} < 1$, artinya produksi berada pada fase kenaikan hasil dengan laju yang makin menurun (Sukartawi, 1990).

Model Regresi Linier Berganda

Persamaan regresi linier berganda yang mengandung *k* variabel bebas dan *n* pengamatan secara umum dapat dirumuskan dalam bentuk :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i \qquad i = 1,2,...,n$$

Keterangan :

Y_i	:	Nilai variabel takbebas ke-i
$X_1, X_2, ..., X_k$	:	Variabel bebas
$\beta_0, \beta_1, ..., \beta_k$	:	Koefisien regresi
ε_i	:	Galat/Sisaan ke-i

Dalam Analisis Fungsi Produksi, produk marjinal dinyatakan sebagai suatu tambahan satu satuan input X yang dapat menyebabkan penambahan atau pengurangan satu satuan output Y. Produk marjinal merupakan turunan pertama dari nilai produksi dan dinyatakan sebagai berikut :

$$PM = \frac{\partial Y}{\partial X}$$

Elastisitas Produksi pada Model Regresi Linier Berganda dapat dituliskan sebagai berikut:

$$E_{p_{X_{ki}}} = \beta_k \cdot \frac{\overline{Y}_i}{\overline{X}_{ki}}$$

Keterangan:

$E_{p_{X_{ki}}}$	:	Elastisitas produksi variabel ke-k pengamatan ke-i
β_k	:	Koefisien regresi ke-k
$\overline{Y}_i$	:	Rata-rata dari variabel tak bebas pengamatan ke-i
$\overline{X}_{ki}$	:	Rata-rata dari variabel bebas ke-k pengamatan ke-i

Pada fungsi linier, nilai elastisitasnya berubah sesuai dengan besarnya faktor produksi dan produksi yang digunakan.

Pendugaan parameter untuk Model Regresi Linier Berganda adalah dengan metode Kuadrat Terkecil Biasa *(Ordinary Least Square)*.

Model Cobb-Douglas untuk Fungsi Produksi

Model *Cobb-Douglas* secara umum dapat dituliskan sebagai berikut :

$$Q_i = AX_{1i}^{\alpha_1} X_{2i}^{\alpha_2} ... X_{mi}^{\alpha_m} e^{u_i}$$

Keterangan:

Q_i : Nilai produksi ke-*i*

A : Indeks Teknologi

X_{mi} : Faktor produksi ke-m pengamatan ke-*i*

α_m : Elastisitas faktor produksi ke-*m*

u_i : Error ke-*i*

Karena pada fungsi produksi *Cobb-Douglas* koefisien regresinya sudah menunjukkan nilai Elastisitas Produksinya, maka dapat dituliskan rumus dari Elastisitas Produksi dari fungsi produksi *Cobb-Douglas* sebagai berikut :

$$E_{p_{X_i}} = \alpha_m$$

Keterangan:

$E_{p_{X_i}}$: Elastisitas Produksi variabel ke-*i*

α_m : Koefisien Regresi ke-*m*

Dan produk marjinalnya adalah sebagai berikut :

$$PM_{X_{ii}} = \alpha_m \frac{\overline{Q}_i}{\overline{X}_{mi}} \qquad i = 1,2,\ldots,n$$

Keterangan:

$PM_{X_{ki}}$: Produk Marjinal untuk variabel ke-*m* dan pengamatan ke-*i*

$\overline{Q}_i$: Mean dari nilai produksi ke-*i* $\qquad i = 1,2,...,n$

$\overline{X}_{mi}$: Mean dari faktor produksi ke-*m* pengamatan ke-*i*

α_m : Koefisien Regresi ke-*m*

m : Banyaknya faktor produksi

Pengujian Koefisien Regresi

Untuk melihat pengaruh variable bebas terhadap variable tak bebas baik secara simultan maupun secara individual maka dilakukan pengujian koefisien regresi.

Koefisien Korelasi

Koefisien Korelasi adalah koefisien yang menggambarkan tingkat keeratan hubungan antara dua variabel atau lebih. Besaran dari Koefisien Korelasi tidak menggambarkan hubungan sebab akibat antara dua variabel atau lebih semata-mata menggambarkan keterkaitan linier antar variabel.

Koefisien Korelasi antara variabel x *dan* y dapat dirumuskan sebagai berikut :

$$r = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$$

dimana,

$$S_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1},\ S_x{}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1},\ S_y{}^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

Keterangan :

S_{xy}	:	Standar Deviasi variabel x dan y
S_x	:	Standar Deviasi variabel x
n	:	Banyak pengamatan
S_y	:	Standar Deviasi variabel y
r	:	Koefisien Korelasi
X_i	:	Variabel bebas ke-i

Koefisien Determinasi

Koefisien Determinasi (R^2) adalah persentase keragaman variabel bebas yang dapat dijelaskan oleh model persamaan regresi. Nilai (R^2) persamaan regresi yang makin mendekati 100% menunjukkan bahwa makin banyak keragaman variabel bebas yang dapat dijelaskan dari persamaan regresi tersebut. Koefisien Determinasi juga digunakan untuk menilai seberapa baik garis regresi sampel mencocokkan data. Jika semua pengamatan terletak pada garis regresi, maka dikatakan sebagai kecocokan sempurna, tetapi ini merupakan kasus yang langka. Koefisien Determinasi memiliki dua sifat penting yaitu :

- Koefisien determinasi $\left(R^2\right)$ merupakan besaran non negatif artinya selalu bernilai positif, batasnya adalah $0 \le R^2 \le 1$.
- Jika R^2 bernilai 1 berarti suatu kecocokan sempurna, sedangkan R^2 bernilai 0 berarti tidak ada hubungan antara variabel tak bebas dengan variabel yang menjelaskannya.

Secara umum, penggunaan R^2_{adj} (disesuaikan) lebih baik daripada R^2, karena R^2 cenderung memberikan gambaran terlalu optimis terhadap kesesuaian regresi, khususnya jika perbandingan antara variabel bebas dengan jumlah observasi nilainya besar.

Pemeriksaan Sisaan

Kegunaan dari pemeriksaan sisaan antara lain yaitu untuk melihat kecocokan model dengan data yang ada dan untuk mengetahui apakah asumsi sisaan dalam model regresi terbaik yang telah dibentuk terpenuhi.

Adapun beberapa asumsi sisaan yang mendasari model regresi terbaik adalah :

1. Asumsi Normalitas

Asumsi ini mensyaratkan bahwa nilai kesalahan dari penduga harus menyebar secara normal dengan rata-rata 0 dan varians σ^2. Uji terhadap asumsi ini dapat dilakukan dengan melihat plot dari probabilitas normal. Apabila mengikuti/mendekati garis lurus maka dapat disimpulkan asumsi kenormalan sudah terpenuhi. Sedangkan secara formal, uji normalitas

dapat dilakukan dengan Uji Kolmogorof Smirnov (KS). Asumsi sisaan berdistribusi normal akan terpenuhi apabila nilai $KS_{hitung.} < KS_{tabel}$.

2. Asumsi tidak adanya Outlier

Untuk mendapatkan model regresi yang terbaik, perlu dilakukan uji terhadap data outlier, langkah-langkah pengujiannya sebagai berikut:

1. Buat Hipotesis
   H_0 = data ke-*j* adalah outlier
   H_1 = data ke-*j* bukan outlier, $j = 1, 2, ..., n$
2. Statistik uji
   $$F = \frac{\left(R^2{}_{p+1} - R^2{}_p\right) n - k - 2}{1 - R^2{}_{p+1}}$$
   Dimana $R^2{}_{p+1}$ adalah Koefisien Determinasi jika modelnya H_1 dan $R^2{}_p$ Koefisien Determinasi Terkoreksi jika modelnya H_0.
3. Kriteria penolakan
   H_0 ditolak jika $F_{hitung} > F_{tabel} = F_{(\alpha;1;n-k)}$ artinya tidak ada outlier dalam sisaan.

3. Asumsi Homoskedastisitas

Asumsi ini mensyaratkan bahwa $E(u_i) = \sigma^2$ untuk $i = 1,2..., n$. Hal ini berarti Varians dari unsur pengganggu sama dengan σ^2. Asumsi ini diuji dengan melihat *predicted value* dengan *standar residual*-nya atau dengan melihat pola gambar Scatterplot model tersebut.

4. Asumsi Multikolinieritas

Multikolinieritas adalah kondisi dimana satu atau lebih variabel penjelas dapat dinyatakan sebagai kombinasi linier dari variabel penjelas lainnya atau secara singkat multikolinieritas berarti adanya hubungan linier antar variabel penjelas. Ada tidaknya multikolinieritas dapat dideteksi dengan melihat nilai R^2, F_{hitung} dan t_{hitung}. Kemungkinan adanya multikolinieritas :

(1) Nilai R^2 tinggi tetapi tidak satupun atau sangat sedikit koefisien yang diduga, signifikan secara statistik.

(2) F_{hitung} tinggi (signifikan) akan tetapi tidak satupun koefisien yang signifikan secara parsial.

5. Asumsi Autokorelasi

Autokorelasi dapat didefenisikan sebagai korelasi antara anggota serangkaian observasi yang diurutkan menurut waktu (data deret waktu) atau ruang (data crossestional). Pengujian autokorelasi dilakukan dengan menggunakan statistik uji dari *Durbin-Watson (d)*:

$$D-W=d=\frac{\sum_{i=2}^{n}\left(e_i-e_{i-1}\right)^2}{\sum_{i=1}^{n}e_i^2}$$

Keterangan :

d : Statistik *Durbin-Watson*

e_i : Nilai residual ke-i $\quad i=1,2,...,n$

Pemilihan Model Regresi Terbaik

Untuk memilih persamaan regresi terbaik dapat digunakan beberapa prosedur, yaitu prosedur *stepwise*, prosedur *forward selection*, dan prosedur *backward elimination*. Dalam *forward selection*, pembentukkan model terbaik dilakukan dengan menambahkan variabel satu persatu. Regresi linier sederhana memulai tahap awal dengan memasukkan satu variabel bebas. Tahap selanjutnya adalah menambahkan variabel bebas terbaru sehingga ada dua variabel bebas dalam model. Penambahan diulangi sampai semua variabel masuk ke dalam model. Dalam *backward elimination,* pembentukkan model terbaik dilakukan dengan membuat terlebih dahulu model regresi untuk semua variabel bebas. Selanjutnya, mengurangi variabel satu per satu sampai tinggal satu variabel bebas. Sedangkan prosedur *stepwise* merupakan gabungan dari prosedur *forward selection* dan *backward elimination.* Prosedur *stepwise* hampir mirip dengan prosedur *forward selection,* hanya saja dalam prosedur *stepwise*, apabila ada dua variabel bebas saling berkorelasi, maka hanya salah satu variabel yang dimasukkan ke dalam model. Pemilihan variabel yang dimasukkan berdasarkan variabel bebas yang memiliki korelasi lebih besar dengan variabel tak bebas.

HASIL DAN PEMBAHASAN

Pendugaan Fungsi Produksi Model Regresi Linier Berganda

Fungsi produksi Model Regresi Linier Berganda dapat dituliskan sebagai berikut :

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\varepsilon$$

Keterangan :

Y : Produksi Padi (Kg)

X_1 : Luas Lahan (Ha)

X_2 : Jumlah Jam Kerja (Jam)

X_3 : Jumlah Benih (Kg)

X_4 : Jumlah Pupuk (Kg)

X_5 : Jumlah Pestisida (Liter)

β_0 : Konstanta

β_i : Nilai Produk Marjinal (Koefisien Regresi)

ε : Residual (sisaan)

Berdasarkan analisis awal secara umum menunjukkan adanya hubungan positif yang signifikan antara faktor-faktor produksi yang ada dengan produksi padi, maka akan dilakukan pembentukkan model fungsi produksi dengan model regresi linier berganda untuk melihat pengaruh faktor-faktor produksi terhadap produksi padi. Pendugaan Model Regresi Linier Berganda yaitu menggunakan Metode Kuadrat Terkecil *(Ordinary Least Square)*.

Model Pendugaan regresi linier berganda yang diperoleh dengan Metode Kuadrat Terkecil dapat dilihat pada Tabel 4.

Tabel 4. Nilai-nilai koefisien untuk Model Regresi Linier Berganda

Variabel	Koefisien model	
	B	Std. Error
(Constant)	-6.196	22.475
Luas lahan	266.414	71.324
Jam Kerja	0.543	0.079
Benih	3.942	1.962
Pupuk	0.597	0.189
Pestisida	81.513	33.451

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi dilakukan dengan Uji-F. Hasil Uji-F secara simultan untuk melihat pengaruh faktor-faktor produksi secara bersama-sama terhadap hasil produksi padi diperoleh nilai F-hitung sebesar 486.075 dengan *p-value* 0.0001. Karena *p-value* lebih kecil dari taraf nyata 5%, maka dapat dinyatakan bahwa pengaruh faktor-faktor produksi secara bersamaan untuk semua fungsi produksi yang dicobakan adalah nyata pada taraf 5%. Hasil pendugaan dan pengujian secara individual Model Regresi Linier Berganda dengan metode OLS didapat bahwa semua faktor produksi nyata pada taraf 5% kecuali konstanta.

Menentukan Model Regresi Terbaik pada Model Regresi Linier Berganda

Metode yang dapat digunakan untuk menentukan model regresi terbaik adalah metode *forward selection*, *backward elimination*, dan *stepwise*. Ke tiga metode tersebut semuanya dibahas dalam skripsi ini.

***Forward Selection Procedure*:**

1. Variabel Jumlah Pestisida (X_5) dimasukkan ke dalam model.

   Berdasarkan Lampiran 2 pada output SAS 6.12 metode *Forward Selection* diperoleh $R^2 = 0.935$ dan R^2-*adjusted* = 0.93 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 93%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F dan uji t. Diperoleh Nilai $F_{hitung} = 791.16 > F_{tabel} = 2.37$, dan $t_{hitung} = 791.16 > t_{tabel} = 1.671$ yang berarti variabel Pestisida berpengaruh secara signifikan terhadap produksi padi. Hal ini juga didukung oleh nilai *p-value* variabel Pestisida yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak

signifikan karena nilainya lebih besar dari 0.05 yaitu 0.4616. Model yang diperoleh pada langkah 1 sebagai berikut:

$$Y = 413,63X_5$$

2. Variabel Jumlah Jam Kerja (X_2) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.960$ dan R^2-*adjusted* = 0.959 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 95.9%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 653.43 > F_{tabel} = 3.15$ yang berarti secara bersama-sama, variabel Pestisida dan Jumlah Jam Kerja berpengaruh secara nyata terhadap produksi padi. Secara individual, kedua variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.4321. Model yang diperoleh pada langkah 2 sebagai berikut:

$$Y = 0,61X_1 + 232,42X_5$$

3. Variabel Luas Lahan (X_1) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.972$ dan R^2-*adjusted* = 0.971 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.1%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 616.00 > F_{tabel} = 2.76$ yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, dan Luas Lahan berpengaruh secara nyata terhadap produksi padi. Secara individual, ketiga variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.1711. Model yang diperoleh pada langkah 3 sebagai berikut:

$$Y = 360,85X_1 + 0,55X_2 + 119,02X_5$$

4. Variabel Jumlah Pupuk (X_4) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.978$ dan R^2-*adjusted* = 0.976 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.6%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 573.13 > F_{tabel} = 2.53$ yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, Luas Lahan, dan Pupuk berpengaruh secara nyata terhadap produksi padi. Secara individual, keempat variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.9101. Model yang diperoleh pada langkah 4 sebagai berikut:

$$Y = 309,05X_1 + 0,57X_2 + 0,69X_4 + 85,48X_5$$

5. Variabel Jumlah Benih (X_3) dimasukkan ke dalam model.

   Diperoleh $R^2 = 0.979$ dan R^2-*adjusted* = 0.977 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.7%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 486.08 > F_{tabel} = 2.37$ yang berarti secara bersama-sama, seluruh variabel berpengaruh secara nyata terhadap produksi padi. Secara individual, kelima variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.7839. Model yang diperoleh pada langkah 5 sebagai berikut:

$$Y = 266,414X_1 + 0,543X_2 + 3,94X_3 + 0,597X_4 + 81,51X_5$$

   Berdasarkan analisis di atas, dapat disimpulkan bahwa model regresi terbaik pada Model Regresi Linier Berganda terdapat pada langkah 5 yaitu model yang mengikutsertakan seluruh variabel ke dalam model karena baik pengujian secara simultan maupun secara individual, memberikan kesimpulan yang sama yaitu signifikan pada taraf nyata 5% kecuali pada konstanta.

***Backward Elimination Procedure*:**

Sesuai dengan prosedur *Backward Elimination*, langkah pertama adalah memasukkan seluruh variabel bebas ke dalam model. Berdasarkan Lampiran 2 pada output SAS 6.12 metode *Backward Elimination,* diperoleh $R^2 = 0.979$ dan R^2-*adjusted* = 0.977 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.7%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 486.08 > F_{tabel} = 2.37$ yang berarti secara bersama-sama, seluruh variabel berpengaruh secara nyata terhadap produksi padi. Secara individual, kelima variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.7839. Seluruh variabel signifikan terhadap hasil produksi baik secara simultan ataupun individual, karena itu proses ***Backward Elimination*** dihentikan. Model yang diperoleh sebagai berikut:

$$Y = 266,414X_1 + 0,543X_2 + 3,94X_3 + 0,597X_4 + 81,51X_5$$

***Stepwise Procedure*:**

1. Variabel Jumlah Pestisida (X_5) dimasukkan ke dalam model.

   Berdasarkan Lampiran 2 pada output SAS 6.12 metode *stepwise elimination* diperoleh $R^2 = 0.935$ dan R^2-*adjusted* = 0.93 menunjukkan bahwa kontribusi pengaruh

yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 93%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F dan uji t. Diperoleh Nilai $F_{hitung} = 791.16 > F_{tabel} = 2.37$, dan $t_{hitung} = 791.16 > t_{tabel} = 1.671$ yang berarti variabel Pestisida berpengaruh secara signifikan terhadap produksi padi. Hal ini juga didukung oleh nilai *p-value* variabel Pestisida yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.4616. Model yang diperoleh pada langkah 1 sebagai berikut:

$$Y = 413,63X_5$$

2. Variabel Jumlah Jam Kerja (X_2) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.960$ dan R^2-*adjusted* = 0.959 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 95.9%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 653.43 > F_{tabel} = 3.15$ yang berarti secara bersama-sama, variabel Pestisida dan Jumlah Jam Kerja berpengaruh secara nyata terhadap produksi padi. Secara individual, kedua variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.4321. Model yang diperoleh pada langkah 2 sebagai berikut:

$$Y = 0,61X_1 + 232,42X_5$$

3. Variabel Luas Lahan (X_1) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.972$ dan R^2-*adjusted* = 0.971 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.1%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 616.00 > F_{tabel} = 2.76$ yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, dan Luas Lahan berpengaruh secara nyata terhadap produksi padi. Secara individual, ketiga variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.1711. Model yang diperoleh pada langkah 3 sebagai berikut:

$$Y = 360,85X_1 + 0,55X_2 + 119,02X_5$$

4. Variabel Jumlah Pupuk (X_4) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.978$ dan R^2-*adjusted* = 0.976 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.6%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 573.13 > F_{tabel} = 2.53$ yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, Luas Lahan, dan Pupuk berpengaruh secara nyata terhadap produksi padi. Secara individual, keempat variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.9101. Model yang diperoleh pada langkah 4 sebagai berikut:

$$Y = 309,05X_1 + 0,57X_2 + 0,69X_4 + 85,48X_5$$

5. Variabel Jumlah Benih (X_3) dimasukkan ke dalam model.

Diperoleh $R^2 = 0.979$ dan R^2-*adjusted* = 0.977 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.7%.

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai $F_{hitung} = 486.08 > F_{tabel} = 2.37$ yang berarti secara bersama-sama, seluruh variabel berpengaruh secara nyata terhadap produksi padi. Secara individual, kelima variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Kecuali nilai *p-value* konstanta yang tidak signifikan karena nilainya lebih besar dari 0.05 yaitu 0.7839. Model yang diperoleh pada langkah 5 sebagai berikut:

$$Y = 266,414X_1 + 0,543X_2 + 3,94X_3 + 0,597X_4 + 81,51X_5$$

Pendugaan Fungsi Produksi Model Regresi Cobb-Douglas

Fungsi produksi Model Regresi *Cobb-Douglas* dapat dituliskan sebagai berikut :

$$\ln Y_i = \hat{\beta}_0 + \hat{\beta}_1 \ln X_{1i} + \ldots + \hat{\beta}_m \ln X_{mi} + e_i$$

Keterangan :

$\hat{\beta}_0$	:	Nilai dugaan Ln A
$\hat{\beta}_m$	:	Nilai dugaan koefisien regresi variabel ke-*m*
e_i	:	*Error* ke-*I* *i=1,2,...,n*
Y_i	:	Variabel takbebas ke-i
X_{mi}	:	Variabel bebas ke-m pengamatan ke-i

Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi dilakukan dengan Uji-F. Hasil Uji F secara simultan untuk melihat pengaruh faktor-faktor produksi secara bersama-sama terhadap hasil produksi padi diperoleh nilai F-hitung sebesar 412.769 dengan *p-value* 0.0001. Sehingga dapat dinyatakan bahwa pengaruh faktor-faktor produksi secara bersamaan untuk semua fungsi produksi yang dicobakan pada model *Cobb-Douglas* adalah nyata pada taraf 5%. Hasil pendugaan dan pengujian secara individual Model Regresi *Cobb-Douglas* dengan metode OLS tersebut diperoleh bahwa faktor produksi seperti Luas Lahan, Jam Kerja, dan Benih nyata pada taraf 5%, sedangkan

faktor Pupuk dan Pestisida tidak nyata pada taraf 5%. Untuk melihat seberapa besar faktor-faktor produksi pada Model Regresi *Cobb-Douglas* dapat menjelaskan variasi data pada produksi padi, maka dapat dilihat dari output SAS.6.12 pada lampiran yaitu nilai R^2= 0.9759 dan R^2-*adjusted* = 0.9735, ini menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.35 %.

Menentukan model regresi terbaik dari fungsi produksi *Cobb-Douglas.*
***Forward Selection Procedure*:**

1. Variabel Jumlah Pestisida (X_5) dimasukkan ke dalam model.
   Berdasarkan output SAS 6.12 metode *Forward Selection* diperoleh R^2= 0.940 dan R^2-*adjusted* = 0.939 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 93.9%.
   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F dan uji t. Diperoleh Nilai F_{hitung} = 869.18 > F_{tabel} = 2.37, dan t_{hitung} = 869.18 > t_{tabel} = 1.671 yang berarti variabel Pestisida berpengaruh secara signifikan terhadap produksi padi. Hal ini juga didukung oleh nilai *p-value* variabel Pestisida yang lebih kecil dari 0.05. Model yang diperoleh pada langkah 1 sebagai berikut:
   $$Y = 5,947 + 1,062X_5$$

2. Variabel Jumlah Jam Kerja (X_2) dimasukkan ke dalam model.
   Diperoleh R^2= 0.9633 dan R^2-*adjusted* = 0.962 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 96.2%.
   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 709.37 > F_{tabel} = 3.15 yang berarti secara bersama-sama, variabel Pestisida dan Jumlah Jam Kerja berpengaruh secara nyata terhadap produksi padi. Secara individual, kedua variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Model yang diperoleh pada langkah 2 sebagai berikut:
   $$Y = 3,65 + 0,41X_2 + 0,58X_5$$

3. Variabel Luas Lahan (X_1) dimasukkan ke dalam model.
   Diperoleh R^2= 0.971 dan R^2-*adjusted* = 0.969 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 96.9%.
   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 594.24 > F_{tabel} = 2.76 yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, dan Luas Lahan berpengaruh secara nyata terhadap produksi padi. Secara individual, ketiga variabel

berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Model yang diperoleh pada langkah 3 sebagai berikut:

$$Y = 4{,}16 + 0{,}29X_1 + 0{,}40X_2 + 0{,}24X_5$$

4. Variabel Jumlah Benih (X_3) dimasukkan ke dalam model.

   Diperoleh R^2 = 0.974 dan R^2-*adjusted* = 0.972 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.2%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 495.15 > F_{tabel} = 2.53 yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, Luas Lahan, dan Jumlah Benih berpengaruh secara nyata terhadap produksi padi. Secara individual, hanya variabel Pestisida yang tidak berpengaruh secara signifikan terhadap hasil produksi, karena nilai *p-value* nya adalah 0.0586 yaitu lebih besar dari 0.05. Model yang diperoleh pada langkah 4 sebagai berikut:

$$Y = 3{,}99 + 0{,}25X_1 + 0{,}37X_2 + 0{,}11X_3 + 0{,}22X_5$$

5. Variabel Jumlah Benih (X_3) dimasukkan ke dalam model.

   Diperoleh R^2 = 0.976 dan R^2-*adjusted* = 0.974 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.4%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 412.77 > F_{tabel} = 2.37 yang berarti secara bersama-sama, seluruh variabel berpengaruh secara nyata terhadap produksi padi. Secara individual, variabel Pupuk dan Pestisida tidak berpengaruh secara signifikan terhadap hasil produksi. Hal ini didukung oleh nilai *p-value* kedua variabel yang lebih besar dari 0.05. Menurut *Summary of Forward Selection Procedure*, model terbaik adalah model yang tidak mengikutsertakan variable Pupuk ke dalam model.

***Backward elimination Procedure*:**

1. Seluruh variabel dimasukkan ke dalam model.

   Diperoleh hasil sama seperti pada langkah 5 prosedur *forward selection*, yaitu variabel yang tidak signifikan dihilangkan dari model, yaitu variabel Pestisida.

2. Variabel Pestisida (X_5) dieliminasi dari model.

   Berdasarkan output SAS 6.12 metode *backward elimination* diperoleh R^2 = 0.975 dan R^2-*adjusted* = 0.973 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.3%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F dan uji t. Diperoleh Nilai F_{hitung} = 502.6 > F_{tabel} = 2.53, yang berarti secara bersama-sama, variabel Jumlah Jam Kerja, Luas Lahan, Jumlah Benih, dan Jumlah Pupuk berpengaruh secara nyata terhadap produksi padi. Secara individual,

keempat variabel tersebut juga berpengaruh secara signifikan terhadap hasil produksi, karena nilai *p-value*nya lebih kecil dari 0.05. Model yang diperoleh pada langkah 2 sebagai berikut:

$$Y = 3,17 + 0,298X_1 + 0,441X_2$$
$$+0,1X_3 + 0,114X_4$$

***Stepwise Procedure*:**

1. Variabel Jumlah Pestisida (X_5) dimasukkan ke dalam model.

   Berdasarkan Lampiran 3 pada output SAS 6.12 metode *stepwise* diperoleh R^2 = 0.940 dan R^2-*adjusted* = 0.939 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 93.9%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F dan uji t. Diperoleh Nilai F_{hitung} = 869.18 > F_{tabel} = 2.37, dan t_{hitung} = 869.18 > t_{tabel} = 1.671 yang berarti variabel Pestisida berpengaruh secara signifikan terhadap produksi padi. Hal ini juga didukung oleh nilai *p-value* variabel Pestisida yang lebih kecil dari 0.05. Model yang diperoleh pada langkah 1 sebagai berikut:

   $$Y = 5,947 + 1,062X_5$$

2. Variabel Jumlah Jam Kerja (X_2) dimasukkan ke dalam model.

   Diperoleh R^2 = 0.9633 dan R^2-*adjusted* = 0.962 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 96.2%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 709.37 > F_{tabel} = 3.15 yang berarti secara bersama-sama, variabel Pestisida dan Jumlah Jam Kerja berpengaruh secara nyata terhadap produksi padi. Secara individual, kedua variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai *p-value* kedua variabel yang lebih kecil dari 0.05. Model yang diperoleh pada langkah 2 sebagai berikut:

   $$Y = 3,65 + 0,41X_2 + 0,58X_5$$

3. Variabel Luas Lahan (X_1) dimasukkan ke dalam model.

   Diperoleh R^2 = 0.971 dan R^2-*adjusted* = 0.969 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 96.9%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 594.24 > F_{tabel} = 2.76 yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, dan Luas Lahan berpengaruh secara nyata terhadap produksi padi. Secara individual, ketiga variabel berpengaruh secara signifikan terhadap hasil produksi. Hal ini juga didukung oleh nilai

p-value kedua variabel yang lebih kecil dari 0.05. Model yang diperoleh pada langkah 3 sebagai berikut:

$$Y = 4{,}16 + 0{,}29X_1 + 0{,}40X_2 + 0{,}24X_5$$

4. Variabel Jumlah Benih (X_3) dimasukkan ke dalam model.

   Diperoleh R^2= 0.974 dan R^2-*adjusted* = 0.972 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.2%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 495.15 > F_{tabel} = 2.53 yang berarti secara bersama-sama, variabel Pestisida, Jumlah Jam Kerja, Luas Lahan, dan Jumlah Benih berpengaruh secara nyata terhadap produksi padi. Secara individual, hanya variabel Pestisida yang tidak berpengaruh secara signifikan terhadap hasil produksi, karena nilai *p-value* nya adalah 0.0586 yaitu lebih besar dari 0.05. Model yang diperoleh pada langkah 4 sebagai berikut:

$$Y = 3{,}99 + 0{,}25X_1 + 0{,}37X_2 + 0{,}11X_3 + 0{,}22X_5$$

5. Variabel Jumlah Benih (X_3) dimasukkan ke dalam model.

   Diperoleh R^2= 0.976 dan R^2-*adjusted* = 0.974 menunjukkan bahwa kontribusi pengaruh yang dapat dijelaskan oleh faktor-faktor produksi yang ada terhadap variasi data pada produksi padi sebesar 97.4%.

   Untuk melihat pengaruh faktor-faktor produksi secara simultan terhadap hasil produksi, dilakukan uji-F. Diperoleh Nilai F_{hitung} = 412.77 > F_{tabel} = 2.37 yang berarti secara bersama-sama, seluruh variabel berpengaruh secara nyata terhadap produksi padi. Secara individual, variabel Pupuk dan Pestisida tidak berpengaruh secara signifikan terhadap hasil produksi. Hal ini didukung oleh nilai *p-value* kedua variabel yang lebih besar dari 0.05. Menurut *Summary of Forward Selection Procedure*, model terbaik adalah model yang tidak mengikutsertakan variable Pupuk ke dalam model.

   Berdasarkan analisis pada tiga metode, dapat disimpulkan bahwa model regresi terbaik pada Model Regresi *Cobb-Douglas* adalah model yang tidak mengikutsertakan variabel Pestisida dan Pupuk ke dalam model karena baik pengujian secara simultan maupun secara individual, memberikan kesimpulan yang sama yaitu signifikan pada taraf nyata 5%. Modelnya adalah sebagai berikut:

$$Y = 3{,}78 + 0{,}353X_1 + 0{,}435X_2 + 0{,}118X_3$$

   dengan nilai R^2=0.973. Nilai ini lebih kecil daripada model regresi terbaik pada analisis regresi linier berganda menurut prosedur seleksi maju yaitu R^2=0.979. Sehingga, dapat ditarik kesimpulan bahwa model terbaik yang dapat digunakan untuk menduga hasil produksi pada kasus produksi padi adalah model terbaik pada regresi linier berganda yaitu model yang mengikut sertakan seluruh variabelnya ke dalam model. Oleh karena itu juga analisis pada model *Cobb-Douglas* tidak dilanjutkan.

Pemeriksaan asumsi sisaan pada model yang terbentuk

Asumsi kenormalan

Model yang terbaik adalah model yang memenuhi semua asumsi sisaan. Hasil Plot kenormalan untuk model linier berganda dan model *Cobb-Douglas* dapat dilihat pada gambar output dari *Software* Minitab 14. Berdasarkan gambar plot kenormalan dapat dilihat bahwa gambar Plot Kenormalan yang diperoleh baik Model Linier Berganda ataupun Model *Cobb-Douglas* telah memenuhi asumsi sisaan yaitu asumsi kenormalan, hal ini dapat dilihat dari plot nilai-nilai sisaan pada gambar mengikuti atau berada diser garis diagonal lurus, *p-value* yang lebih besar dari taraf nyata, dan nilai *Ks hitung < Ks Tabel.* Akan tetapi gambar Plot Kenormalan pada Regresi Linier Berganda relatif lebih baik atau lebih mengikuti garis lurus diagonal apabila dibandingkan dengan gambar Plot Kenormalan pada Regresi *Cobb-Douglas*.

Asumsi tidak adanya outlier

Berdasarkan analisis tentang outlier, diperoleh adanya dugaan data outlier yaitu pada data ke-24 dari 57 sampel data.

1. Buat Hipotesis

H_0 = data ke-24 adalah outlier

H_1 = data ke-24 bukan outlier

2. Statistik uji

Berdasarkan Nilai Koefisien Determinasi dan Koefisien Determinasi Terkoreksi tentang data outlier, diperoleh:

$$F = \frac{(0,982 - 0,98)57 - 6 - 2}{1 - 0,982} = \frac{(0,002)49}{0,018} = 5,44$$

3. Kriteria penolakan

$F_{hitung} = 5{,}44$ dan $F_{tabel} = 4{,}08$

Jadi, $F_{hitung} > F_{tabel}$, artinya adalah menolak H_0. Dan data ke-24 bukanlah outlier. Dapat disimpulkan juga bahwa tidak ada outlier.

Multikolinieritas

Salah satu asumsi yang harus dipenuhi untuk mendapatkan model yang terbaik adalah tidak adanya multikolinieritas antar variabel bebas. Apabila terjadi multikolinieritas, maka akan terjadi ketidaksesuaian dalam model. Untuk memeriksa adanya Multikolinieritas di antara variabel bebas adalah dengan menggunakan nilai VIF *(Varians Inflation Factor).* Jika nilai $VIF > 10$, maka di dalam model tersebut terjadi multikolinieritas. Hasil perhitungan dengan bantuan program Minitab 14 diperoleh nilai-nilai VIF untuk masing-masing variabel bebas dalam model Regresi Linier Berganda terbaik dirangkum dalam tabel berikut:

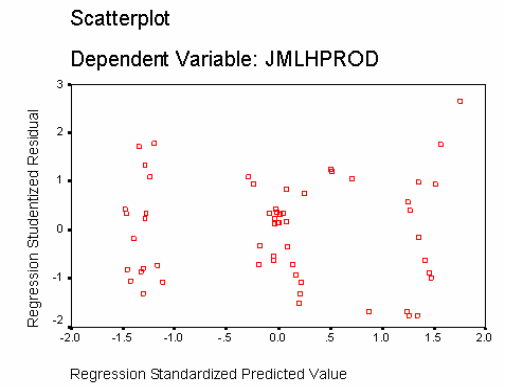
Tabel 10. Nilai VIF variabel bebas model regresi linier berganda

Variabel	Model Linier Berganda Terbaik
	Nilai VIF
Luas Lahan	10.3
Jam Kerja	8.5
Benih	5.0
Pupuk	3.3
Pestisida	15.2

Berdasarkan tabel di atas dapat dilihat bahwa nilai VIF dari masing-masing variabel bebas ada yang bernilai >10, artinya terdapat multikolinieritas dalam model tersebut. Akan tetapi tidak langsung disimpulkan bahwa dalam model tersebut terdapat multikolinieritas, menyimpulkan apakah di dalam model regresi yang digunakan terdapat multikolinieritas atau tidak, dapat dilihat dari tanda koefisien korelasi masing-masing variabel. Dari model regresi terlihat bahwa semua koefisien regresi untuk model terbaik Linier Berganda semua koefisien regresinya bertanda positif, dan koefisien korelasi pada lampiran antara variabel bebas juga bertanda positif. Berdasarkan dari beberapa teori di atas, maka dapat disimpulkan bahwa multikolinieritas yang terdapat dalam model regresi linier berganda di atas pada data petani padi tersebut dapat diabaikan karena tanda koefisien regresi dan koefisien korelasi dari masing-masing variabel bebas adalah sama yaitu bertanda positif.

Asumsi homoskedastisitas

Asumsi yang harus dipenuhi berikutnya adalah asumsi bahwa varians sama untuk semua pengamatan atau homoskedastisitas, dari hasil output pada program SPSS 11.6 didapat Scatterplot sebagai berikut :



Gambar 3. Scatterplot homoskedastisitas

Dari gambar di atas dapat dilihat bahwa titik-titik data menyebar di atas dan di bawah atau diser angka 0, serta tidak membentuk pola tertentu. Maka dapat disimpulkan bahwa Model Regresi Linier Berganda yang akan digunakan untuk menduga produksi padi terbebas dari masalah Heteroskedastisitas.

Asumsi autokorelasi

Pengujian yang selanjutnya adalah masalah autokorelasi. Dari hasil perhitungan, nilai *Durbin Watson* pada output Minitab 14 adalah sebesar 1.540. Menurut kriteria uji Durbin Watson yang hipotesanya disederhanakan, untuk model regresi linier berganda berarti berada di daerah tidak terjadi autokorelasi karena $1.540 < d_U = 2.35$. Sehingga kesimpulannya data sisaan terbebas dari masalah autokorelasi.

Elastisitas Produksi dan Produk Marjinal

Berdasarkan model dugaan fungsi *Cobb-Douglas* di atas dapat diketahui nilai elastisitas dan produk marjinal dari masing-masing faktor produksi. Nilai Elastisitas Produksi dan Produk Marjinal dari masing-masing faktor produksi untuk Model Linier Berganda dan model *Cobb-Douglas* diperoleh bahwa Produk Marjinal dari semua faktor produksi kedua model bernilai positif. Artinya penggunaan faktor-faktor produksi tersebut masih dapat ditingkatkan guna mendapatkan hasil produksi padi yang meningkat pula. Pada model linier berganda pada faktor Luas Lahan, nilai Elastisitas Produksinya paling besar daripada faktor produksi yang lain yaitu sebesar 369064,22. Ini berarti luas lahan yang digarap petani padi masih dapat ditingkatkan. Artinya jika luas lahan yang digarap petani padi diperluas sebesar 1%, maka akan meningkatkan hasil produksi padi sebesar 369064,22%. Dilihat dari Nilai Elastisitas dan Produk Marjinal untuk kedua model terlihat bahwa model linier berganda lebih memiliki arti dari pada model *Cobb-Douglas*. Karena, pada model *Cobb-Douglas* Nilai Elastisitas dan Produk Marjinalnya mendekati nol.

Pada model regresi *Cobb-Douglas*, nilai Elastisitas Produksi terbesar adalah faktor luas lahan yaitu sebesar 3.44 berarti jika faktor produksi lain dianggap tetap, maka penambahan 1% Luas lahan akan meningkatkan produktivitas produksi padi sebesar 3.44 %. Nilai Elastisitas faktor benih sebesar 0.39, artinya apabila faktor produksi lain dianggap tetap, maka penambahan 1 % jumlah benih akan meningkatkan produksi padi sebesar 0.39 %.

Jumlah elastisitas produksi menggambarkan fase pergerakan usaha *(return to scale)*. Pendugaan fungsi produksi padi di desa darat sawah ini memberikan hasil jumlah elastisitas produksi ($\sum E_{pi} = 4.25$) yang lebih dari 1. Hal ini berarti apabila faktor-faktor produksi ditambah secara bersamaan akan menghasilkan tambahan produksi yang proporsinya lebih besar dari proporsi pertambahan faktor-faktor produksi tersebut.

Pada model linier berganda, Produk Marjinal tertinggi terjadi pada faktor produksi Jumlah Luas Lahan. Hal ini menunjukkan bahwa untuk setiap penambahan 1 unit faktor produksi Luas Lahan, maka penambahan terbesar oleh faktor produksi Luas Lahan sebesar 266,414 unit. Hal ini menunjukkan bahwa luas lahan yang digarap para petani padi di Desa Darat Sawah Ulu masih dapat ditingkatkan lagi. Sedangkan Nilai Produk Marjinal terendah terjadi pada faktor Jam Kerja, yaitu sebesar 0,543. Artinya untuk setiap penambahan 1 unit faktor produksi Jam Kerja, maka penambahan terbesar oleh faktor produksi Jam Kerja adalah sebesar 0,543 unit.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil Analisis dan pembahasan di atas, maka dapat ditarik beberapa kesimpulan sebagai berikut :

5. Model yang digunakan sebagai model pendugaan adalah Model Regresi Linier Berganda penuh yaitu :

$$Y = 266.4X_1 + 0.54X_2 + 3.94X_3 + 0.60X_4 + 81.51X_5$$

6. Berdasarkan hasil pengujian baik secara simultan maupun secara parsial untuk model di atas menunjukkan kesimpulan yang sama yaitu semua faktor-faktor produksinya berpengaruh secara nyata terhadap produksi padi.
7. Berdasarkan nilai Koefisien Determinasi R^2 dan Koefisien Determinasi Terkoreksi menunjukkan bahwa pengaruh faktor produksi padi untuk Model Linier Berganda relatif lebih besar bila dibandingkan dengan pada Model *Cobb-Douglas* untuk kasus produksi padi..
8. Dari hasil pemeriksaan asumsi menunjukkan bahwa Model Linier Berganda relatif lebih baik dari Model *Cobb-Douglas*.
9. Model Linier berganda relatif lebih berarti dari model *Cobb-Douglas*.
10. Berdasarkan nilai Koefisien Determinasi Terkoreksi, model *Cobb-Douglas* analisisnya tidak dilanjutkan karena menurut prosedur seleksi maju pada Analisis Regresi *Cobb-Douglas*, model regresi terbaiknya adalah model yang tidak mengikutsertakan variabel Pupuk dan Pestisida ke dalam model dengan nilai $R^2 = 0.973$ nilai ini lebih kecil daripada model regresi terbaik pada analisis regresi linier berganda menurut prosedur seleksi maju yaitu $R^2 = 0.979$.

Saran

Untuk kesempurnaan penelitian selanjutnya di masa yang akan datang, maka penulis nenyarankan sebagai berikut :

1. Dalam menentukan variabel atau faktor produksi diharapkan dapat lebih banyak lagi variabel atau faktor produksi yang dilibatkan guna menaksir hasil produksi padi.
2. dalam pembentukkan model perlu dicari bentuk transformasi yang lain seperti fungsi kuadrat, trasenden, dan lain sebagainya.
3. Untuk penelitian di masa yang akan datang dapat dilanjutkan pada kesempurnaan analisis asumsi sisaannya.

DAFTAR PUSTAKA

Djarwanto, 1996, *Mengenal Beberapa Uji Statistik Dalam Penelitian,* Liberty, Yogyakarta

Draper, N.R, 1992, *Analisis Regresi Terapan*. Edisi ke-2. PT. Gramedia Pustaka Utama, Jakarta.

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

Sugiyono, 2004, *Statistik Nonparametris Untuk Penelitian,* Alfabeta, Bandung.

Sugiyono, 2006, *Metode Penelitian Kuantitatif Kualitatif dan R &D,* Alfabeta, Bandung.

Sukartawi, 1990. *Analisis Fungsi Produksi*, Jakarta: Gramedia Jakarta.

Sukartawi, 1990. *Teori Ekonomi Produksi dengan Pokok Bahasan Analisis Fungsi Produksi Cobb-Douglas*, Rajawali Pers, Jakarta.

Sukartawi, 1994, *Alokasi Faktor Produksi Dengan Pokok Bahasan Analisis Fungsi Cobb Douglas*, PT. Raja Grafindo Persada, Jakarta.