Jurnal Matematika dan Pengetahuan Alam
Vol. 5, No. 1, Th. 1996 ISSN: 0853-0033

PEMILIHAN MODEL TERBAIK DENGAN ANALISIS REGRESSI BERGANDA

(Chosing of the Best Model by Multiple Regression Analysis)

Maiyastri

Jurusan Matematika, FMIPA Universitas Andalas

ABSTRACT

To chose the model of the multiple regression does not only look at the significancy of the explanatory variables, but also need to see the best model criteria from R^2_{adj}, C_p Mallow and PRESS (these values measure the accurateness of the model). It is followed by assumption test such as collinearity, homoscedasticity, etc.

This paper is purposed to know the best model by multiple regression analysis and to apply the model to the group of data. Data, which are used, IP in the first year (semester I and II) as the response variable and NEM SMA for Mathematics, Biology, Physics, Chemistry, and English as the explanatory variabels on Pharmacy student 1990.

Therefore, the best model is: IP = 1,41 + 0,0789 (Math) + 0,117 (Physics). So, it is said that, especially for the student of Pahrmacy FMIPA-UNAND 1990, the successful of a student in the first year is influenced by the his/her basic knowledge about Mathematics and Physics.

PENDAHULUAN

Analisis regressi adalah salah satu analisis statistika yang berguna untuk melihat hubungan antara dua peubah atau lebih, hubungan ini bisa berupa antara satu peubah respon dan satu peubah bebas atau antara satu peubah respon dan beberapa peubah bebas. Analisis regressi untuk tipe kedua dinamakan Analisis Regressi Berganda (*Multiple Regression Analysis*), dan modelnya adalah sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} + \epsilon_i \qquad (1)$$

$i = 1, 2, \dots, n$

$\beta_0, \beta_1, \dots, \beta_p$ = parameter yang akan diduga

ϵ_i = sisaan.

Mencari model terbaik dari persamaan di atas berarti mencari peubah bebas yang mana saja yang betul-betul memberikan pengaruh kepada peubah respon Y_i, masalah ini dinamakan **pemilihan model**.

Tujuan karya tulis ini adalah untuk mempelajari pemilihan model terbaik dengan teknik analisis regressi berganda dan menerapkannya pada suatu kelompok data.

MODEL ANALISIS REGRESSI BERGANDA

1. Pendugaan

Dalam analisis regressi berganda akan dibahas hubungan antara peubah respon Y dengan p buah peubah penjelas, modelnya seperti persamaan (1) dan dalam bentuk vektor dan matriks menjadi:

$$\underline{Y} = X \underline{\beta} + \underline{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

dengan asumsi:

$$\underline{Y} = N(X\underline{\beta}, \sigma^2 I)$$
$$\underline{\epsilon} = N(0, \sigma^2 I)$$

Pendugaan koefisien $\underline{\beta}$ dilakukan dengan metode kuadrat terkecil dengan meminimumkan Q^2 sedangkan $Q = \underline{\epsilon}'\underline{\beta}$, yaitu dengan membuat turunan pertamanya sama dengan nol. Dugaan $\underline{\beta}$ didapatkan sebagai berikut:

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y} \text{ ; dengan syarat } (X'X)^{-1} \text{ ada,}$$

Sifat penduga ini tak bias, karena:

$$E(\hat{\underline{\beta}}) = \underline{\beta} \text{ ; dengan ragamnya}$$

$$\text{Ragam}(\hat{\underline{\beta}}) = \sigma^2(X'X)^{-1} \text{ ; di mana:}$$

$$\sigma^2 = s^2 = \underline{Y}'(I - X(X'X)^{-1}X')\underline{Y} = \underline{Y}'\underline{Y} - \hat{\underline{\beta}}'X'\underline{Y}$$

2.Analisis Ragam

Analisis ragam pada analisis regresi berganda berguna untuk menguji hipotesis di bawah ini:

$$H_0 : \underline{\beta} = \underline{0}$$
$$H_0 : \underline{\beta} = \underline{0}$$

Bentuk tabel analisis ragam untuk model di atas (persamaan (1)) dapat dilihat pada Tabel 1.

Tabel 1. Analisis Ragam

Sumber Keragaman	Derajat Bebas	Jumlah Kuadrat	Kuadrat Tengah	F hitung
Regresi	p	JK(R)	KT(R)	KT(R)/KT(S)
Sisaan	n-p-1	JK(S)	KT(S)	
Total	n-1	JK(T)		

PEMILIHAN MODEL TERBAIK

Pada bagian ini akan dibahas beberapa kriteria yang digunakan dalam pemilihan model terbaik pada analisis regressi berganda.

1. Kuadrat Tengah Sisaan = s^2: semakin kecil s^2 model semakin bagus.
2. R^2 dan $R^2_{adjusted}$: semakin besar nilainya semakin bagus modelnya.
3. PRESS : model terbaik dengan PRESS paling kecil.
4. C_p Mallow : model terbaik nilai $C_p = p$.
5. Regressi Bertatar (*The Stepwise Regression*)

MASALAH SISAAN DAN DATA BERPENGARUH

1. Sisaan

Dalam menerima suatu model regressi tidak cukup hanya dengan melihat hasil analisis ragam saja, akan tetapi perlu dilakukan pemeriksaan asumsi, apakah asumsi yang digunakan terpenuhi atau tidak. Uji asumsi tentang kehomogenan ragam (homoskedasitas), kelinearan dan kebebasan dapat dilihat dari plot antara sisaan e_i dengan Y_i (Y_i duga).

2. Pencilan dan Data Berpengaruh

Mendeteksi pencilan dan data berpengaruh berguna untuk melihat apakah kesimpulan yang diambil berubah dengan masuknya data tersebut. Pencilan dilihat dari e_i dan data berpengaruh dari D_i (*Cook Distance*).

3. Transformasi Data

Transformasi data tidak selalu harus dilakukan. Apabila pada plot sisaan terlihat bahwa pola data tidak linear atau masih ada asumsi yang belum terpenuhi, maka data perlu ditransformasi, sampai semua asumsi-asumsi terpenuhi.

BAHAN DAN METODA

Untuk penerapan pada data digunakan data dari mahasiswa Jurusan Farmasi FMIPA-UNAND Padang angkatan 1990. Sebagai peubah respon adalah: Indeks Prestasi (IP) tahun I (Semester I dan II) dan peubah bebas adalah: NEM dari lima mata pelajaran SMA yaitu: X_1 = Matematika, X_2 = Biologi, X_3 = Fisika, X_4 = Kimia, dan X_5 = Bahasa Inggris. Jumlah data 60 buah.

Metode pemilihan model terbaik pertama dilakukan dengan mencari sebagian persamaan regressi terbaik (*best subset regression*) dari melihat nilai-nilai: $s = \sqrt{s^2}$, R^2, dan R^2_{adj}, dan C_p. Dari hasil ini dipilih persamaan paling baik, untuk mendukung hasil ini dilakukan juga analisis regressi bertatar (*stepwise regression*). setelah itu dilakukan pengujian asumsi dan dilihat apakah ada pencilan dan data berpengaruh. Di sini apabila pembuangan data pencilan dan data berpengaruh telah "mengobati" masalah yang ada, maka transformasi data tidak diperlukan lagi.

HASIL DAN PEMBAHASAN

Hasil regressi sebagian terbaik (*best subset regression*) dan nilai PRESS untuk beberapa model yang telah diseleksi dapat dilihat pada Tabel 2.

Tabel 2. Nilai PRESS, C_p, R^2_{adj}, dan s^2 untuk Beberapa Model Persamaan

Model Persamaan	PRESS	Cp	R²adj	s²
Y=f(X1,X3)	0,3847	1,7	34,0%	0,36548
Y=f(X3,X4)	0,6911	2,8	33,3%	0,36962
Y=f(X1,X2,X3)	0,5492	2,4	35,0%	0,36485
Y=f(X1,X3,X4)	0,5458	2,7	34,7%	0,36570
Y=f(X2,X3,X4)	0,9954	4,3	33,7%	0,37124

Nilai PRESS terkecil adalah untuk model Y = f(X1,X3), tetapi nilai PRESS model persamaan regresi pada Tabel 3 hampir sama.

$$Y = 1.31 + 0.0793\ X1 + 0.140\ X3$$

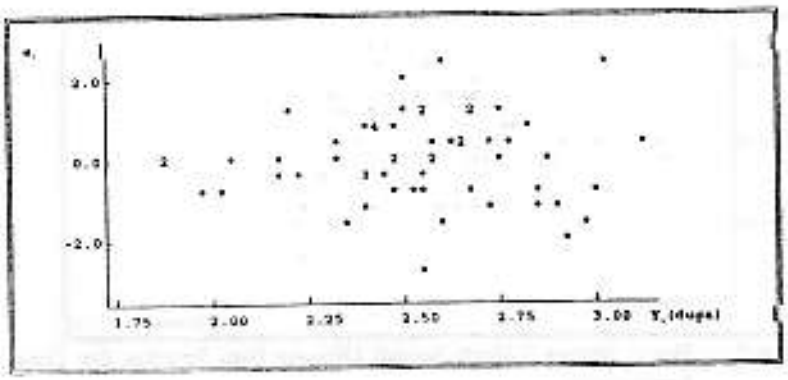
Peubah (X)	Koefisien	Galat Baku	Nilai-t	Nilai
Intercep	1.3085	0.2130	6.14	0.000
C1	0.07929	0.03879	2.04	0.046
C3	0.14006	0.04516	3.10	0.003

$s = 0.3655$ $R^2 = 37.0\%$ $R^2_{adj} = 4.8\%$

Tabel 3. Analisis Ragam

Sumber Keragaman	Derajat Bebas	Jumlah Kuadrat	Kuadrat Tengah	F hitung
Regresi	2	4,4656	2,2328	16,72*
Sisaan	57	7,6137	0,1336	
Total	59	12,0793		

Dari hasil di atas terlihat model Y = f(X1,X3) cukup layak untuk diterima, tetapi uji asumsi belum dilakukan karena kita tidak bisa menerima begitu saja hasil analisis tersebut. Untuk menganalisa uji asumsi (kehomogenan ragam dan kelinearan) dilihat dari plot sisaan e_i dengan Y_i(duga). Pada Gambar 1 dapat dilihat plot e_i dengan Y_i(duga) pada model Y = f(X1,X3).



Gambar 1. Plot e_i dengan Y_i (duga)

Dari Gambar 1 terlihat bahwa uji asumsi kehomogenen ragam belum dipenuhi dan

ada beberapa sisaan yang lebih besar dari |2| (data pencilan), karena itu perlu dilakukan analisis lebih lanjut, yaitu melihat apakah ada data yang berpengaruh dan perlukah dilakukan transformasi data. Dari 60 buah data ternyata ada beberapa data yang dianggap pencilan dan data yang berpengaruh.

Analisis selanjutnya adalah menduga persamaan regresi untuk model yang sama yaitu Y=f(X1,X3) dengan membuang data pencilan dan data berpengaruh, tabel analisis ragamnya dapat dilihat pada Tabel 4 dan persamaannya adalah:

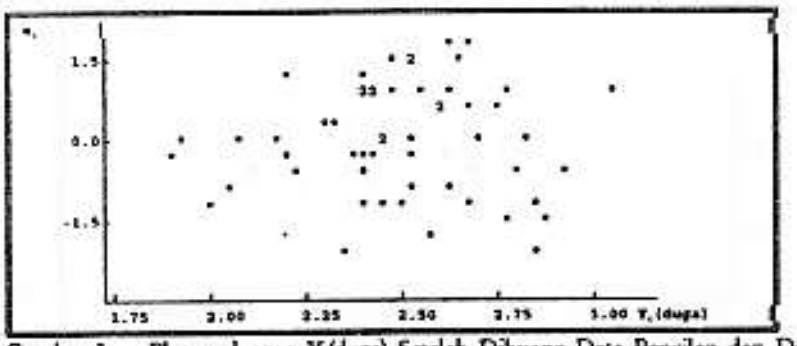
$$Y = 1.41 + 0.0789\ X1 + 0.117\ X3$$

Peubah (X)	Koefisien	Galat Baku	Nilai-t Hitung	Nilai p
Constant	1.4051	0.1776	7.91	0.000
C1	0.07890	0.03733	2.11	0.039
C3	0.11564	0.04187	2.79	0.007
s = 0.2970	R^2 = 42.2%	R^2_{adj} = 40.0%		

Tabel 4. Analisis Ragam (Setelah Dibuang Data Pencilan dan Data yang Berpengaruh)

Sumber Keragaman	Derajat Bebas	Jumlah Kuadrat	Kuadrat Tengah	F hitung
Regresi	2	3,3515	1,6757	19,00*
Sisaan	52	4,5854	0,0882	
Total	54	7,9369		

Plot e_i dengan Y_i(duga) dapat dilihat pada Gambar 2.



Gambar 2. Plot e_i dengan Y_i(duga) Setelah Dibuang Data Pencilan dan Data Berpengaruh

Dari Gambar 2 terlihat bahwa asumsi kehomogenen ragamnya telah terpenuhi

sehingga model ini cukup baik untuk diambil sebagai model terbaik.

KESIMPULAN

Pemilihan model terbaik dengan analisis regresi berganda dilakukan melalui suatu proses yang cukup panjang dan dengan kriteria tertentu seperti: Kuadrat Tengah Sisaan (KTS), R^2, R^2_{adj}, C_p Mallow, PRESS, regresi bertatar (*stepwise regression*). Setelah model didapat dilakukan uji asumsi kehomogenen ragam, kelinearan, kebebasan dan lain-lain. Apabila ada asumsi yang dilanggar perlu dilihat apakah data yang berpengaruh atau data pencilan yang menyebabkan asumsi dilanggar atau apakah perlu dilakukan transformasi data.

Dari contoh data yang digunakan model terbaiknya adalah :

$$IP = 1{,}41 + 0{,}0789\ (\text{Matematika}) + 0{,}117\ (\text{Fisika})$$

Jadi dapat dikatakan bahwa k.berhasilan seseorang di tahun I (semester I dan II) khusus pada jurusan Farmasi FMIPA UNAND Padang berkaitan erat dengan kemampuannya/keberhasilannya pada mata pelajaran Matematika dan Fisika. Kalau dilihat dari koefisien regressinya Fisika relatif lebih besar pengaruhnya dari Matematika.

DAFTAR PUSTAKA

Aunuddin. 1989. *Analisis Data*. Departemen Pendidikan dan Kebudayaan Dirjen Dikti. Pusat Antar Universitas. Institut Pertanian Bogor.

Drapper, N. dan H. Smith. 1981. *Analisis Regresi Terapan*, (edisi kedua), terjemahan Ir. Bambang Sumantri. P.T. Gramedia. Jakarta.

Montgomery, D. C. & E. A. Peck. 1992. *Introduction to Linear Regression Analysis*, (2nd ed). John Wiley. New York.

Weisberg, S. 1985. *Applied Linear Regression*. John Wiley. New York.