**S-7**

# Seleksi Variabel Dalam Analisis Regresi Multivariat Multipel

**Neneng Sunengsih**

**Staf Jurusan Statistika FMIPA UNPAD**

**Abstrak**

Salah satu tujuan analisis regresi adalah untuk tujuan prediksi. Semakin banyak variabel yang masih dalam model akan semakin baik model tersebut dalam melakukan fungsi prediksinya. Namun, banyaknya variabel yang masuk memberikan permasalahan dalam sulitnya mengumpulkan data dan kontrol setiap variabel. Sehingga diperlukan adanya seleksi variabel yaitu memilih variabel yang benar-benar memberikan informasi dalam keakuratan prediksi. Dalam makalah ini dijelaskan bagaimana seleksi variabel dalam analisis regresi multivariate multipel melalui pendekatan prosedur pemilihan variabel dan semua bagian regresi yang mungkin.

Kata Kunci : Analisis Regresi Multivariat, Prosedur Seleksi, Best Subset Regression

## I. PENDAHULUAN

Pemilihan variabel prediktor yang tepat dalam pembuatan model regresi khususnya untuk tujuan prediksi merupakan satu hal yang sangat penting. (Breiman & Friedman, 1997; Bilodeau & Brenner, 1999). Umumnya seleksi model dalam pembentukan model regresi terbaik untuk regresi univariate dilakukan dengan metode maju (forward selection), metode mundur (backward selection) dan juga metode bertahap (Stepwise selection). Seleksi variabel dalam analisis regresi yang melibatkan variabel dependen lebih dari satu atau yang dikenal dengan model regresi multivariat relatif lebih kompleks karena setiap variabel prediktor tidak hanya berhubungan dengan satu variabel dependen namun lebih dari satu variabel dependen.

Dalam model regresi multivariat multiple (MMR) terdapat $q$ variabel dependen ($y_1, y_2, \ldots, y_q$) yang diprediksi melalui hubungan linier $k$ variabel independen ($x_1, x_2, \ldots, x_k$).

Model statistic untuk MMR adalah :

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times(k+1)}\mathbf{B}_{(k+1)\times q} + \mathbf{E}_{n\times q} \quad (1)$$

Dengan Y, adalah matriks variabel respon dengan *n* unit pengamatan independen dan *q* variabel acak normal. Selanjutnya X merupakan matriks variabel prediktor (matriks desain) dengan ordo *k+1* dengan kolom pertama adalah vektor satuan, **B** adalah matriks paramter regresi yang akan diestimasi dan **E** adalah matriks kekeliruan (*error term*)

Permasalahan yang sering dihadapi dalam pembentukan model regresi baik model univariat maupun multivariat adalah menentukan himpunan variabel prediktor terbaik sehingga dapat memberikan hasil prdiksi yang paling akurat. Dalam kaitan pemilihan variabel independen terdapat dua kriteria yang saling bertentangan satu dengan yang lainnya :

1. Agar persamaan bermanfaat bagi tujuan prediksi, peneliti biasanya ingin memasukkan sebanyak mungkin variabel independen ***X*** kedalam model sehingga diperoleh nilai prediksi yang handal
2. Karena untuk memperoleh informasi dari banyak variabel serta pemonitorannya seringkali diperlukan biaya yang tinggi, maka salah satu caranya adalah dengan memasukkan sedikit mungkin variabel independen kedalam model. Atau dengan kata lain, dalam pembentukan model selalu diinginkan model yang paling sederhana.

Kompromi kedua ekstrim tersebut yang biasanya disebut *pemilihan model regresi terbaik*. Seperti yang telah digambarkan bahwa permasalahan yang akan dihadapi dalam pembentukan model MMR adalah adalah memilih variabel prediktor yang tepat untuk dimasukkan dalam model MMR sehingga prediksi yang dilakukan memiliki tingkat akurasi yang tinggi (McQuarrie & Tsai, 1998). Terdapat dua pendekatan yang selama ini dilakukan :

1. Menemukan himpunan bagian variabel perdiktor terbaik "best" ***X*** untuk setiap satu variabel respon **Y** dengan menggunakan satu atau beberapa kriteria pemilihan model yang telah tersedia dalam beberapa paket program statistik. Pemilihan model didasarkan pada prosedur univariat sebanyak ***q*** kali sesuai dengan banyaknya variabel respon ***Y*** dalam model. Tentunya dengan cara ini akan diperoleh dengan ***q*** subset variabel prediktor berbeda, satu set untuk setiap variabel respon ***Y***.
2. Menemukan himpunan variabel prediktor untuk semua variabel respon ***Y*** secara simultan dimana satu set himpunan variabel prediktor yang diperoleh merupakan himpunan variabel prediktor terbaik "*the best*" untuk semua variabel respon ***Y*** dengan menggunakan ekspresi matriks, kriteria pemilihan model regresi terbaik secara univariat dapat digunakan.

Sparks et al. (1985) telah melakukan perbandingan metodologi pemilihahan model secara univariat dibandingkan dengan pemilihan model secara simultan dan

menyimpulkan bahwa pemilihan variabel prediktor terbaik lebih baik dilakukan secara simultan dengan dua alasan. Alasan pertama adalah bahwa proses perhitungan akan lebih efisien karena waktu yang dibutuhkan untuk set himpunan variabel prediktor terbaik akan berkurang dair ***q*** kali menjadi hanya satu kali. Alasan kedua adalah bahwa peneliti terkadang ingin menetapkan himpunan bagian variabel variabel prediktor yang diharapkan merangkum keseluruhan variabel prediktor yang menjadi target penelitian. Hal ini akan memperkecil biaya yang terkait dengan masalah sampling.

Dalam makalah ini, akan dijelaskan komputasi pemilihan himpunan bagian variabel prediktor terbaik "*best*" yang dapat digunakan untuk memprediksi semua variabel respon ***y*** secara simultan menggunakan kriteria pemilihan model mutltivariat yang saat ini sudah berkembang. Beberapa metode pemihan variabel yang diperkenalkan diantaranya adalah

1. Prosedur Stepwise (*Stepwise Procedure*)
   a. *Forward Selection,*
   b. *Forward Stepwise Regression,*
   c. *Backward Elimination,*
2. Prosedur Semua Komungkinan Regresi *(All-Posible Regression)*
   a. *Mean Square Error (MSE),*
   b. *Coefficient of Multiple Determination ($R^2$),*
   c. *Adjusted Coefficient of Multiple Determination ($AdjR^2$),*
   d. *Akaike's Information Criterion (AIC),*
   e. *the Corrected Form of Akaike's Information Criterion (AICc),*
   f. *Hannan and Quinn Information Criterion (HQ),*
   g. *Corrected Form of Hannan and Quinn ($HQ_c$) Information Criterion,*
   h. *Schwarz's Criterion (SC),*
   i. *Mallow's* $\mathbf{C_P}$ .

## TINJAUAN PUSTAKA

### Pemilihan Model Regresi Terbaik

Prosedur regresi bertahap (*Stepwise Regression*) dan semua kemungkinan regresi (*all-possible-regression*) adalah dua prosedur pemilihan himpunan variabel prediktor terbaik. Dalam aplikasinya pemilihan model regresi terbaik dilakukan dengan menghilangkan atau memasukkan variabel independent secara bertahap (*Stepwise)* dan selanjutnya melakukan pengujian semua himpunan bagian variabel prediktor terbaik yang memenuhi beberapa criteria yang ditetapkan dan memilih satu model regresi terbaik.

Tabel di bawah ini menunjukkan notasi dan defnisi dari vairabel dan fungsi yang akan digunakan dalam mendefinisikan kriteria pemilihan model.

**Table 1** Notasi dan Defnisi dari Variabel dan Fungsi Yang Digunakan

| Simbol | Defnisi |
|---|---|
| *N* | Banyaknya unit observasi/pengamatan |
| *P* | Banyaknya parameter regresi termasuk intercept |
| *K* | Banyaknya variabel predictor **x** dalam model penuh "*full model*" |
| *Q* | Banyaknya variabel respon **y** |
| **Y** | Matriks variabel dependen/respon |
| **X** | Matriks variabel independent/predictor yang akan dimasukkan dalam model dengan kolom pertama adalah vector 1. |
| $\mathbf{X_p}$ | Sub Matriks **X** yang berisi vector satu dan kolom yang lain berisi variabel predictor yang terpilih $x_p$ dalam model. |
| **J** | Matriks satuan dengan ordo $q \times q$ |
| $\Lambda$ | Statistik Wilks' $\Lambda$, sejalan dengan variabel acak **F**, didefinisikan sebagai rasio dari dua variabel acak independent chi-square dibagi dengan masing-masing derajat bebasnya. |
| $\hat{\Sigma}$ | Jumlah kuadrat error untuk model penuh dengan memasukkan intercept |
| $\hat{\Sigma}_p$ | Jumlah kuadrat error /galat dari model dengan *p* parameter termasuk intercept. |
| Ln | Logaritma natural |
| $\lvert\bullet\rvert$ | Fungsi determinan |

**Metode Regresi Bertahap (Stepwise Regression Method)**

*Stepwise regression* terdiri dari tiga prosedur yaitu *Forward Selection, Forward Stepwise Regression*, dan *Backward Elimination* (Barrett & Gray, 1994; Rencher, 1995).

Pada umumnya criteria yang digunakan untuk menambahkan ataupun membuang variabel independen $x$ dalam pembentukan model regresi terbaik salah satunya adalah partial Wilks' $\Lambda$ atau partial **F**.

Partial Wilks $\Lambda$ diformulasikan sebagai berikut :

$$\Lambda(x_1, x_2, x_3, ..., x_p) = \frac{\left|\mathbf{Y}'[\mathbf{I} - \mathbf{X_p}(\mathbf{X_p}'\mathbf{X_p})^{-1}\mathbf{X_p}']\mathbf{Y}\right|}{\left|\mathbf{Y}'[\mathbf{I} - \frac{1}{N}\mathbf{J}]\mathbf{Y}\right|} \quad (2)$$

dikenal dengan distribusi Wilks $\Lambda$ dengan derjaat bebas ($q$,1,$n$-$p$-1) atau ditulis $\Lambda_{q,1,n-p-1}$

Sebuah variabel independen akan dimasukkan ke dalam model jika hasil perhitungan partial Wilks' $\Lambda$ diperoleh nilai yang lebih kecil dari *cut point* yang dietapkan sebelumnya. Sedangkan variabel akan dibuang dari model jika dari hasil perhitungan diperoleh nilai partial Wilks' $\Lambda$ terlalu besar.

**Prosedur Maju (*Forward Selection*)**

Tahapan prosedur Prosedur *forward selection* adalah :

1. Menghitung model regresi tanpa variabel independen $\boldsymbol{Y} = \beta_0$

2. Menghitung nilai partial Wilks'$\Lambda$ untuk setiap variabel independen yang dimasukkan satu-persatu ke dalam model.
3. Membandingkan nilai partial Wilks'$\Lambda$ yang diperoleh dengan nilai Wilks' $\Lambda$ terkecil pada tingkat signifikansi tertentu misalkan saja $\Lambda_o$
    a. Jika partial Wilks'$\Lambda$ < $\Lambda_o$, masukkan variabel independen yang diuji ke dalam model.
    b. Jika partial Wilks'$\Lambda$ > $\Lambda_o$, variabel independen yang diuji tidak layak dimasukkan ke dalam model.
4. Kemudian dilakukan kembali perhitungan statistik partial Wilks' $\Lambda$ untuk variabel-variabel yang belum dimasukkan dalam model sampai tidak ditemukan lagi variabel yang menghasilkan nilai statistik Wilks' $\Lambda$ yang signifikan. Setiap varibel yang sudah masuk dalam model akan tetap berada dalam model.

**Regresi Maju Bertahap (*Forward Stepwise Regression*)**

Prosedur *forward stepwise regression* adalah modifikasi dari prosedur *forward selection* dimana perbedaannya terletak pada variabel yang sudah dimasukkan dalam model mungkin akan dikeluarkan lagi dari model jika dinilai setelah dimasukkan beberapa variabel independen yang lain tidak terlalu penting. Sama halnya dalam metode *forward selection*, variabel-variabels independen dimasukkan satu-persatu ke dalam model dengan menggunakan kriteria partial Wilks' $\Lambda$. Variabel yang akan dimasukkan ke dalam model adalah variabel yang memiliki nilai statistik partial Wilks' $\Lambda$ yang signifikan atau memiliki nilai partial Wilks' $\Lambda$ yang lebih kecil dari *cut poit* yang telah ditetapkan sebelumnya.

**Prosedur Mundur (*Backward Elimination )***

Metode *Eleminasi Backward* dilakukan sebagai berikut :

1. Menghitung persamaan regresi dengan memasukkan semua variabel $x$ ke dalam model.
2. Menghitung nilai partial Wilks'$\Lambda$ untuk setiap vairabel independen, seolah-olah variabel tersebut merupakan variabel terakhir yang dimasukkan ke dalam model regresi.
3. Membandingkan nilai partial Wilks'$\Lambda$ terkecil dengan $\Lambda$ pada taraf signifikansi yang ditetapkan sebut saja $\Lambda_o$.
    a. Jika partial Wilks'$\Lambda$ > $\Lambda_o$ variabel independen yang mengasilkan partial Wilks'$\Lambda$ dibuang dari persamaan regresi.
    b. Jika partial Wilks'$\Lambda$ < $\Lambda_o$ ambilah persamaan regresi tersebut.
3. Lakukan perhitungan partial Wilks' $\Lambda$ untuk setiap variabel sisanya $q-1$. Dan variabel yang dinilai tidak terlalu penting dimasukkan dalam model dikeluarkan dari model. Proses ini dilakukan sampai diperoleh nilai terbesar partial Wilks' $\Lambda$

signifikan. Hasil ini menunjukkan bahwa hubungan variabel dalam model tidak redundan atau tumpang tindih dengan variabel yang lain dalam model.

***Semua Kemungkinan Regresi (*All-Possible-Regression*)***

Prosedur semua kemungkinan regresi (*all-possible-regression*) mempertimbangkan semua himpunan variabel prediktor untuk dimasukkan dalam model diawali dengan memasukkan variabel konstanta $x_0$ sampai variabel ke-n $x_n$ dan melakukan pemilihan himpunan variabel prediktor terbaik dengan menggunakan kriteria rata-rata kuadrat error /*Residual mean square error* (**MSE**), koefisien determinasi multiple/ *coefficient of multiple determination* ($\mathbf{R^2}$), koefisien regresi yang disesusaikan / *adjusted coefficient of multiple determination* (***Ajd*R** $^2$), kriteria Informasi Akaike's/ *Akaike's information criterion* (**AIC**), Kriteria Informasi Hannan dan Quinn (**HQ**), Kriteria Schwarz (**BIC**), dan Mallow's $C_P$.

**Kriteria Residual Mean Square Error (MSE)**

*Residual mean square error (MSE)* adalah taksiran varians galat untuk setiap model yang dapat diformulasikan sebagai berikut :

$$\mathbf{MSE} = \left| \frac{\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}}{n-p} \right| \tag{3}$$

dengan $\hat{\boldsymbol{\Sigma}}_{\mathbf{p}} = \mathbf{Y}'[\mathbf{I} - \mathbf{X}_{\mathbf{P}}(\mathbf{X}'_{\mathbf{P}}\mathbf{X}_{\mathbf{P}})^{-1}\mathbf{X}'_{\mathbf{P}}]\mathbf{Y}$ adalah jumlah kuadrat error untuk model dengan *p* parameters termasuk intercept. Model terbaik adalah model dengan nilai ***MSE*** minimum

**Kriteria Koefisien Determinasi Multiple ($\mathbf{R}^2$)**

Krieteria $\mathbf{R}^2$ adalah juga merupakan metode menemukan himpunan variabel prediktor terbaik untuk memprediksi variabel dependen melalui model regresi linier yang diperoleh dari data sampel. Metode ini dinilai efisien dengan menyajikan semua kemungkinan model regresi dan menunjukkan nilai $R^2$ sesuai dengan banyaknya variabel independen dalam model. Koefisien determinasi multipel $\mathbf{R}^2$ dapat dihitung dengan menggunakan formulasi berikut :

$$\mathbf{R}^2 = \left| [\mathbf{Y}'(\mathbf{I} - \tfrac{1}{\mathbf{n}}\mathbf{J})\mathbf{Y}]^{-1}[\mathbf{Y}'(\mathbf{X}_{\mathbf{p}}(\mathbf{X}'_{\mathbf{p}}\mathbf{X}_{\mathbf{p}})^{-1}\mathbf{X}'_{\mathbf{p}} - \tfrac{1}{\mathbf{n}}\mathbf{J})\mathbf{Y}] \right| \tag{4}$$

Metode koefisien determinasi multipel selalu menetapkan model terbaik adalah model dengan $\mathbf{R}^2$ terbesar untuk setiap unit variabel prediktor yang dipertimbangkan dalam model.

**Kriteria *Adjusted* R** $^2$

Karena banyaknya parameter dalam model regresi tidak dimasukkan dalam perhitungan Since **R** $^2$, sehingga **R** $^2$ tidak mungkin menurun pada saat banyaknya parameter *p* bertambah. Hal ini merupakan satu kelemahan metode **R$^2$**. Sebagai satu bentuk penyempurnaan diperkenalkan metode koefisien determinasi multipel yang disesuaikan /*adjusted coefficient of multiple determination* (***Ajd*R** $^2$) sebagai alternatif krieteria dalam pemilihan model terbaik. Metode ***Ajd*R** $^2$ sama halnya dengan metode **R**$^2$ yaitu menetapkan model terbaik adalah model yang memiliki ***Adj*R** $^2$ terbesar. Fromulasi dari ***Adj*R** $^2$ dapat dituliskan sebagai berikut :

$$\mathrm{AdjR}^2 = 1 - \frac{(n-1)(1-\mathrm{R}^2)}{n-p} \tag{5}$$

**Kriteria Informasi Akaike's (AIC)**

Prosedur **AIC** (Akaike, 1973) digunakan untuk mengevaluasi seberapa baik model sementara dibandingkan dengan model sebenarnya dengan melihat perbedaan antara nilai ekpektasi dari vektor ***y*** dari model sebenarnya dengan model sementara dengan menggunakan jarak Kullback-Leibler (K-L). Jarak Kullback-Leibler (K-L) adalah jarak antara densitas sebenarnya dan densitas taksiran untuk setiap model dengan formulasi sebagai berikut :

$$\mathbf{AIC} = \ln|\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}| + \frac{2pq + q(q+1)}{n} \tag{6}$$

Model terbaik dalam memprediksi ***y*** secara simulatan adalah model yang memilki nilai **AIC**'terkecil.

**Penyesuaian Pada Kriteria Informasi Akaike's (AIC$_c$)**

Bedrick & Tsai (1994) memberikan catatan bahwa kriteria informasi Akaike's mungkin akan memberikan hasil yang bias untuk sampel kecil, sehingga dilakukan perbaikan pada kriteria AIC dan menghasilkan kriteria AICc dengan formulasi sebagai berikut :

$$\mathbf{AIC}_{\mathbf{C}} = \ln|\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}| + \frac{(n+p)q}{n-p-q-1} \tag{7}$$

Himpunan variabel prediktor terbaik ***x*** adalah himpunan variabel yang memiliki nilai **AIC**$_C$'s minimum.

**Kriteri Informasi Hannan Dan Quinn (HQ)**

Kriteria informasi **HQ** yang diperkenalkan oleh Hannan dan Quinn (1979), dan telah banyak digunakan dalam model autoregressive dan untuk model regresi linier (McQuarrie & Tsai, 1998). Formulasi dari HQ dapat dituliskan sebagai beirkut :

$$\mathbf{HQ} = \ln|\widehat{\mathbf{\Sigma}}_{\mathbf{p}}| + \frac{2ln(ln(n))pq}{n} \tag{8}$$

Model terbaik adalah model yang memiliki nilai **HQ** terkecil.

**Penyesuaian Kriteria Informasi Hannan dan Quinn ($HQ_c$)**

Kriteria Informasi Hannan dan Quinn (**HQ**) akan bias untuk ukuran sampel kecil McQuarrie & Tsai (1998). Sehingga McQuarrie & Tsai (1998) melakukan perbaikan untuk metode ini dengan hasil HQc yang diformulasikan sebagai berikut :

$$\mathbf{HQ}_{\boldsymbol{C}} = \ln|\widehat{\mathbf{\Sigma}}_{p}^{2}| + \frac{2ln(ln(n))pq}{n-p-q-1} \tag{9}$$

Model terbaik adalah model yang memiliki nilai nilai HQc terkecil

**Kriteria Schwarz's (BIC)**

Perhitungan kriteria infomrasi Schwarz 's Bayesian untuk setiap model menggunakan jarak Kullback-Leibler (K-L) yang dapat digunakan untuk mengidentifikasi model terbaik. Kriteri ini dapat diformulasikan sebagai berikut :

$$\mathbf{BIC} = \ln|\widehat{\mathbf{\Sigma}}_{p}^{2}| + \frac{ln(n)p}{n} \tag{10}$$

Model terbaik adalah model yang memiliki nilai BIC minimum.

**Mallow's $C_P$**

Kiteria $\boldsymbol{C_P}$ diperkenalkan oleh Mallow's (1973) untuk regresi univariat dan dikembangkan oleh Spark et al. (1983) untuk model Regresi Multivariat Multiple (MMR). Kriteria $C_P$ dilakukan dengan mengevaluasi total rata-rata kuadrat galat *n* nilai yang sesuai untuk setiap himpunan bagian regresi. Kriteri $C_P$ diperoleh dengan formulasi sebagai berikut :

$$\boldsymbol{C_P} = (n-k)\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{p}} + (2p-n)I \tag{11}$$

dengan $\hat{\boldsymbol{\Sigma}} = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$, dan **I** adalah matriks identitas dengan ukuran ($q \times q$). Prosedur identifikasi himpunan bagian terbaik dari variabel preiktor ***x*** dengan satu variabel memberikan dua nilai $C_P$ minimum dan dekat $p$**I** (Spark et al., 1983; Rencher, 1995). Jika $2p - n < 0$, akan menghasilkan nilai determinan $|\mathbf{C_P}|$ negatif dan tidak reliable (Spark et al., 1983). Oleh karena itu, dilakukan modifikasi untuk $|\mathbf{C_P}|$ yang telah dikemukakan oleh Spark et al. (1983) sebagai solusi dari masalah ini, agar nilai dari $|\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}|$ selalu positif dan dapat ditulis sebagai berikut :

$$\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}} = \frac{\mathbf{C_p} + (n-2p)\mathbf{I}}{n-k} \tag{12}$$

Ketika bias adalah 0, $\mathbf{C_p} = p\mathbf{I}$, dan (2.12) menjadi

$$\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}} = \frac{n-p}{n-k}\mathbf{I} \tag{13}$$

Oleh karena itu, himpunan bagian yang dihasilkan selalu memuaskan dengan kriteria sebagai berikut :

$$\left|\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}\right| \le \left(\frac{n-p}{n-k}\right)^{q} \tag{14}$$

**CONTOH APLIKASI**

Contoh dalam makalah ini mengenai kandungan kimia dalam daun tembako. Dalam penelitian ini diambil sampel daun tembakao sebanyak 25 daun. Terdapat tiga variabel dependen yaitu :

$Y_1$ : Tingkat rokok terbakar per inci dalam 1000 detik

$Y_2$ : Kandungan gula dalam daun (%)

$Y_3$ : Kandungan nikotin dalam daun (%)

Variabel independennya adalah :

$X_1$ : Kandungan Nitrogen (%)
$X_2$ : Kandungan Chlorine (%)
$X_3$ : Kandungan Potassium (%)
$X_4$ : Kandungan Phosphorus (%)
$X_5$ : Kandungan Calcium (%)
$X_6$ : Kandungan Magnesium (%)

Data hasil penelitian ini disajikan dalam Tabel 3.1

Spark et al. (1983) menggunakan data ini dan menemukan variabel-variabel independen terbaik yang akan dimasukkan dalam model didasarkan pada kriteria multivariat ***Cp***. Dari data di atas akan dilakukan perhitungan untuk mendapatkan variabel-variabel independen yang akan dimasukkan dalam model regresi sebagai prediktor dengan mempertimbangkan semua kriteria. Untuk mempermudah perhitungan telah dibuatkan program SAS/IML untuk masing-masing prosedur.

Ada tiga tahap penting yang dibutuhkan dalam menggunakan program SAS IML untuk semua prosedur pemilihan model regresi terbaik sehingga data dapat dianalisis oleh program tersebut.

1. Baca data menggunakan DATA statement
2. Lakukan pemberian nama untuk setiap variabel
3. Jalankan program

**Tabel 2** Data Tembakau

| Subject ID | Variabel Dependen | | | Variabel Independen | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Y1** | **Y2** | **Y3** | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** |
| 1 | 1.55 | 20.05 | 1.38 | 2.02 | 2.90 | 2.17 | 0.51 | 3.47 | 0.91 |
| 2 | 1.63 | 12.58 | 2.64 | 2.62 | 2.78 | 1.72 | 0.50 | 4.57 | 1.25 |
| 3 | 1.66 | 18.56 | 1.56 | 2.08 | 2.68 | 2.40 | 0.43 | 3.52 | 0.82 |
| 4 | 1.52 | 18.56 | 2.22 | 2.20 | 3.17 | 2.06 | 0.52 | 3.69 | 0.97 |
| 5 | 1.70 | 14.02 | 2.85 | 2.38 | 2.52 | 2.18 | 0.42 | 4.01 | 1.12 |
| 6 | 1.68 | 15.64 | 1.24 | 2.03 | 2.56 | 2.57 | 0.44 | 2.79 | 0.82 |
| 7 | 1.78 | 14.52 | 2.86 | 2.87 | 2.67 | 2.64 | 0.50 | 3.92 | 1.06 |
| 8 | 1.57 | 18.52 | 2.18 | 1.88 | 2.58 | 2.22 | 0.49 | 3.58 | 1.01 |
| 9 | 1.60 | 17.84 | 1.65 | 1.93 | 2.26 | 2.15 | 0.56 | 3.57 | 0.92 |
| 10 | 1.52 | 13.38 | 3.28 | 2.57 | 1.74 | 1.64 | 0.51 | 4.38 | 1.22 |
| 11 | 1.68 | 17.55 | 1.56 | 1.95 | 2.15 | 2.48 | 0.48 | 3.28 | 0.81 |
| 12 | 1.74 | 17.97 | 2.00 | 2.03 | 2.00 | 2.38 | 0.50 | 3.31 | 0.98 |
| 13 | 1.93 | 14.66 | 2.88 | 2.50 | 2.07 | 2.32 | 0.48 | 3.72 | 1.04 |
| 14 | 1.77 | 17.31 | 1.36 | 1.72 | 2.24 | 2.25 | 0.52 | 3.10 | 0.78 |
| 15 | 1.94 | 14.32 | 2.66 | 2.53 | 1.74 | 2.64 | 0.50 | 3.48 | 0.93 |
| 16 | 1.83 | 15.05 | 2.43 | 1.90 | 1.46 | 1.97 | 0.46 | 3.48 | 0.90 |
| 17 | 2.09 | 15.47 | 2.42 | 2.18 | 0.74 | 2.46 | 0.48 | 3.16 | 0.86 |
| 18 | 1.72 | 16.85 | 2.16 | 2.16 | 2.84 | 2.36 | 0.49 | 3.68 | 0.95 |
| 19 | 1.49 | 17.42 | 2.12 | 2.14 | 3.30 | 2.04 | 0.48 | 3.28 | 1.06 |
| 20 | 1.52 | 18.55 | 1.87 | 1.98 | 2.90 | 2.16 | 0.48 | 3.56 | 0.84 |
| 21 | 1.64 | 18.74 | 2.10 | 1.89 | 2.82 | 2.04 | 0.53 | 3.56 | 1.02 |
| 22 | 1.40 | 14.79 | 2.21 | 2.07 | 2.79 | 2.15 | 0.52 | 3.49 | 1.04 |
| 23 | 1.78 | 18.86 | 2.00 | 2.08 | 3.14 | 2.60 | 0.50 | 3.30 | 0.80 |
| 24 | 1.93 | 15.62 | 2.26 | 2.21 | 2.81 | 2.18 | 0.44 | 4.16 | 0.92 |
| 25 | 1.53 | 18.56 | 2.14 | 2.00 | 3.16 | 2.22 | 0.51 | 3.73 | 1.07 |
| **Sum** | **42.20** | **415.39** | **54.03** | **53.92** | **62.02** | **56.00** | **12.25** | **89.79** | **24.10** |

**Ouput dari Program SAS IML**

**TABLE 3-2** Himpunan Bagian Variabel Prediktor Terbaik Untuk Semua Kriteria Pemilihan Model

| Kriteria Pemilihan Model | | Himpunan Bagian Prediktor |
|---|---|---|
| Stepwise Regression | FORWARD | 1246 |
| | BACKWARD | 1234 |
| | STEPWISE | 1246 |
| Semua Kemungkinan Regresi (*All-Possible-Regression*) | MSE | 1246 |
| | ADJRSQ | 24 |
| | AIC | 126 |
| | AICC | 126 |
| | HQ | 126 |
| | HQC | 126 |
| | BIC | 123456 |
| | Cp | 126 |
| | | 1236 |
| | | 1246 |
| | | 1256 |
| | | 12346 |
| | | 12356 |
| | | 12456 |

Tabel 2 menunjukkan himpunan bagian terbaik dari variabel prediktor yang diperoleh dari prosedur regresi bertahap (*stepwise*) dan semua kemungkinan regresi (*all-possible-regression*). Pada tabel tersebut ditunjukkan bahwa melalui prosedur *forward stepwise regression* dan semua kemungkinan regresi dengan kritera *means square error* diperoleh model dengan variabel prediktor $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$, dan $\mathbf{x}_6$, sebagai himpunan bagian terbaik variabel prediktor dalam memprediksi variabel respon $\boldsymbol{y}$. Melalui prosedur mundur (*Backward elimination method*) diperoleh model dengan himpunan variabel prediktor yang terpilih adalah $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, dan $\mathbf{x}_6$, sebagai himpunan bagian terbaik variabel prediktor guna memprediksi $\boldsymbol{y}$. Kriteria adjusted $R^2$ memilih model dengan variabel prediktor $x_2$ dan $x_4$ . Kriteria informasi Akaike's (AIC), (AICc), Hannan dan Quinn (HQ), (HQc) memilih model dengan himpunan variabel prediktor $\mathbf{x}_1$, $\mathbf{x}_2$, dan $\mathbf{x}_6$. Sedangkan kriteria Schwarz's Bayesian memilih model dengan semua vairabel prediktor **x**.

Dalam hal lain, kriteria Mallow's $\mathbf{C}_p$, telah memberikan sebuah susunan pemilihan model terbaik dengan bantuan komputer dapat dilakukan perhitungan dari $\left|\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}\right|$ untuk $(\frac{n-p}{n-k})^q$ dan memilih model dengan $\left|\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}\right| \leq (\frac{n-p}{n-k})^q$ sebagai satu model terbaik. Dalam penggunaan kriteria $\mathbf{C}_p$, kita mencoba mengidentifikan himpunan bagian dari vairabel prediktor **x** yang memenuhi kondisi :

(1) Nilai| $\mathbf{C}_p$ | relatif kecil

(2) Nilai $\mathbf{C}_p$ mendekati $p$**I** ( $\left|\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}\right| \leq (\frac{n-p}{n-k})^q$ ).

Table 3-3 menunjukkan nilai $(\frac{n-p}{n-k})^q$ untuk setiap *p*

**TABLE 3-3** Batas atas untuk setiap *p*

| *P* | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| $(\frac{n-p}{n-k})^q$ | 2.086248 | 1.825789 | 1.587963 | 1.371742 | 1.176097 | 1 |

Dari semua prosedur pemilihan himpunan bagian variabel prediktor terbaik, untuk prosedur stepwise yaitu *Prosedur Forward*, *Forward Stepwise Regression,* dan prosedur semua kemungkinan regresi yaitu dengan kriteria *MSE* dan $C_P$ menghasilkan hipunan bagian variabel prediktor terbaik untuk memprediksi variabel respon ***Y*** adalah $X_1$, $X_2$, $X_4$ dan $X_6$. Model regresi multivariat multiple terbaik untuk memprediksi kandungan kimia dalam daun tembakau adalah model yang melibatkan variabel prediktor $X_1$, $X_2$, $X_4$ dan $X_6$.

## KESIMPULAN

Seleksi variabel dalam analisis regresi multivariate multiple sebaiknya dilakukan secara simultan dengan alasan akan lebih cepat dalam proses perhitungan. Dari contoh aplikasi, diketahui bahwa *Prosedur Forward*, *Forward Stepwise Regression,* dan prosedur semua kemungkinan regresi yaitu dengan kriteria *MSE* dan $C_P$ dapat dijadikan rujukan dalam menentukan variabel independen mana yang harus dimasukkan ke dalam model.

Penulis menyarankan bahwa dalam pembentukan model regresi khususnya untuk tujuan prediksi sangat penting melakukan investigasi dan mempelajari prilaku

dari vairabel prediktor. Dalam pembentukan model tidak disarankan hanya didasarkan pada kriteria pemilihan model terbaik yang ada karena semua kriteria tersebut tidak ada yang sempurna dan sangat bergantung pada berbagai faktor. Sebuah simulasi yang dibuat oleh Bedrick dan Tsai (1994) bahwa ukuran sampel, jumlah vairabel independen, dan korelasi antara variabel respon ***y*** memiliki peran penting dalam menentukan kriteria pemilihan model mana yang harus digunakan. Sehingga setiap peneliti dalam melakukan pemilihan model regresi terbaik membutuhkan informasi lebih mendalam mengenai vairabel independen yang didasarkan pada teori yang relevan, tidak adanya hubungan yang kuat antara variabel independen dan memiliki korelasi yang kuat dengan semua variabel respon ***y*** . Peneliti membutuhkan penggunaan lebih dari satu kriteria dalam mengevaluasi himpunan variabel independen yang layak dimasukkan dalam model. Tahap terakhir adalah peneliti harus melakukan evaluasi pada model terbaik menggunakan beberapa prosedur diagnosis model regresi sehingga diperoleh model regresi terbaik untuk tujuan prediksi.

## DAFTAR PUSTAKA


[1] Al-Subaihi, Ali A. (2002), "Variable Selection in Multivariable Regression Using SAS/IML", *Journal of Statistics Software Volume 7 Issue 12*.

[2] Akaike, H. (1973), "Information Theory and an Extension of The Maximum Likelihood Principle", In B.N. Petrov and F. Csaki ed., 2nd *International Symposium on Information Theory*, pp. 267-281, Akademia Kiado, Budapest.

[3] Al-Subaihi, Ali A. (2002), "Univariate Variabel Selection Criteria Available In SAS or SPSS", paper presented at the American Statistical Association annual meeting- August 11-15, 2002, New York, NY.

[4] Anderson, R. L. and Bancroft, T. A. (1952), *Statistical Theory in Research*, McGraw-Hill Book Company, Inc., New York, NY.

[5] Barrett, B. E. and Gray, J. B. (1994), "A Computational Framework for Variabel Selection in Multivariate Regression", *Statistics and Computing*, **4**, 203-212.

[6] Bilodeau, M. and Brenner, D. (1999), *Theory of Multivariate Statistics*, Springer-Verlag New York, Inc., New York.

[7] Breiman, L. and Friedman, J. H. (1997), "Predicting Multivariate Responses in Multiple Linear Regression", *Journal of the Royal Statistical Society*, **59** (No. 1), 3-54.

[8] Fujikoshi, Y.; and Satoh, K. (1997), " Modified AIC and Cp in Multivariate Linear Regression", *Biometrika*, **84** (3), 707-716.

[9] Hannan, E. J. and Quinn, B. G. (1979), "The Determination of The Order of an Autoregression", *Journal of the Royal Statistical Society*, B **41**, 190-195.

[10] Mallows, C. L., (1973), "Some Comments on Cp", *Technometrics*, **15** (4), 661-675.

[11] McQuarrie A. D., and Tsai, C. (1998), "*Regression and Time Series Model Selection*", World Scientific Publishing Co. Pte. Ltd., River Edge, NJ.

[12] Miller, A. J. (1990), *Subset Selection in Regression*, Chapman and Hall, New York, NY.

[13] Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996), "*Applied Linear Statistical Models*", McGraw-Hill Companies, Inc., NY.

[14] Rencher, A. C. (1995), "*Methods of Multivariate Analysis*", John Wiley & Sons Inc., New York, New York.

[15] Rencher, A. C. (1998), " *Multivariate Statistical Inference and Applications*", John Wiley & Sons Inc., New York, New York.

[16] *SAS/STAT User's Guide*, Version 6, 4th Edition, SAS Institute Inc., Cary, NC (1990).

[17] Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461 -464.

[18] Sparks, R. S.; Coutsourides, D.; and Troskie, L. (1983), " The Multivariate Cp", *Commun. Statistis. –Theor. Meth.,* **12** (15), 1775-1793.

[19] Sparks, R. S.; Zucchini, W.; and Coutsourides, D. (1985), " On Variabel Selection in Multivariate Regression" , *Commun. Statistis. –Theor. Meth.,* **14** (7), 1569-1587.