LAPORAN PENELITIAN

PEMILIHAN MODEL REGRESI LINIER TERBAIK DENGAN METODE VALIDASI-SILANG

Oleh:

Drs. Tarno, M.Si.
Drs. Rukun Santoso, M.Si.

DIBIAYAI OLEH BAGIAN PROYEK PENINGKATAN KUALITAS SUMBERDAYA MANUSIA,
DIREKTORAT JENDERAL PENDIDIKAN TINGGI, DEPARTEMEN PENDIDIKAN NASIONAL
TAHUN ANGGARAN 2002

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS DIPONEGORO SEMARANG
OKTOBER 2002

SISTEMATIKA LAPORAN AKHIR
HASIL PENELITIAN DOSEN MUDA

PEMILIHAN MODEL REGRESI LINIER TERBAIK DENGAN METODE VALIDASI-SILANG

Oleh : Tarno, Rukun Santoso
Jurusan Matematika FMIPA UNDIP

RINGKASAN

Salah satu model statistik yang sangat berguna dalam berbagai bidang aplikasi adalah model linier umum :

$$y_i = x_i'\beta + \varepsilon_i, i = 1,2,...,n \qquad (1)$$

dengan y_i adalah respon ke-i , x_i : p-vektor variabel prediktor yang berkaitan dengan y_i, β: p-vektor parameter yang tidak diketahui dan ε_i : sesatan random. Masalah regresi linier dapat diformulasikan sebagai kasus khusus dari model (1) tersebut. Dalam masalah regresi semua variabel prediktor yang sebelumnya diduga berpengaruh terhadap respon, pada prakteknya belum tentu berpengaruh secara signifikan terhadap respon. Dengan demikian perlu dilakukan pemilihan variable prediktor yang berpengaruh secara signifikan terhadap variable respon. Berdasarkan hal tersebut, yang menjadi permasalahan dalam penelitian ini adalah bagaimana prosedur pemilihan model regresi terbaik, yaitu model yang memiliki kesalahan prediksi paling kecil dan melibatkan variable predictor sesedikit mungkin. Dan tujuan penelitian ini adalah menentukan model regresi terbaik dari semua model yang mungkin dengan metode yang berdasarkan resampling yaitu metode validasi silang.

Prosedur pemilihan model terbaik dilakukan dengan menentukan estimasi sesatan prediksi yang minimal atas semua model yang mungkin yaitu ada sebanyak 2^p-1 model dengan p: banyaknya prediktor. Prosedur pemilihan model dengan menggunakan validasi silang lepas-1 dikembangkan dari metode jackknife lepas-1. Metode validasi silang lepas-1 membagi data menjadi dua bagian yaitu data konstruksi terdiri dari data asli dengan melepaskan (mengeluarkan) datum yang ke-i (i=1,2,3,...,n) dan data validasi {i}. Dan metode validasi silang lepas-1 secara umum dikembangkan menjadi metode validasi silang lepas-d dengan d lebih kecil dari

ukuran sampel n. Untuk validasi silang lepas-d, data konstruksi terdiri dari (n-d) dan data validasi terdiri dari d. Sesatan prediksi diperoleh dari pengurangan **y** (data validasi) dengan estimasi **y** (dari data konstruksi). Dengan mengambil $\frac{d}{n} \to 1$ untuk $(n-d) \to \infty$ metode validasi silang lepas-d konsisten. Hasil simulasi yang merupakan implementasi secara praktis dengan menggunakan sistem SPLUS-2000 diberikan juga pada penelitian ini. Dari hasil simulasi menunjukkan bahwa pemilihan variabel dengan menggunakan validasi silang lepas-d cenderung memilih model dengan ukuran lebih kecil dibandingkan dengan metode validasi silang lepas-1.

SELECTION OF BEST LINEAR REGRESSSION MODEL BY CROSS-VALIDATION METHODE

By : Tarno, Rukun Santoso
Jurusan Matematika FMIPA UNDIP

SUMMARY

One of very useful statistical model in application is generalized linear model :

$$y_i = x_i'\beta + \varepsilon_i \,, i = 1,2,\ldots, n \qquad (1)$$

where y_i is i-th response , x_i : p-vector predictor variable related with y_i, β: p-vector unknown parameter and ε_i: error random. Linear regression problem can be formulated as special case of equation (1). Predictor variables factually may not be influence to dependent variable **y**. Then we need select predictor variable **x** that significantly influence to response **y**. Based on the argument the problem of this research is how to select the best linear model that have mean square error minimum and involve smallest variable. And the goal of this research is select the best linear regression of all possible models by cross-validation methode that based on resampling data .

Procedure of model selection is being done by minimized error prediction of all possible model, there are 2^p-1 where p : number of predictor. Procedure of model selection by cross-validation is extended from jackknife delete-1. Cross-validation delete-1 divide data to two parts : construction data and validation data. Cross-validation delete-1 is extended to cross-validation delete-d where d is less than the number of sample n. Error prediction is differences between response y dan their estimation. Cross-validation delete-d is consistent as $\frac{d}{n} \to 1$ and $(n-d) \to \infty$. Results from a simulation with SPLUS system is also presented. From simulation results, cross-validation delete-d selects a model that have smaller predictor than cross-validation delete-1.

PRAKATA

Puji syukur kehadlirat Allah SWT atas limpahan rahmat dan taufiq-Nya, seehingga peneliti dapat menyelesaikan penyusunan Laporan Akhir Penelitian yang berjudul " **Pemilihan Model Regresi Linier Terbaik Dengan Metode Validasi Silang**". Peneliti menyadari bahwa dalam rangka pelaksanaan penelitian hingga penyusunan Laporan ini beberapa pihak telah membantu peneliti baik berupa material maupun spiritual. Untuk itu pada kesempatan ini peneliti sampaikan terima kasih kepada:

1. Prof. Dr. dr. Ign Riwanto, SP.BD selaku Ketua Lembaga Penelitian UNDIP Semarang
2. Prof. Drs. Mustafid, M.Eng, Ph.D selaku Dekan Fakultas MIPA UNDIP
3. Direktorat Jenderal Pendidikan Tinggi, Departemen Pendidikan Nasional yang telah memberikan dana penelitian ini
4. Drs. Bayu Surarso, M.Sc, Ph.D selaku Ketua Jurusan Matematika FMIPA UNDIP
5. Pengelola Laboratorium Kumputer Jurusan Matematika FMIPA UNDIP
6. Rekan-rekan di kelompok Laboratorium Statistika serta pihak-pihak yang tidak dapat peneliti sebut satu per satu.

Selanjutnya peneliti mengharapkan masukan serta saran dari para pembaca demi kesempurnaan penelitian untuk masa yang akan datang. Dan akhirnya peneliti berharap semoga Laporan penelitian ini dapat bermanfaat bagi para pembaca.

Peneliti

DAFTAR TABEL :

LAMPIRAN

PEMILIHAN MODEL REGRESI LINIER TERBAIK DENGAN METODE VALIDASI SILANG

I. PENDAHULUAN

Dalam berbagai bidang aplikasi seringkali digunakan model-model statistik . Salah satu model yang sangat berguna adalah model linier umum :

$$y_i = x_i'\beta + \varepsilon_i, i = 1,2,...,n \quad (1)$$

dengan y_i adalah respon ke-i , x_i : p-vektor variabel prediktor yang berkaitan dengan y_i, β: p-vektor parameter yang tidak diketahui dan ε_i : sesatan random. Masalah regresi linier dapat diformulasikan sebagai kasus khusus dari model (1) tersebut.

Apabila x_i dalam model tersebut tertentu (deterministik), diasumsikan bahwa ε_i independen dengan mean 0 dan variansi σ^2. Sedangkan apabila x_i tersebut random, maka model (1) dikatakan sebagai model korelasi. Dalam suatu model korelasi, (y_i, x_i') diasumsikan independen dan berdistribusi identik dengan momen kedua berhingga dan $E(y_i \mid x_i) = x_i'\beta, \sigma_i^2$ menyatakan variansi bersyarat dari y_i diberikan x_i .

Parameter β dalam model tersebut dikenal sebagai parameter regresi. Dalam model regresi dengan asumsi bahwa unsur sesatan berdistribusi normal dengan mean 0 dan variansi konstan, dengan menggunakan metode kuadrat terkecil, estimasi parameter dapat ditentukan dengan mendefinisikan :

$$\hat{\beta} = (X'X)^{-1}X'y$$

dengan $X' = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)'$.

Jika estimasi parameter telah diperoleh berarti telah diperoleh estimasi model untuk respon **y** yang tergantung pada prediktor **x**, yang dapat digunakan untuk melakukan prediksi untuk nilai y yang akan datang berdasarkan prediktor **x**. Beberapa komponen dari **x** mungkin tidak menghasilkan prediksi yang akurat karena tidak berpengaruh secara signifikan terhadap respon y, oleh karena itu perlu dilakukan

pemilihan model terbaik (dalam hal ini sama dengan pemilihan variabel prediktor), yang memiliki kemampuan prediksi yang paling akurat. Menurut Shao (1993), pemilihan variabel dalam model regresi linier ini dapat dilakukan dengan beberapa metode antara lain : Akaike Information Criterion (AIC) , Cp (Mallows), Bayesian Information Criterion (BIC) dan metode yang berdasarkan resampling data pengamatan yaitu Cross-Validation (CV) dan bootstrap.

Menurut Hjorth (1994), kriteria AIC secara eksak atau pendekatan merupakan estimator tak bias untuk model dalam kategori II (semua parameter tak nol), tetapi jika digunakan untuk memilih model dalam kategori I (ada komponen parameter yang sama dengan nol) kriteria ini kadang-kadang tidak konsisten (bias). Sedangkan untuk kriteria BIC secara asimptotis tidak konsisten untuk data pengamatan berukuran besar (banyaknya pengamatan antara 50-500) dan lebih baik apabila diterapkan pada model runtun waktu. Berdasarkan latar belakang tersebut, untuk mengatasi kekurangan diatas dalam penelitian ini dibahas metode pemilihan model regresi terbaik berdasarkan resampling data pengamatan yaitu metode Validasi –Silang / Cross-Validation (CV)

II. TINJAUAN PUSTAKA

Menurut Shao (1993), Allen dan Stone mengusulkan metode pemilihan model linier yang dikenal dengan metode Croos-Validation (CV) yang secara esensial merupakan suatu metode yang berdasarkan metode jackknife terhapus-1 (jackknife delete-1). Karena metode jackknife terhapus-1 dianggap terlalu konservatif maka metode tersebut diperbaiki dengan metode jackknife terhapus-d (jackknife delete-d) (Shao dan Tu, 1995), (Tarno, 2000). Sedangkan metode Cp, BIC dan AIC untuk pemilihan variabel telah dikembangkan oleh Hjorth (1994). Metode Cp, AIC, BIC maupun CV memilih model dengan meminimalkan jumlah kuadrat sesatan (mse) atas semua model yang mungkin. Untuk menentukan mse tersebut, metode CV membagi data dalam dua bagian yaitu data validasi dan data konstruksi, sedangkan untuk Cp, BIC dan AIC berdasarkan kumpulan data secara keseluruhan .