e-Jurnal Matematika, Vol. 1, No. 1, Agustus 2012, 75-80

PERBANDINGAN ANALISIS *LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR* DAN *PARTIAL LEAST SQUARES* (Studi Kasus: Data Microarray)

KADEK DWI FARMANI[1], **I PUTU EKA NILA KENCANA**[2], **KOMANG GDE SUKARSA**[3]

[1,2,3,] Jurusan Matematika, Fakultas MIPA, Universitas Udayana
e-mail: [1]dwifarmani_magical@yahoo.com, [2]i.putu.enk@gmail.com, [3]sukarsakomang@yahoo.com

Abstract

Linear regression analysis is one of the parametric statistical methods which utilize the relationship between two or more quantitative variables. In linear regression analysis, there are several assumptions that must be met that is normal distribution of errors, there is no correlation between the error and error variance is constant and homogent. There are some constraints that caused the assumption can not be met, for example, the correlation between independent variables (multicollinearity), constraints on the number of data and independent variables are obtained. When the number of samples obtained less than the number of independent variables, then the data is called the microarray data. Least Absolute shrinkage and Selection Operator (LASSO) and Partial Least Squares (PLS) is a statistical method that can be used to overcome the microarray, overfitting, and multicollinearity. From the above description, it is necessary to study with the intention of comparing LASSO and PLS method. This study uses coronary heart and stroke patients data which is a microarray data and contain multicollinearity. With these two characteristics of the data that most have a weak correlation between independent variables, LASSO method produces a better model than PLS seen from the large RMSEP.

Keywords*: microarray, overfitting,* RMSEP, LASSO, PLS.

1. Pendahuluan

Analisis regresi berganda digunakan untuk mengetahui hubungan yang melibatkan lebih dari satu variabel bebas dan satu variabel tak bebas. Dalam regresi berganda harus dipenuhi beberapa asumsi, yaitu galat berdistribusi normal, antara galat-galat tidak berkorelasi atau bersifat saling bebas, dan ragam suku galat konstan dan homogen. Skala pengukuran dari variabel bebas dan variabel tak bebas adalah metrik (interval dan ratio). Tidak terpenuhinya asumsi tersebut dapat diakibatkan oleh korelasi antar variabel bebas (multikolinearitas), kendala pada jumlah data, dan jumlah variabel bebas yang diperoleh.

Ketika data yang diperoleh terdiri dari variabel bebas yang lebih banyak daripada banyaknya data, maka data semacam ini disebut data *microarray*. Kendala yang timbul

[1] Mahasiswa Jurusan Matematika FMIPA Universitas Udayana
[2,3] Staf Pengajar Jurusan Matematika FMIPA Universitas Udayana

ketika membangun model regresi dengan data *microarray* adalah terjadinya *overfitting*. Menurut Izenman (2008: 13), *overfitting* adalah suatu kejadian di mana jumlah parameter yang masuk ke dalam model terlalu besar dibandingkan dengan ukuran data yang digunakan untuk membangun model (*learning set*). Model tersebut menghasilkan galat yang sangat kecil untuk data *learning set*, namun galat yang besar untuk data validasi.

Least Absolute Shrinkage and Selection Operator (LASSO) dan *Partial Least Squares* (PLS) adalah metode statistika yang dapat digunakan untuk mengatasi *microarray*, *overfitting*, dan multikolinearitas. Adapun penelitian terdahulu yang menggunakan dua metode tersebut pada data *microarray* yaitu prediksi waktu tahan hidup pasien penyakit jantung koroner dengan metode PLS yang dilakukan Kusuma (2011) dan prediksi waktu tahan hidup penderita stroke dengan metode LASSO yang dilakukan Wulandari (2011). Berdasarkan uraian di atas perlu dilakukan penelitian dengan maksud membandingkan metode LASSO dan PLS dalam menganalisis data penderita jantung koroner dan stroke.

Tujuan dari penelitian ini adalah: (1) Mengetahui model persamaan regresi LASSO dan PLS pada analisis data waktu tahan hidup pasien jantung koroner; (2) Mengetahui model persamaan regresi LASSO dan PLS pada analisis data waktu tahan hidup pasien stroke; (3) Mengetahui perbedaan antara metode LASSO dan PLS dalam menganalisis data pasien jantung koroner dan stroke.

LASSO adalah salah satu teknik regresi pengkerutan variabel bebas. LASSO dapat digunakan untuk mengatasi masalah pada data *microarray*. LASSO mengkerutkan koefisien (parameter β) yang berkorelasi menjadi nol atau mendekati nol. Sehingga menghasilkan penduga dengan varian yang lebih kecil dan model akhir yang lebih representatif (Tibshirani, 1996).

Menurut Yongdai (2004) dalam Wulandari (2011), misalkan $(X_i, Y_i), \dots, (X_n, Y_n)$ adalah *n* pasangan variabel bebas atau variabel tak bebas dengan $Y_i \epsilon\, Y$ dan $X_i \epsilon\, X$, di mana Y dan X adalah input dan output dengan Y variabel tak bebas, *p* variabel bebas X_1, X_2, …, X_p, dan *n* banyaknya pengamatan. Estimasi parameter pada LASSO adalah sebagai berikut:

$$\begin{aligned} R(\beta_0, \beta_j) &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \sum_{j=1}^{p} X_{ij}\hat{\beta}_j\right)^2 \\ (\beta)_{l_1} &= \sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\left|\hat{\beta}_j\right| \\ &= (Y - X\beta)^T(Y - X\beta) + \lambda\left\|\hat{\beta}\right\|_1 \end{aligned} \quad (1)$$

λ adalah parameter yang mengontrol koefisien LASSO yang diatur dengan batasan $D = \sum_{j=1}^{p}\left|\hat{\beta}_j\right| \leq t$. Sehingga model LASSO dapat dinyatakan sebagai berikut:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip} + D \quad (2)$$

Partial Least Squares (PLS) dikembangkan pada tahun 1960-an oleh Herman Wold seorang ahli ekonometrika. Faktor pertama dari model PLS dipilih untuk memaksimalkan nilai kovarian dengan variabel tak bebas dan selanjutnya digunakan sebagai regressor dalam model regresi dengan metode kuadrat terkecil biasa. Faktor selanjutnya dipilih untuk memaksimumkan nilai kovarian dengan sisaan penduga dari metode kuadrat terkecil biasa.

Banyaknya variabel laten yang digunakan pada akhir model PLS dapat ditentukan dengan *validasi silang*. Dalam studi ini, pengamatan dibagi menjadi dua set data, satu digunakan untuk kecocokan model dan lainnya untuk validasi model dengan perbandingan prediksi dan nilai kebenarannya (Mevik, 2007).

Di sini hanya beberapa komponen utama yang digunakan, berapa banyak komponen yang optimal harus ditentukan biasanya dengan validasi silang.

Koefisien determinasi (r^2) merupakan suatu nilai yang digunakan untuk

mengukur sejauh mana model yang diperoleh mampu menjelaskan keadaan data yang sebenarnya atau seberapa besar variabel bebas *X* mampu menjelaskan varian dari variabel tak bebas *Y*. Koefisien determinasi berkisar antara nol sampai dengan satu ($0 \leq r^2 \leq 1$) (Neter,1997:91).

Untuk melihat keakuratan prediksi dan kebaikan suatu model digunakan metode *Root Mean Squares Error of Prediction* (RMSEP). Pada metode PLS RMSEP digunakan untuk menentukan banyaknya komponen pada model prediksi, yaitu dengan membagi dua data yaitu satu sebagai model kalibrasi dan satu untuk model validasi.

2. Metode Penelitian

Pada jantung koroner terdapat satu variabel tak bebas (*Y)* dan sebelas variabel bebas (*X)* yang digunakan, yaitu:Y=waktu tahan hidup; X_1=Keturunan; X_2= Jenis kelamin; X_3=Usia; X_4=Stres; X_5=Kadar gula darah; X_6=Tekanan darah; X_7=Jumlah batang rokok/hari; X_8=Kolesterol; X_9=Obesitas; X_{10}=Olahraga; X_{11}=Waktu pertama kali sakit sampai diteliti.

Variabel stroke terdiri dari satu variabel tak bebas (*Y)* dan delapan variabel bebas (*X)* yaitu Y=waktu tahan hidup ; X_1= Jenis kelamin; X_2=Kondisi pertama kali diperiksa; X_3=Usia; X_4=Berat badan; X_5=Kebiasaan mengkonsumsi alkohol; X_6=Jumlah batang rokok/hari; X_7=Jumlah saraf yang terganggu; X_8=Waktu pertama kali sakit sampai diteliti.

Proses analisis data pada penelitian ini menggunakan *software* R 2.14.1 dan SPSS. Kedua data dianalisis dengan metode regresi LASSO dan PLS. Langkah awal yang dilakukan adalah memasukkan data pengamatan dalam bentuk matriks pada program R di mana baris menunjukkan *n* banyaknya data dan kolom yang menunjukkan *p* variabel bebas. Melakukan uji multikolinearitas dengan memeriksa matriks korelasi dan signifikansi dengan SPSS. Jika data terbukti mengandung multikolinearitas, maka analisis dapat dilanjutkan pada LASSO dan PLS. Setelah kedua data dianalisis dengan LASSO dan PLS, selanjutnya dapat ditentukan nilai r^2 dan RMSEP masing-masing model. RMSEP digunakan untuk membandingkan kinerja kedua metode.

3. Hasil dan Pembahasan

3.1 Matriks Korelasi

Matriks korelasi berguna untuk memeriksa adanya multikolinearitas pada data. Dari matriks korelasi dapat diketahui variabel bebas yang mengalami multikolinearitas. Pada data pasien jantung koroner terdapat tujuh pasang variabel bebas yang memiliki korelasi signifikan. Sedangkan pada data pasien stroke terdapat dua pasang variabel bebas yang saling berkorelasi signifikan.

Dari pemaparan mengenai koefisien korelasi data pasien jantung koroner dan stroke, terbukti bahwa data mengandung multkolinearitas. Hal ini menunjukkan bahwa LASSO dan PLS dapat diterapkan pada kedua data.

3.2 Analisis Prediksi Waktu Tahan Hidup Pasien Jantung Koroner dengan Metode LASSO

LASSO memiliki batasan yaitu $D = \sum_{j=1}^{p} |\hat{\beta}_j| \leq t$, dengan $t \geq 0$ yang merupakan parameter *tuning* pada LASSO. Parameter *t* yang digunakan merupakan nilai minimum validasi silang. Nilai *t* dan *D* yang diperoleh dari proses LASSO adalah 1,251182 dan

0,8316876. Batasan LASSO terpenuhi karena nilai *t* dan *D* yang diperoleh telah memenuhi syarat $t \geq 0$ dan $D \leq t$. Diperoleh variabel bebas yang berpengaruh signifikan terhadap waktu tahan hidup pasien jantung koroner adalah jenis kelamin, stres, tekanan darah, jumlah batang rokok, kolesterol, obesitas, olahraga, dan waktu pertama kali sakit sampai diteliti. Sehingga model LASSO pasien jantung koroner adalah

$$\hat{Y} = 1{,}118 + 0{,}024247644X_2 - 0{,}025178664X_4 - 0{,}0002959958X_6 + 0{,}0081139062X_7 + 0{,}0020701941X_8 + 0{,}0053380618X_9 - 0{,}018205108X_{10} + 0{,}048145312X_{11} + 0{,}8316876 \quad (3)$$

3.3 Analisis Prediksi Waktu Tahan Hidup Pasien Jantung Koroner dengan Metode PLS

Pemilihan model terbaik dilakukan dengan memperhatikan pola RMSEP minimum dari pembentukan model dan validasi serta nilai koefisien determinasi sebagai indikator bahwa model yang dipilih mampu mengatasi *overfitting*. Diperoleh model terbaik untuk memprediksi waktu tahan hidup pasien jantung koroner adalah model dengan menggunakan data ke-5, 6, dan 9 sebagai data validasi. Berdasarkan RMSEP minimum dari validasi silang diperoleh model dengan satu komponen. Model PLS yang diperoleh untuk pasien jantung koroner adalah

$\hat{Y} = 9{,}6 + 0{,}0824w_1$ dengan w_1 adalah komponen (4)

3.4 Analisis Prediksi Waktu Tahan Hidup Pasien Stroke dengan Metode LASSO

Batasan LASSO untuk pasien stroke yaitu $D = \sum_{j=1}^{p} |\hat{\beta}_j| \leq t$, dengan $t \geq 0$ yang merupakan parameter *tuning* pada LASSO. Nilai *D* yang diperoleh sebesar 0,863862 dan nilai *t* sebesar 6,92. Karena nilai $t \geq 0$ dan $D \leq t$ maka batasan LASSO terpenuhi. Diperoleh variabel bebas yang berpengaruh signifikan terhadap waktu tahan hidup pasien stroke adalah kondisi pada saat pertama kali diperiksa, usia, berat badan, jumlah batang rokok, dan jumlah saraf yang mengalami gangguan. Model LASSO untuk pasien stroke adalah

$$\hat{Y} = 2{,}645714 - 0{,}006231345X_2 - 0{,}0044966X_3 - 0{,}00113078X_4 + 0{,}000490742X_6 - 0.1314523X_7 + 0{,}8636862 \quad (5)$$

3.5 Analisis Prediksi Waktu Tahan Hidup Pasien Stroke dengan Metode PLS

Model terbaik untuk memprediksi waktu tahan hidup pasien stroke adalah model dengan data ke-5, 6, dan 7 sebagai data validasi. Berdasarkan RMSEP minimum dari validasi silang diperoleh model dengan satu komponen. Model PLS untuk pasien stroke adalah

$\hat{Y} = 2136 + 0{,}9674w_1$ dengan w_1 sebagai komponen (6)

3.6. Perbandingan LASSO dan PLS

Berdasarkan uraian sebelumnya mengenai analisis waktu tahan hidup pasien jantung koroner dan stroke dengan metode LASSO dan PLS diperoleh beberapa hal yaitu:

1. Prosedur analisis
   Pembangunan model pada metode PLS menggunakan sebagian data. Sedangkan pada LASSO pembangunan model memanfaatkan seluruh data.
2. Metode PLS dan LASSO terbukti memiliki kemampuan sama baiknya dalam mengatasi *overfitting* pada data *microarray*. Hal ini dapat dilihat dari nilai r^2 untuk

kedua metode yang masih berada pada rentang nol dan satu. Tabel 1 menyajikan r^2 kedua metode untuk data pasien jantung koroner dan stroke.

Tabel 1. Nilai r^2 Metode LASSO dan PLS

Metode	r^2 Jantung Koroner	r^2 Stroke
LASSO	0,77	0,8
PLS	0,51	0,79

Sumber: Data diolah (2012)

3. RMSEP
   Dalam memprediksi waktu tahan hidup pasien jantung koroner dan stroke kedua metode menghasilkan RMSEP yang berbeda. Tabel 2 menunjukkan RMSEP masing-masing metode.

Tabel 2. RMSEP Metode LASSO dan PLS

Metode	RMSEP Jantung Koroner	RMSEP Stroke
LASSO	0,408	29,9
PLS	1,712	45,8

Sumber: Data diolah (2012)

4. Kesimpulan

Berdasarkan hasil analisis data waktu tahan hidup pasien jantung koroner dan stroke dengan metode PLS dan LASSO pada Bab IV, diperoleh beberapa hal sebagai berikut:

1. Metode LASSO dan PLS mampu mengatasi masalah *overfitting* pada kasus data *microarray*.
2. Data pasien jantung koroner dan stroke memiliki kemiripan karateristik dilihat dari tingkat korelasi antar variabel bebas. Sebagian besar korelasi yang dimiliki adalah korelasi sangat rendah yaitu pada selang 0,00 – 0,20.

Dengan karakteristik data tersebut, diperoleh nilai RMSEP LASSO lebih kecil dibandingkan RMSEP PLS.

Daftar Pustaka

Abdi, H. 2006. *Partial Least Squares Regression (PLSR)*. University of Texas.

Izenman, A.J. 2008. *Modern Multivariate Statistical Techniques (Regression, Classification, and Manifold Learning)*. USA: Springer.

Mevik, H. 2007. "The PLS Package:Principla Component and Partial Least Square Regression in R". *Journal of Statistical Software*, Januari vol 18(2),pp.1-24. Norwegian. Available: http://www.jstatsoft.org/ (Accessed: 31 Oktober 2011)

Neter, Jhon, dkk. 1997. *Model Linear Terapan – Buku I: Analisis Regresi Linear*

Sederhana. Penerjemah: Bambang Sumantri. Bogor: Jurusan Statistika FMIPA-IPB.

Somnath, D and Susmita, D. 2007. "Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO". *Journal of Biometrics*, Maret vol 63(1), pp.259-271. USA. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2006.00660.x/citedby (Accessed: 31 Oktober 2011)

Tibshirani, R. 1996. *Regression Shrinkage and Selection via LASSO*. Universitas Toronto, Kanada: JSTOR.

Wulandari, P.R. 2011. *Penerapan Metode Regresi Least Absolute Shrinkage and Selection Operator (LASSO) terhadap Waktu Tahan Hidup Penderita Stroke*. Universitas Udayana: Jurusan Matematika Fakultas MIPA