OLS, LASSO dan PLS Pada data Mengandung Multikolinearitas

OLS, LASSO dan PLS Pada data Mengandung Multikolinearitas

Yuliani Setia Dewi
Jurusan Matematika FMIPA Universitas Jember

ABSTRACT

Correlation between predictor variables (multicollinearity) become a problem in regression analysis. There are some methods to solve the problem and each method has its own complexity. This research aims to explore performance of OLS, LASSO and PLS on data that have correlation between predictor variables. OLS establishes model by minimizing sum square of residual. LASSO minimizes sum square of residual subject to sum of absolute coefficient less than a constant and PLS combine principal component analysis and multiple linear regression. By analyzing simulation and real data using R program, results of this research are that for data with serious multicollinearity (there are high correlations between predictor variables), LASSO tend to have lower bias average than PLS in prediction of response variable. OLS method has the greatest variance of MSEP, that is mostly not consistent in estimating the Mean Square Error Prediction (MSEP). MSEP that is resulted by using PLS is less than that by using LASSO.

Keywords : OLS, LASSO, PLS, bias, MSEP, multicollinearity

PENDAHULUAN

Dalam analisis regresi, terkadang kita jumpai kondisi terdapatnya korelasi antar variabel bebas (variabel prediktor) atau yang biasa disebut dengan istilah multikolinearitas. Multikolinearitas menjadi suatu masalah dalam analisis regresi, terutama dalam regresi linear standar (OLS). Adanya multikolinearitas yang tinggi tidak memungkinkan melihat pengaruh variabel bebas terhadap variabel respon secara terpisah (Gujarati 1992).

Terdapat beberapa metode untuk mengatasi masalah multikolinearitas ini. Masing-masing metode mempunyai kekomplekan. Metode-metode yang diusulkan untuk mengatasi masalah multikolinearitas tersebut antara lain LASSO dan PLS.

PLS dapat digunakan untuk pemodelan yang mengandung sejumlah besar regressor/varilabel bebas. PLS pertama kali populer penerapannya dalam bidang kemometrik (Geladi 1992). Kemudian berkembang dan digunakan dalam bidang-bidang lain. Datta (2001) menggunakan PLS untuk konteks data *microarray*. Namun demikian, meskipun metode ini sudah lama diperkenalkan (tahun 1960an) sifat-sifat statistikanya relatif baru dipelajari (Frank & Friedman 1993). Metode regresi lain yang baru-baru ini populer *adalah Least Absolute Shrinkage & Selection Operator* (LASSO), diusulkan oleh Tibshirani pada tahun 1996. Efron (2004) memperkenalkan skema regresi yang lebih umum dengan nama *Least Angle Regression* (LAR) yang melibatkan LASSO sebagai salah satu di dalamnya. Datta *et al.* (2007) menggunakan metode PLS dan LASSO untuk memodelkan waktu daya tahan hidup pasien dalam konteks data *microarray* tersensor.

Regresi PLS merupakan teknik baru yang menjeneralisasi dan mengkombinasikan analisis komponen utama dan regresi berganda (Abdi 2006). PLS mereduksi dimensi variabel-variabel penjelas asal melalui pembentukan variabel-variabel laten dengan dimensi yang lebih kecil yang merupakan kombinasi linier dari variabel-variabel penjelas asal, kemudian metode kuadrat terkecil diaplikasikan pada variabel-variabel baru tersebut. Sedangkan LASSO merupakan teknik regresi yang melakukan pendugaan dengan meminimumkan jumlah kuadrat error dengan suatu kendala L_1, $\sum_{j=1}^{p}\left|\hat{\beta}_j\right| \le s$ dengan s adalah parameter tuning yang ditentukan oleh pengguna. Karena kendala tersebut, LASSO mengurangi sejumlah koefisien dengan membuatnya menjadi 0.

Berdasarkan hal-hal tersebut di atas, dengan adanya korelasi antara variabel-variabel bebas (multikolinearitas) dan kaitannya dengan metode-metode untuk mengatasi multikolinearitas, dengan menggunakan data

simulasi dan data riil, penelitian ini bertujuan untuk mengetahui *performance* metode "*Ordinary Least Square*" (OLS), "*Partial Least Squares"* (PLS) dan "*Least Absolute Shrinkage And Selection Operator"* (LASSO), ketepatan dan ketelitian metode-metode tersebut dalam menduga model.

METODE

Multikolinearitas

Multikolinearitas dikatakan ada ketika terdapat 2 atau lebih variabel bebas yang digunakan dalam regresi saling berkorelasi (Mendenhall & Sincich 1996). Salah satu cara untuk mendeteksi multikolinearitas dengan menggunkan *Variance Inflation Factor* (VIF).

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

VIF merupakan unsur-unsur diagonal utama matriks korelasi $\mathbf{C} = \left(\mathbf{X'X}\right)^{-1}$ dengan R_j^2 merupakan koefisien determinasi yang didapat dari variabel bebas X_j diregresikan terhadap p variabel bebas lain. Jika X_j tidak berkorelasi dengan variabel bebas lain, maka R_j^2 akan bernilai kecil dan VIF_j mendekati 1. Sebaliknya jika X_j mempunyai korelasi dengan variabel bebas lain, maka R_j^2 akan mendekati 1 dan VIF_j menjadi besar. Jika nilai VIF_j lebih dari 10,maka ini menunjukkan data mengalami masalah multikolinearitas (Montgomery & Peck 1991)

***Ordinary Least Square* (OLS)**

Misal hubungan antara variabel respon (Y) dan variabel bebas (X) dirumuskan dengan $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, dengan **Y** adalah vektor pengamatan berordo (nx1), **X** adalah variabel bebas berordo (nx(p+1)), **β** adalah koefisien regresi berordo ((p+1)x1) dan ε adalah vektor variabel random berordo (nx1). Pendugaan koefisien regresi dengan metode kuadrat terkecil dengan meminimumkan jumlah kuadrat sisa $(\sum_{i=1}^{n} Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2$.

Jika **X'X** tidak singular maka solusi dari penduga kuadrat terkecil dari β adalah $\hat{\beta} = \left(X^{'}X\right)^{-1} X^{'}Y$ dan $\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}^{'}\mathbf{X}\right)^{-1}$ dengan σ^2 diduga dari varian sisaan yang diperoleh dari *Mean Square Error* (MSE), $\mathrm{MSE} = \dfrac{Y^{'}Y - \hat{\beta}^{'}X^{'}Y}{n - p - 1}$

***Least Absolute Shrinkage and Selection Operator* (LASSO)**

LASSO diperkenalkan oleh Tibshirani (1996), merupakan teknik regresi penyusutan yang berguna dalam hal yang berurusan dengan sejumlah besar regressor (variabel prediktor). LASSO menduga model linier $\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$ melalui maksimisasi jumlah kuadrat sisaan $(\sum_{i=1}^{n} Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2$ dengan mengacu pada suatu kendala L_1, $\sum_{j=1}^{p} \left|\hat{\beta}_j\right| \le s$. Karena kendala tersebut, LASSO mengurangi sejumlah koefisien dengan membuatnya menjadi 0. Efron (2004) memperkenalkan skema regresi yang lebih umum dengan nama *Least Angle Regression* (LAR), yang melibatkan LASSO sebagai salah satu di dalamnya. Algoritmanya dimulai dengan model trivial dengan semua koefisien β dijadikan nol. Kemudian variabel-variabel yang paling berkorelasi dengan sisaan pada tahap sebelumnya ditambahkan. Jumlah variabel yang dilibatkan dalam model, berkaitan dengan pemilihan parameter tuning s. Ukuran numerik s disarankan berdasarkan hasil cross validasi (Datta *et al*. 2007).

***Partial Least Square* (PLS)**

Regresi PLS merupakan teknik yang menjeneralisasi dan mengkombinasikan sifat-sifat dari analisis komponen utama dan regresi berganda (Abdi 2006). Tujuannya adalah menduga atau menganalisa variabel-variabel tak bebas dari variabel-variabel bebas atau variabel prediktor. PLS diperkenalkan oleh Herman Wold dkk pada tahun 1960 an dan kemudian menjadi populer dalam bidang kemometrik dan juga digunakan dalam bidang-bidang lain. Regresi PLS terutama berguna untuk menduga variabel-variabel tak bebas dari sejumlah besar variabel-variabel bebas (variabel-variabel prediktor). Dalam hal seperti itu PLS mereduksi dimensi variabel-variabel penjelas asal dengan cara membentuk variabel-variabel laten yang merupakan kombinasi linier dari variabel-variabel penjelas asal dengan dimensi yang lebih kecil. Kemudian regresi OLS diaplikasikan terhadap variabel-variabel baru tersebut (Datta *et al.* 2007).

Misal Y merupakan variabel respon tunggal dan X_1, X_2,...,X_p merupakan p variabel prediktor. Pertama-tama vektor $\mathbf{X}_{.j} = (X_{1j},...,X_{nj})'$, $1 \le j \le p$ dan $\mathbf{Y} = (Y_1, ...,Y_n)$ dibakukan. Kemudian p variabel X_1, X_2,..., X_p direduksi menjadi faktor-faktor laten ($t^{(k)}$) ortogonal, ($t^{(k)}$) = (X_1,X_2,...,X_p) $c^{(k)}$ untuk k = 1,2,...,q dengan q adalah parameter tuning. Dalam praktek, cross validasi telah direkomendasikan untuk memilih q (Datta *et al*. 2007).

Parameter tuning q lebih kecil daripada p dan n sehingga Y dapat diregresikan terhadap $t^{(1)},..., t^{(q)}$ menggunakan regresi linier OLS. Variabel-variabel $t^{(k)}$ dibentuk secara rekursiv dari variabel-variabel X_j, $1 \le j \le p$, demikian juga dengan Y melalui suatu cara sebagai berikut : setelah diperoleh $t^{(1)},..., t^{(k-1)}$, dicari vektor konstanta c^k yang mempunyai panjang satu sehingga kombinasi linier $t^{(k)} = (X_1,X_2,...,X_p)\ c^{(k)}$ ortogonal untuk semua $t^{(i)}$ sebelumnya ($\{t^{(k)}\}' t^{(i)} = 0$, $i < k$) dan ($\{t^{(k)}\}'Y$ dengan Y mempunyai kovarian terbesar. Setelah q faktor laten ditemukan, Y diregresikan terhadap $t^{(1)},..., t^{(q)}$ dalam bentuk $\hat{Y} = \sum_{k=1}^{q} \hat{\gamma}_k t^{(k)}$. Misal $\hat{\beta} = C\hat{\gamma}$ dengan C adalah matriks dengan kolom $c^{(1)}, c^{(2)},..., c^{(q)}$ maka dapat diperoleh $\hat{Y} = \sum_{j=1}^{p} \hat{\beta}_j X_j$, hubungan yang dapat dengan mudah diekspresikan kembali ke dalam variabel X dan Y asal.

Langkah-langkah untuk mencapai tujuan

Untuk mencapai tujuan, penelitian ini menggunakan data simulasi dan data riil. Data simulasi diperoleh dengan membangkitkan variabel-variabel (prediktor dan respon) berukuran n = 100 dan p = 10 berdistribusi normal. Terdapat dua jenis data simulasi yang dibangkitkan. Data simulasi 1 berukuran n = 100 dan p = 10 dan terdapat korelasi tinggi diantara variabel prediktor. Parameter yang digunakan untuk membangkitkan variabel respon adalah (1,0 ; 4,0 ; 3,0; 2,0 ; 1,5 ; 1,0 ; 1,0 ; 4,0 ; 3,0 ; 2,0 ;1,5). Dengan ukuran yang sama data simulasi 2 dibangkitkan dan terdapat korelasi sedang diantara variabel prediktor. Parameter yang digunakan untuk membangkitkan variabel respon pada data simulasi 2 adalah (10,0; 4,0; 3,0; 2,0; 1,5 0,0; 2,0; 1,2; 6,0; 2,3; 0,0). Sedangkan data riil yang digunakan berasal dari data pendapatan petani pisang peserta Kelompok Usaha Bersama Agribisnis di Kecamatan Ajung Kabupaten Jember tahun 1998 dengan variabel responnya adalah pendapatan (rupiah). Variabel-variabel prediktornya adalah X_1 = umur, X_2 = jumlah anggota keluarga, X_3 = luas lahan (Ha), X_4 = biaya produksi (rupiah), X_5 = produksi dan X_6 = harga jual (rupiah).

Berdasarkan data-data tersebut di atas, metode OLS, LASSO dan PLS digunakan untuk menduga parameter. Metode OLS menduga parameter dengan meminimumkan jumlah kuadrat sisa yaitu meminimumkan $(\sum_{i=1}^{n} Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2$, sehingga diperoleh $\hat{\beta} = \left(X'X\right)^{-1} X'Y$. LASSO menduga parameter dengan meminimumkan jumlah kuadrat sisa $((\sum_{i=1}^{n} Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2)$ dengan batasan $\sum_{j=1}^{p} |\hat{\beta}_j| \le s$, dengan s adalah parameter penyusutan yang ditentukan oleh pengguna. Ukuran numerik s diperoleh melalui proses cross validasi. Sedangkan PLS menduga parameter dengan terlebih dahulu membentuk variabel baru (variabel laten) dengan dimensi yang lebih kecil dari dimensi variabel-variabel prediktor, kemudian meregresikan variabel-variabel tersebut terhadap variabel respon. Pemilihan dimensi variabel laten diperoleh berdasarkan proses cross validasi. Untuk mengetahui tingkat multikolinearitas variabel-variabel prediktor digunakan kriteria *Variance Inflation Factor* (VIF). Menurut Montgomery & Pack (1991) data mengalami multikolinearitas serius jika nilai *Variance Inflation Factor* lebih dari 10.Untuk mencari performance bias dari ketiga metode tersebut dalam menduga variabel respon digunakan persen bias mutlak $\left(\left|\frac{\hat{Y} - Y_{observasi}}{Y_{observasi}}\right|\right)$. Untuk mencari daya ramal ketiga metode tersebut digunakan kriteria *Mean Square Error Prediction* (MSEP). Pengolahan data dilakukan dengan menggunakan bantuan paket program R versi 2.7.

HASIL DAN PEMBAHASAN

Mendeteksi multikolinearitas

Multikolinearitas dideteksi menggunakan nilai *Variance Inflation Factor* (VIF). Tabel 1 menunjukkan nilai *Variance Inflation Factor* dari masing-masing data yang digunakan.

Tabel 1. Nilai VIF data simulasi.

Variabel Prediktor	VIF SIMULASI1	VIF SIMULASI 2
X_1	72,5	1,7
X_2	72,2	1,8
X_3	1,2	1,0
X_4	1,1	1,0
X_5	1,1	1,0
X_6	1,2	1,0
X_7	1,1	1,1
X_8	1,1	1,1
X_9	1,2	1,1
X_{10}	1,1	1,1

Data simulasi 1 dibangkitkan dengan terdapat korelasi yang tinggi antara variabel X_1 dan X_2 (0,993). Dari tabel di atas nilai *Variance Inflation Factor* variabel X_1 dan X_2 sangat besar (VIF = 72,5 untuk X_1 dan VIF = 72,2 untuk X_2). Jika nilai VIF lebih besar dari 10 artinya terjadi masalah multikolinearitas pada variabel bebas (Montgomery & Peck 1991). Untuk

data simulasi 2, data dibangkitkan dengan korelasi sedang antara X_1 dan X_2 (korelasi = 0,632). Untuk data simulasi 2 tersebut, nilai VIF untuk variabel X_1 adalah 1,7 dan nilai VIF untuk variabel X_2 adalah 1,8.

Berikut ini nilai *Variance Inflation Factor* (VIF) dari data riil penghasilan petani pisang Desa Ajung Jember.

Tabel 2. Nilai VIF data penghasilan petani pisang.

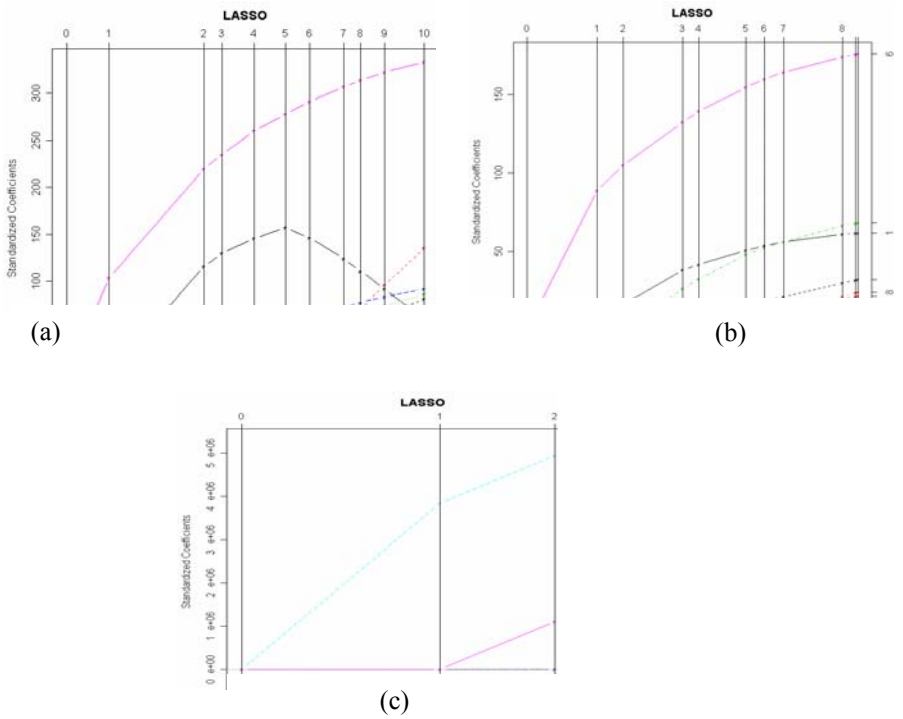
Variabel Prediktor	VIF
X_1	1,3
X_2	1,3
X_3	24,1
X_4	5,2
X_5	30,1
X_6	1,3

Untuk data riil, Nilai VIF terbesar dimiliki oleh variabel bebas X_5 yaitu 30,1 dan X_3 yaitu 24,1. Jadi pada data penghasilan petani pisang terjadi multikolinearitas yang tinggi antara variabel X_3 (luas lahan) dan X_5 (produksi).

Pendugaan koefisien regresi

Metode OLS menduga koefisien regresi dengan meminimumkan jumlah kuadrat sisa yaitu dengan menurunkan fungsi jumlah kuadrat sisa terhadap parameter regresi. Hasil pendugaan koefisien regresi dengan menggunkan OLS untuk data simulasi 1, simulasi 2 dan data riil dapat dilihat pada Tabel 5 sampai Tabel 7

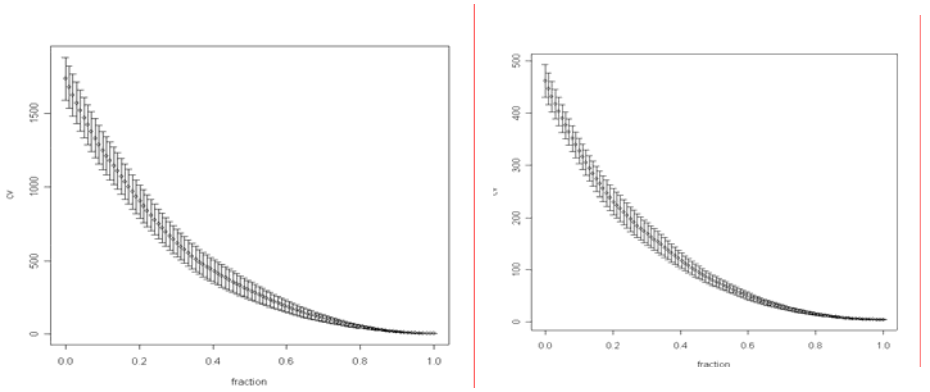
Pendugaan koefisien LASSO dilakukan secara bertahap dengan menetapkan koefisien tahap awal semuanya bernilai 0.Tahapan LASSO dapat dilihat pada gambar1. Pendugaan koefisien regresi LASSO diperoleh dengan menentukan batas yang dibakukan, yaitu $s = t / \sum \left|\hat{\beta}_j^0\right|$ dengan $t = \sum \left|\hat{\beta}_j\right|$ dan $\hat{\beta}_j^0$ adalah penduga kuadrat terkecil untuk model penuh atau pada gambar ditulis sebagai |beta|/max|beta|. Nilai optimal s dapat diperoleh melalui *cross* validasi. Melalui proses *cross* validasi diperoleh bahwa dari data simulasi 1, optimal pada *fraction* (s = 1), data simulasi 2 optimal pada s = 0,99 dan data penghasilan petani pisang optimal pada s = 0,91 (Gambar 2).



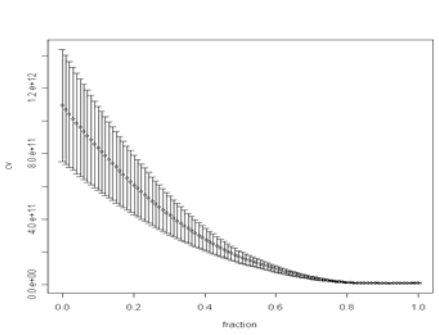
(a) (b)

(c)

Gambar 1. Tahapan Lasso untuk data simulasi 1 (a), data simulasi 2 (b) dan data penghasilan petani pisang (c)



(a) (b)



(c)

Gambar 2. Nilai CV untuk data simulasi 1 (a), data simulasi 2 (b) dan data penghasilan petani pisang (c).

Tabel 3. Nilai *Mean Square Error* metode PLS untuk data Simulasi 1 dan 2.

Komponen	Simulasi 1	Simulasi 2
Intersep	1686,732 (41,48)	448,647 (21,40)
Komponen 1	339,689 (19,26)	123,779 (11,50)
Komponen 2	54,635 (7,835)	42,928 (7,455)
Komponen 3	29,802 (5,833)	29,709 (5,943)
Komponen 4	8,361 (3,073)	10,911 (3,784)
Komponen 5	6,553 (2,730)	9,461 (3,442)
Komponen 6	5,713 (2,544)	6,345 (2,837)
Komponen 7	4,339 (2,248)	4,175 (2,286)
Komponen 8	4,005 (2,206)	3,709 (2,174)
Komponen 9	3,801 (2,164)	**3,704 (2,168)**
Komponen 10	**3,777 (2,158)**	3,701 (2,171)

Keterangan : Nilai dalam tanda kurung merupakan nilai *Root Mean Square Error* (RMSE) dari Proses *Cross* Validasi.

Dengan demikian, nilai koefisien terpilih untuk data simulasi 1 adalah tahap ke-10, simulasi 2 adalah tahap ke-9 dan untuk data penghasilan petani pisang adalah tahap ke-4 (Gambar 1). Nilai koefisien untuk model terpilih dapat dilihat pada Tabel 5 sampai Tabel 7.

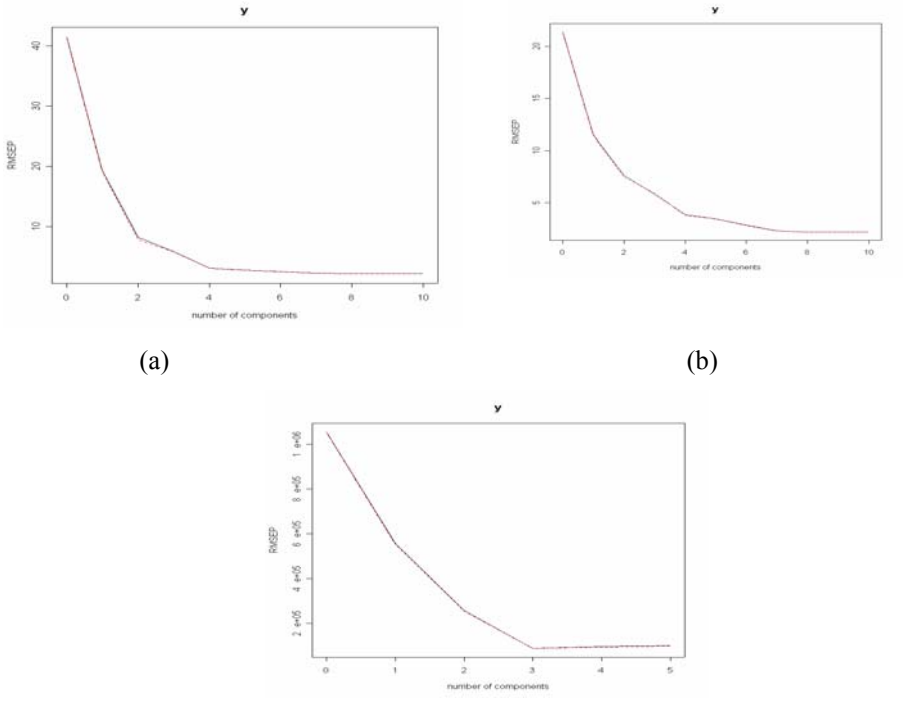
Metode PLS menduga koefisien regresi melalui prosedur pemilihan jumlah komponen yang digunakan dalam model dengan Mean Square Error Optimal (dipilih MSE minimum). Nilai-nilai MSE untuk pemilihan koefisien regresi dapat dilihat pada Tabel 3 dan 4.

Dari Tabel 3 dapat diketahui bahwa untuk data simulasi 1 model optimal dipilih dengan melibatkan 10 komponen. Hal ini sesuai dengan hasil proses cross validasi. Untuk data simulasi 2 pemilihan model tanpa mempertimbangkan cross validasi, model optimal dipilih dengan melibatkan 10 komponen. Akan tetapi model optimal dengan mempertimbangkan proses cross validasi, model optimal melibatkan 9 komponen. Jadi dalam hal ini, model terpilih melibatkan 9 komponen. Untuk data riil pengahasilan petani pisang, model tanpa mempertimbangkan cross validasi, model optimal melibatkan 5 komponen, tetapi jika mempertimbangkan cross validasi model optimal melibatkan 3 komponen. Jadi untuk data penghasilan petani pisang model terpilih melibatkan 3 komponen. Nilai koefisien model terpilih untuk masing-masing jenis data dapat dilihat pada tabel 5 sampai Tabel 7. Sedangkan plot dari Root Mean Square Error Prediction (RMSEP) dari proses cross validasi untuk masing-masing jenis data dapat dilihat pada Gambar 3.

Tabel 4. Nilai *Mean Square Error* metode PLS untuk data penghasilan petani pisang dengan 5 komponen.

Komponen	MSE
Intersep	1,038e+12 (1054199)
Komponen 1	2,491e+11 (544488)
Komponen 2	4,663e+10 (259459)
Komponen 3	**5,628e+09 (87336)**
Komponen 4	5,424e+09 (91022)
Komponen 5	5,416e+09 (94212)

Keterangan : Nilai dalam tanda kurung merupakan nilai *Root Mean Square Err.*



Gambar 3. Plot nilai RMSEP vs komponen untuk data simulasi 1(a) data simulasi 2 (b) dan data penghasilan petani pisang (c).

Tabel 5. Nilai koefisien data simulasi 1 (korelasi tinggi).

Variabel Prediktor	Koefisien asli	Pendugaan OLS	Pendugaan LASSO	Pendugaan PLS	Penduga Paling Besar	Penduga Paling Tepat
intersep	1,0	5,47200	5,47200	5,47244	PLS	OLS/LASSO
X_1		3,04180	3,04184	3,04184	PLS/	PLS/LASSO
X_2	4,0	3,44180	3,44177	3,44177	LASSO	PLS/LASSO
X_3	3,0	2,04456	2,04456	2,04456	OLS	-
X_4	2,0	1,63870	1,63868	1,63868	-	PLS/LASSO
X_5	1,5	1,40960	1,40958	1,40958	OLS	PLS/LASSO
X_6	1,0	3,99546	3,99546	3,99546	OLS	-
X_7	1,0	2,83872	2,83872	2,83872	-	-
X_8	4,0	2,06890	2,06889	2,06889	-	OLS
X_9	3,0	1,35841	1,35841	1,35841	OLS	-
X_{10}	2,0	0,96176	0,96176	0,96176	-	
	1,5					

Tabel 6. Nilai koefisien data simulasi 2 (korelasi sedang).

Variabel Prediktor	Koefisien asli	Pendugaan OLS	Pendugaan LASSO	Pendugaan PLS	Penduga Paling Kecil	Penduga Paling Tepat
intersep	10,0	28,3500	28,3500	27,2636	PLS	PLS
X_1	4,0	3,8374	3,8293	3,8634	LASSO	PLS
X_2	3,0	3,4482	3,3920	3,3467	LASSO	PLS
X_3	2,0	1,6284	1,5909	1,6251	LASSO	OLS
X_4	1,5	1,2640	1,2263	1,2604	LASSO	OLS
X_5	0,0	-0,0428	0,0000	-0,0152	LASSO	LASSO
X_6	2,0	2,0017	1,9978	2,0014	LASSO	PLS
X_7	1,2	1,1225	1,1094	1,1244	LASSO	PLS
X_8	6,0	5,2766	5,1711	5,3163	LASSO	PLS
X_9	2,3	2,2782	2,2692	2,2817	LASSO	PLS
X_{10}	0,0	0,0352	0,0294	0,0348	LASSO	LASSO

Tabel 7. Nilai koefisien data penghasilan petani pisang (korelasi tinggi).

Variabel Prediktor	Pendugaan OLS	Pendugaan LASSO	Pendugaan PLS	Penduga Paling Besar
intersep	-2573748,00	-2573748,00	-2501297,00	OLS/LASSO
X_1	2220,00	1652,07	380,98	OLS
X_2	3490,00	0,00	60,10	OLS
X_3	-5199,00	0,00	-7,71	OLS
X_4	-0,09	-0,04	-0,09	OLS/PLS
X_5	181,95	174,73	181,49	OLS
X_6	8755,50	8604,20	8817,21	PLS

Ketepatan menduga variabel respon (*Dependent Variable*)

Untuk mengetahui ketepatan dari ketiga metode tersebut dalam menduga variabel respon digunakan rata-rata persen mutlak bias. Tabel 8 menunjukkan persen mutlak bias yang dihasilkan dari data simulasi dengan menggunakan ketiga metode.

Tabel 8. Rata-rata persen mutlak bias dari metode OLS, LASSO dan PLS.

	OLS	LASSO	PLS
• Rata-rata MSEP Model Simulasi 1	4,856	4,791	4,669
• Varian MSEP Model Simulasi 1	2,913	0,007	0,016
• Rata-rata MSEP Model Simulasi 2	4,679	4,766	4,702
• Varian MSEP Model Simulasi 2	2,213	0,011	0,009
• Rata-rata MSEP Model Penghasilan Petani Pisang	1,117 E+10	1,001E+ 10	7,671 E+09
• Varian MSEP Model Penghasilan Petani Pisang	5,825 E+19	5,546E+ 17	1,702 E+17

Dari Tabel 8 di atas dapat diketahui bahwa untuk data yang mengalami masalah multikolinearitas (adanya korelasi tinggi diantara variabel prediktor) LASSO cenderung memiliki rata-rata tingkat bias yang lebih kecil daripada PLS dalam menduga variabel respon.

Daya ramal dan kekonsistenan OLS, PLS dan LASSO

Daya ramal suatu model ditunjukkan oleh nilai *Mean Square Error Prediction* (MSEP) sebagai indikator seberapa baikkah model regresi terpilih bisa meramal amatan di masa akan datang. Semakin kecil nilai MSEP semakin baik model tersebut dalam meramal amatan di masa mendatang. Tabel 9 menunjukkan nilai MSEP dari ketiga metode yang diperoleh dari 10 *cross* validasi. Dari tabel tersebut dapat diketahui bahwa keragaman dari MSEP OLS selalu lebih besar dibanding yang lain. Hal ini menunjukkan bahwa OLS lebih tidak teliti/tidak konsisten dalam menduga nilai MSEP dibanding lainnya. Nilai MSEP yang diperoleh dengan menggunakan metode PLS lebih kecil daripada MSEP dari Metode LASSO.

Tabel 9. Nilai rata-rata dan varian MSEP dari 10 *cross* validasi.

	OLS	LASSO	PLS
• Rata-rata MSEP Model Simulasi 1	4,856	4,791	4,669
• Varian MSEP Model Simulasi 1	2,913	0,007	0,016
• Rata-rata MSEP Model Simulasi 2	4,679	4,766	4,702
• Varian MSEP Model Simulasi 2	2,213	0,011	0,009
• Rata-rata MSEP Model Penghasilan Petani Pisang	1,117E +10	1,001E +10	7,671E +09
• Varian MSEP Model Penghasilan Petani Pisang	5,825E +19	5,546E +17	1,702E +17

KESIMPULAN

Dari hasil-hasil yang diperoleh mengenai metode OLS, PLS dan LASSO pada data dalam penelitian ini dengan mengandung multikolinearitas, dapat disimpulkan bahwa:

1. Metode OLS menduga koefisien regresi dengan meminimumkan jumlah kuadrat sisa yaitu dengan menurunkan fungsi jumlah kuadrat sisa terhadap parameter regresi. Pendugaan koefisien LASSO dilakukan secara bertahap dan pada masing-masing tahap dicari nilai $s = t / \sum \left| \hat{\beta}_j^0 \right|$ dengan $t = \sum \left| \hat{\beta}_j \right|$ dan $\hat{\beta}_j^0$ adalah penduga kuadrat terkecil untuk model penuh, nilai optimal s dapat diperoleh melalui *cross* validasi. Metode PLS menduga koefisien regresi melalui prosedur pemilihan jumlah komponen yang digunakan dalam model dengan *Mean Square Error* Optimal (dipilih MSE minimum), MSE optimal diperoleh melalui proses *cross* validasi.

2. Untuk data yang mengandung multikolinearitas, PLS dan LASSO cocok digunakan untuk menduga koefisien regresi, memberikan hasil yang lebih tepat dibanding OLS. Metode OLS cenderung menduga koefisien regresi lebih besar dibanding pendugaan menggunakan metode lainnya.
3. Untuk data yang mengalami masalah multikolinearitas (adanya korelasi yang tinggi diantara variabel prediktor) LASSO cenderung memiliki rata-rata tingkat bias yang lebih kecil daripada PLS dalam menduga variabel respon.
4. Keragaman dari MSEP OLS selalu lebih besar dibanding yang lain. Hal ini menunjukkan bahwa OLS lebih tidak teliti/tidak konsisten dalam menduga nilai MSEP dibanding lainnya.
5. Nilai MSEP yang diperoleh dengan menggunakan metode PLS lebih kecil daripada MSEP dari Metode LASSO

DAFTAR PUSTAKA

Abdi H. 2006. *Partial Least Squares Regression (PLSR)*. [Online] http://www.statisticssolutions.com/Partial-Least-Squares-Regression [07 Januari 2008].

Datta S. 2001. Exploring Relationship in Gene Expression : A Partial Least Square Approach. *Gene Expression*, **9**: 249 – 255

Datta S, Jennifer LR & Somnath D. 2007. Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO. *Biometrics*, **63**: 259 – 271.

Efron B, Hastie T, Johnstone I & Tibshirani R. 2004. Least Angle Regression (with discussions). *Annals of Statistics*, **32**: 407 – 499.

Frank IE & Friedman JH. 1993. A Statistical View of Some Chemometrics Regression Tools (with discussion). *Technometrics*, **35**: 109 – 148.

Geladi P. 1992. Wold, Herman, the father of PLS. *Chemometrics and Intelligent Laboratory Systems*, **15**: 1, R7 – R8.

Gujarati D. 1992. *Ekonometrik Dasar* (Terjemahan), Edisi ke-2. Alih Bahasa Zeinn, S. Erlangga, Jakarta.

Mendelhall W & Sincich T. 1996. *A Second Course in Statistics Regression Analysis*, 5th. New Jersey.

Montgomery DC & Peck EA. 1991. *Introduction to Linear Regression Analysis*, New York : John Wiley & Sons.

Tibshirani R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of The Royal Statistical Society, Series B*, **58**: 267 – 288.