Forum Statistika dan Komputasi : *Indonesian Journal of Statistics* available online at:
ISSN : 0853-8115 *journal.ipb.ac.id/index.php/statistika*
Vol. 18 No.1, April 2013, p: 21-27

LASSO : SOLUSI ALTERNATIF SELEKSI PEUBAH DAN PENYUSUTAN KOEFISIEN MODEL REGRESI LINIER

(Lasso: An Alternative Solution for Selection and Shrinkage Linear Regression Models)

Agus M Soleh[1], Aunuddin[1]
[1]Departemen Statistika IPB
Email: agusms@ipb.ac.id

Abstract

A new method, known as LASSO, has recently developed for selections and shrinkage linear regression methods. The method gives an alternative solution on high correlated data between independent variables, where the least squares produces high variance. Based on simulation this method is not better than forward selection (in the case the parameters contains many zero values) and ridge regression (in the case all parameter values close to zero). Unknowing the true parameter and consistency estimates for all conditions that put the LASSO is better than ridge or forward selection.

Keywords *: LASSO, least square, forward selection, ridge, cross validation*

PENDAHULUAN

Model regresi linier pada model linier klasik diduga menggunakan metode kuadrat terkecil dengan cara meminimumkan jumlah kuadrat sisaan. Perhatikan sebuah penduga $\tilde{\theta}$ yang digunakan untuk menduga θ, kuadrat tengah galat didefinisikan sebagai berikut:

$$KTG(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 = Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2.$$

Bagian pertama menyatakan ragam sedangkan bagian kedua menyatakan bias kuadrat. Teorema Gauss-Markov mengimplikasikan bahwa penduga kuadrat terkecil memiliki kuadrat tengah galat terkecil dari seluruh penduga linier tak bias (Ryan, 1997). Pada kondisi tertentu, kuadrat terkecil sering tidak memuaskan yang disebabkan oleh dua hal (Thibsirani, 1996), yaitu:

a. Keakuratan prediksi: penduga kuadrat terkecil memiliki bias rendah tetapi ragam besar.
b. Interpretasi: semakin banyak peubah penjelas, maka model menjadi semakin sulit diinterpretasikan.

Pada kondisi ini dimungkinkan penggunaan metode seleksi dan penyusutan dalam menduga model regresi. Metode seleksi menggunakan Subset Terbaik dan Regresi Bertatar dapat mengurangi ragam prediksi dengan mengorbankan sedikit bias. Interpretasi model semakin mudah karena model hanya memuat subset dari keseluruhan peubah penjelas. Kekurangan metode ini adalah penduga model tidak stabil, dimana perubahan kecil pada data dapat menghasilkan model yang berbeda (termasuk subset peubahnya).

Alternatif lain adalah dengan menyusutkan koefisien regresi menggunakan regresi gulud. Seperti pada model seleksi, regresi gulud juga menurunkan ragam prediksi dengan mengorbankan sedikit bias. Penduga model yang dihasilkan oleh regresi gulud lebih stabil, tetapi untuk interpretasi model relatif lebih sulit dibandingkan metode seleksi apabila jumlah peubah yang digunakan sangat banyak.

Thibsirani (1996) mengembangkan metode LASSO (*Least Absolute Shrinkage and Selection Operator*) yang bertujuan mengatasi masalah dalam keakuratan pendugaan dan interpretasi, dengan mempertahankan keuntungan-keuntungan metode seleksi subset dan regresi gulud. Motivasi pengembangan LASSO berasal dari metode *Non-negative Garrote* yang dikembangkan sebelumnya oleh Breiman (1995).

Tulisan ini bertujuan memberikan *state of the art* metode LASSO yang telah dikembangkan Thibsirani (1996) dan membuat suatu simulasi untuk mengetahui penerapan terbaik metode Laswo dibanding kuadrat terkecil, gulud, dan *forward selection*.

LANDASAN TEORI

Regresi Linier

Misalkan terdapat vektor peubah bebas $X^T = (X_1, X_2, \ldots, X_p)$ dan digunakan untuk memprediksi luaran nilai Y yang berupa bilangan real. Model regresi linier memiliki bentuk:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

Untuk menduga $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_j)^T$ digunakan sekumpulan data $(x_1, y_1) \ldots (x_N, y_N)$. Metode Kuadrat Terkecil meminimumkan jumlah kuadrat

sisaan dalam menduga koefisien $\boldsymbol{\beta}$ (Hastie *et al*, 2008), yaitu dengan meminimumkan persamaan:

$$JKS(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2 = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2.$$

Dalam catatan matriks, di mana **X** berukuran *N x (p+1)* dan **y** adalah vektor-*N*, jumlah kuadrat sisaan dapat ditulis sebagai :

$$JKS(\beta) = (y - \boldsymbol{X}\beta)^T(y - \boldsymbol{X}\beta)$$

Dari kalkulus dengan menurunkan JKS(β) terhadap β, diperoleh JKS(β) minimum, yaitu dalam bentuk:

$$\boldsymbol{X}^T y = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

yang disebut sebagai persamaan normal.

Jika $\boldsymbol{X^TX}$ adalah matriks berpangkat penuh, maka β yang diduga oleh $\hat{\beta}$ akan menghasilkan solusi unik, yaitu:

$$\hat{\beta} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^T}\boldsymbol{y}.$$

Regresi Gulud

Regresi gulud diperkenalkan oleh Hoerl dan Kennard (1970) (dalam Draper & Smith, 1998) yang diusulkan untuk menangani ketidakstabilan penduga kuadrat terkecil. Regresi gulud mempenalti ukuran dari koefisien regresi pada *norm* L_2 atau secara spesifik menduga $\hat{\beta}$ dengan meminimumkan $JKS(\beta)$ dengan kendala:

$$\sum_{j=1}^{p}\beta_j^2 \le t.$$

Masalah regresi gulud ini dapat ditulis dengan cara lain yaitu memiminumkan:

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2.$$

Pada kedua persamaan di atas terdapat korespondensi satu-ke-satu antara t dan $\lambda \ge 0$. Solusi regresi gulud didapat dengan cara yang sama seperti kuadrat terkecil, yaitu dengan meminimumkan jumlah kuadrat sisaan (JKS):

$$JKS(\beta, \lambda) = (y - \boldsymbol{X}\beta)^T(y - \boldsymbol{X}\beta) + \boldsymbol{\lambda}\boldsymbol{\beta^T}\beta$$

yang memperoleh persamaan:

$$\boldsymbol{X}^T y = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda I})\beta.$$

Dengan cara seperti ini $(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda I})$ dapat dijamin selalu berpangkat penuh walaupun $\boldsymbol{X}^T\boldsymbol{X}$ tidak berpangkat penuh dengan mengambil $\lambda > 0$. Untuk $\lambda = 0$ persamaan ini adalah persamaan normal seperti yang diperoleh menggunakan kuadrat terkecil. Solusi yang unik dapat diperoleh dalam bentuk tertutup:

$$\hat{\beta}^{gulud} = (\boldsymbol{X^TX} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X^T}y.$$

Penduga koefisien yang diperoleh menggunakan regresi gulud adalah tidak *equivariant* (Hastie *et al*., 2008), artinya penduga koefisien tersebut akan berbeda hasilnya jika peubah asal dibakukan dengan peubah asal tidak dibakukan. Oleh karena itu untuk pendugaan $\hat{\beta}^{gulud}$ ini sebelumnya disarankan membakukan skala dari peubah asal sehingga memiliki nilai harapan nol dan ragam satu.

Forward Selection

Forward selection merupakan salah satu metode untuk seleksi peubah dalam regresi linier, yaitu teknik untuk mendapatkan model regresi dengan cara menseleksi peubah yang memenuhi kriteria tertentu. Teknik yang umum dalam seleksi peubah adalah "Semua Kemungkinan Regresi" (*All Possible Regression*), Subset Terbaik (*Best* Subset) dan Regresi Bertatar (*Stepwise Regression*). Pada "Semua Kemungkinan Regresi" dipilih model dengan mencari kombinasi peubah dari seluruh kemungkinan peubah sebanyak 2^p (termasuk model konstanta) oleh kriteria R^2, JKS atau Cp (Draper & Smith, 1998). Kelemahan metode ini adalah semakin banyak peubah yang dievaluasi akan semakin banyak model yang harus dievaluasi. Oeh karena itu dikembangkan alternatif lain yaitu metode Subset Terbaik yang hanya menghitung sebanyak K terbaik untuk model dengan satu, dua, tiga dan seterusnya peubah (Draper & Smith, 1998). Beberapa paket perangkat lunak membatasi penggunaan metode ini hanya untuk maksimum 20 peubah.

Regresi bertatar menyeleksi peubah dengan cara memasukkan atau membuang satu persatu peubah ke dalam model. Terdapat tiga teknik dalam regresi bertatar, yaitu: *backward elimination*, *forward selection* dan *stepwise* (gabungan dari *backward* dan *forward*). Pada *backward elimination* model pertama diduga dengan memasukkan seluruh peubah. Selanjutnya model dievaluasi dengan mengeluarkan satu persatu peubah yang tidak memenuhi kriteria. *Forward selection* berlaku kebalikan, model dimulai dengan konstanta kemudian dievaluasi dengan memasukkan satu persatu peubah. *Stepwise* menggabungkan keduanya dimulai dari *forward selection* kemudian untuk setiap peubah yang masuk dievaluasi dengan *backward elimination.* Kriteria untuk memasukkan atau membuang peubah didasarkan pada salah satu kriteria statistik : R^2, R^2_{adj}, JKS, F, Cp's Mallows atau AIC (Draper & Smith (1998); Ryan (1997); Venables & Ripley (2002)).

LASSO

Dalam regresi gulud, penduga koefisien regresi disusutkan ke arah nol seiring dengan peningkatan nilai λ. Satu hal yang tidak dapat

dilakukan oleh regresi gulud adalah melakukan seleksi peubah secara otomatis dikarenakan secara simultan koefisien yang diduga mungkin tidak bernilai nol. Perhatikan jika kendala seperti dalam regresi gulud diubah menjadi (Thibsirani, 1996)

$$\sum_{j=1}^{p} |\beta_j| \leq t,$$

atau dalam bentuk persamaan *lagrange* ditulis:

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Untuk mendapatkan solusi penduga koefisien tidak dapat diperoleh dalam bentuk tertutup, tetapi harus menggunakan pemrograman kuadratik (Tibshirani, 1996). Dampak yang terjadi dari pengubahan kendala ini sangat besar, yaitu menyebabkan koefisien menyusut ke arah nol seperti dalam regresi gulud dan beberapa koefisien menghasilkan nilai nol secara tepat.

Ide dasar LASSO berasal dari *Non-negative Garrotte* (Breiman, 1995) yang meminimumkan, terhadap $c = \{c_j\}$:

$\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} c_j x_{ij}\hat{\beta}_j\right)^2$ dengan kendala $c_j \geq 0$,

$$\sum_{j=1}^{p} c_j \leq t,$$

di mana $\hat{\beta}_j$ adalah penduga kuadrat terkecil biasa. NN-Garrotte ini tidak terdefinisikan ketika p > N (yang bukan merupakan topik panas pada tahun 1995) (Thibsirani, 2011). Pada sekitar tahun tersebut, beberapa metode yang mirip dengan LASSO telah dikembangkan berdasarkan penalty-l_1, seperti *ridge regression* (Frank dan Friedman, 1993 dan *basis pursuit (C*hen *et al.* 1998 dalam Thibsirani ,2011). Setelah publikasi pertama tahun 1996, makalah LASSO ini tidak mendapatkan perhatian sampai tahun 2002 setelah berkembangkannya algoritma LAR (*Least Angle Regresion*) oleh Hastie.

Hastie mengembangkan algoritma LAR yang digunakan untuk menduga model regresi linier dalam bentuk model umum:

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = f(x) = \beta_0 + \beta_M\phi_1(\boldsymbol{x}) + \beta_M\phi_2(\boldsymbol{x}) + \cdots + \beta_M\phi_M(\boldsymbol{x}),$$

di mana ϕ_M adalah fungsi nonlinier dari prediktor **X** asli (Hesterberg *et al*, 2008). Modifikasi dari LAR untuk LASSO menghasilkan efisiensi algoritma dalam menduga solusi penduga koefisien LASSO dengan komputasi yang lebih cepat dibandingkan pemrograman kuadratik. Selain untuk menduga koefisien LAR dan LASSO, algoritma LAR ini juga dimodifikasi untuk digunakan dalam menduga koefisien regresi *Forward Stepwise* dan *Forward Selection*, sehingga kemudian namanya dikenal sebagai LARS (untuk LAR, LASSO, *Stagewise* dan *Forward Selection*).

Algoritma LAR asli adalah sebagai berikut (Hastie *et al*, 2008):

1. Bakukan prediktor sehingga memiliki nilai tengah nol dan ragam satu. Mulai dengan sisaan $r = y - \bar{y}, \beta_1, \beta_2, \cdots, \beta_p = 0$.
2. Cari prediktor x_j yang paling berkorelasi dengan r.
3. Ubah nilai β_j dari 0 bergerak menuju koefisien kuadrat terkecil $\langle x_j, r\rangle$, sampai kompetitor lain x_k memiliki korelasi sebesar korelasi x_j dengan sisaan sekarang.
4. Ubah nilai β_j dan β_k bergerak dalam arah yang didefinisikan oleh koefisien kuadrat terkecil bersama dari sisaan sekarang dalam (x_j, x_k) sampai kompetitor x_l lain memiliki korelasi dengan sisaan sekarang dengan besaran yang sama.
5. Teruskan cara ini sampai semua p prediktor telah masuk. Setelah min(N-1, p) langkah, solusi model penuh untuk kuadrat terkecil diperoleh.

Modifikasi algoritma LAR untuk mendapatkan solusi LASSO adalah dengan memodifikasi langkah ke-4, yaitu dengan cara:

4a. Jika koefisien bukan nol mencapai nilai nol, keluarkan peubah tersebut dari gugus peubah aktif dan hitung kembali arah kuadrat terkecil bersama.

LAR selalu mengambil p langkah untuk mendapatkan penduga kuadrat terkecil secara penuh, sedangkan modifikasi LAR untuk LASSO dapat memiliki lebih dari p langkah untuk mendapatkannya. Algoritma LASSO dengan memodifikasi LAR adalah suatu cara yang efisien dalam komputasi solusi masalah LASSO khususnya ketika p >> N (Hastie *et al*, 2008).

Pemilihan Nilai Penalti dalam Gulud dan LASSO

Beberapa metode telah dikembangkan untuk memilih nilai penalti dalam regresi gulud dan LASSO. Metode yang umum digunakan dalam pemilihan nilai penalti ini adalah validasi silang (*Cross Validation*/CV). Ide dari validasi silang adalah membagi data menjadi dua bagian, yaitu: data *training* dan data *test*. Data *training* digunakan untuk mengepas nilai $\hat{\beta}$ dan data *test* digunakan untuk menguji kebaikan prediksi dari $X\hat{\beta}$. Nilai dari validasi silang ini merupakan penduga bagi galat prediksi (*prediction error*) (Izenman, 2008).

Terdapat beberapa tipe dari validasi silang yang mengatur bagaimana data *training* dan data *test*. Tipe validasi silang yang umum adalah validasi silang *leave-one-out* (LOO) dan *k-fold*. Validasi silang LOO menggunakan satu observasi sebagai data *test* dan sisanya sebagai data *training*. Hal ini diulang sampai setiap observasi pernah menjadi data *test*. Dalam validasi silang *k-fold*, semua observasi dipartisi secara acak ke dalam k

sub-contoh. Setiap sub-contoh digunakan sebagai data *test* dan sisanya digunakan sebagai data *training*. Proses validasi silang diulang sampai k kali dan setiap satu sub-contoh digunakan hanya sekali dalam data *test*. Jika k sama dengan banyaknya observasi, maka validasi *k-fold* menjadi validasi silang LOO.

Nilai Galat Prediksi ($\widehat{PE}$) diduga oleh CV dengan menggunakan persamaan:

$$\widehat{PE} = CV = \frac{1}{k}\sum_{i}\left(y_i - \hat{y}_{-i(k)}\right)^2,$$

di mana $\hat{y}_{-i(k)}$ adalah dugaan y pada saat *fold* ke-k tidak digunakan dalam menduga model. Shao (1993) dalam Mahmood & Khan (2009) membuktikan secara *asymptotic* dan simulasi bahwa model dengan nilai CV minimum dari validasi silang LOO sering melebihi dari model yang dispesifikasikan, yaitu terlalu banyak peubah yang tidak signifikan masuk dalam model regresi. Izenman (2008) menyarankan menggunakan validasi silang *10*-fold (juga *5*-fold) disebabkan perbedaan pemilihan nilai *k* (apakah 5, 10 atau n) adalah masalah "bias versus ragam". Validasi silang LOO (k=n) menghasilkan $\widehat{PE}$ yang memiliki bias rendah tetapi ragam tinggi, sedangkan 5-fold dan 10-fold menghasilkan $\widehat{PE}$ dengan bias tinggi tetapi ragam rendah.

Secara Umum untuk mendapatkan nilai CV merupakan suatu langkah yang tidak efisien, sehingga dikembangkan validasi silang terampat (*Generalized Cross Validation*/GCV) yang lebih sederhana, yaitu:

$$GCV = \frac{1}{n}\sum_{i}\left(\frac{y_i - \hat{y}_i}{1 - tr(\boldsymbol{H})/n}\right)^2$$

di mana H adalah matriks *hat*. GCV mensyaratkan adanya matriks H, sehingga metode ini tidak bisa dikenakan pada metode LASSO.

Metode lain yang dapat digunakan selain validasi silang adalah Cp's Mallows. Statistik Cp pertama kali dikemukakan oleh C. L. Mallows dengan menggunakan persamaan (Draper & Smith, 1998):

$$Cp = \frac{JKS_p}{s_k^2} - (n - 2p)$$

di mana JKS_p adalah jumlah kuadrat sisaan dari model yang mengandung p parameter termasuk intersep, dan s_k^2 adalah kuadrat tengah galat untuk model lengkap yang diasumsikan merupakan nilai dugaan tak bias bagi galat σ^2. Nilai Cp diplotkan dengan banyaknya parameter dari subset yang menjadi pertimbangan. Model yang dapat diterima dalam pengertian meminimumkan total bias adalah model yang nilai Cp-nya mendekati nilai p (banyaknya parameter).

CONTOH KASUS : SIMULASI

Pada kondisi peubah-peubah prediktor saling bebas dan peubah tersebut mempengaruhi respons dengan jelas, metode kuadrat terkecil merupakan metode pendugaan tak bias terbaik dalam menduga model regresi. Tetapi pada kondisi sebaliknya alternatif-alternatif seperti seleksi peubah dan penyusutan mungkin diperlukan dalam menduga model. Untuk menentukan mana yang disarankan untuk digunakan, disusun suatu simulasi pada kondisi peubah bebas saling berkorelasi tinggi. Keempat metode kemudian digunakan dalam menduga koefisien regresinya dengan menggunakan software R. Tahapan simulasi dilakukan sebagai berikut:

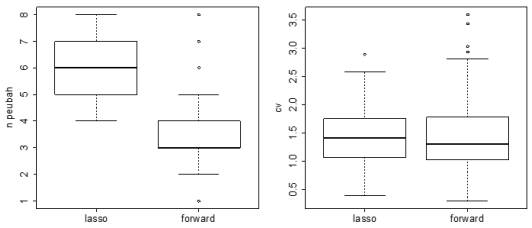
1. Pembangkitan data simulasi.
   a. Membangkitkan peubah X sebanyak 8 peubah yang memiliki nilai korelasi ρ = 0.9 antar peubah dengan jumlah n=20.
   b. Membangkitkan peubah Y dengan model persamaan regresi linier dari 8 peubah yang dihasilkan pada bagian 1(a) ditambah sisaan ~ Normal(0,0.5). Parameter koefisiennya berturut-turut:
      - bernilai 3, 2, 0, 2, 0, 0.7, 0 dan 0. Pemilihan nilai ini sembarang, tetapi memiliki keterwakilan nilai > 1, < 1 dan 0;
      - bernilai dekat dengan nol, yaitu sebesar 0.7;
      - bernilai > 1 untuk semua, yaitu sebesar 3.
2. Menduga koefisien model dengan metode kuadrat terkecil, regresi gulud, LASSO dan *forward selection*. Pemilihan model untuk gulud menggunakan kriteria GCV, sedangkan LASSO dan *forward selection* menggunakan nilai CV minimum yang pertama.
3. Tahap 1(b) dan 2 diulang sebanyak 100 kali.
4. Hasil pendugaan di plot menggunakan boxplot untuk setiap penduga β.

Skenario simulasi pertama, di mana digunakan parameter koefisien regresi β = $(3, 2, 0, 2, 0, 0.7, 0, 0)^t$, menghasilkan grafik yang disajikan pada Lampiran 1. Pendugaan dengan menggunakan metode kuadrat terkecil menghasilkan keragaman yang lebih besar atau relatif sama dibanding dengan metode pendugaan gulud, LASSO ataupun *forward selection*. Pada metode gulud, keragamaan pendugaan koefisien yang dihasilkan relatif lebih kecil di banding ketiga metode lainnya, kecuali pada saat koefisien parameter sebenarnya bernilai nol.

LASSO dan *forward selection* pada simulasi ini dapat menyeleksi peubah dengan baik. Kedua metode ini memberikan sebaran yang lebih sempit pada parameter koefisien regresi bernilai nol. Dalam hal ini, *forward selection* menseleksi penduga lebih baik dibanding dengan LASSO,

kecuali terhadap peubah yang memiliki kontribusi yang sangat kecil. Pada peubah ini, *forward selection* menghasilkan dugaan yang sangat jauh dari nilai sebenarnya.

Gambar 1 menyajikan sebaran banyaknya peubah yang terseleksi oleh metode LASSO dan *forward selection* dan sebaran nilai cv-nya. Seleksi peubah yang dilakukan metode LASSO lebih stabil dibanding dengan seleksi peubah oleh *forward selection*, hal ini terlihat dari kecenderungan bentuk sebaran yang simetrik untuk LASSO dibanding *forward selection* yang menjulur ke kanan. Selain itu, LASSO memiliki kecenderungan untuk melakukan *over fitting (*terlihat dari median peubah yang terseleksi adalah 6), sedangkan *forward selection* memiliki kecenderungan *under fitting* dengan median peubah terseleksi bernilai 3. Galat prediksi yang diduga menggunalan nilai cv, kedua metode menunjukkan nilai yang relatif sama baiknya dalam pendugaan maupun dalam keragamannya.

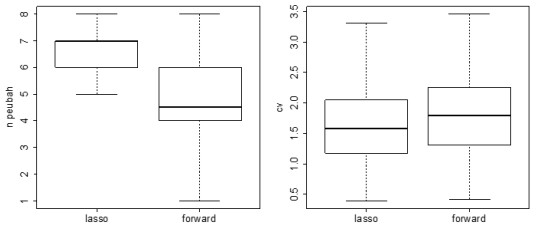


(a) (b)

Gambar 1 Sebaran (a) banyaknya peubah yang terseleksi, (b) nilai cv validasi silang hasil simulasi untuk metode LASSO dan *forward selection* pada saat koefisien $\beta = (3, 2, 0, 2, 0, 0.7, 0, 0)^t$.

Skenario simulasi kedua, di mana semua koefisien **β** mendekati nilai nol, metode gulud merupakan metode yang unggul dibanding metode lainnya. Terlihat pada Lampiran 2, pendugaan menggunakan metode gulud konsisten menghasilkan nilai dugaan yang biasnya kecil secara empirik dengan keragaman yang paling kecil di antara metode-metode lainnya. Metode selanjutnya yang terbaik adalah LASSO yang memberikan pendugaan lebih baik dengan keragaman yang lebih kecil dibanding kuadrat terkecil atau *forward selection*. Metode yang memberikan keragaman pendugaan parameter paling besar pada skenario ini adalah *forward selection*.

Pada kondisi skenario simulasi kedua ini, *forward selection* cenderung untuk melakukan *under fitting*, di mana peubah yang terseleksi hampir setengahnya kurang (Gambar 2) dengan keragaman yang sangat tinggi. Bentuk sebaran yang diperoleh dari *forward selection* cenderung menjulur ke kanan (median=4.5) dibanding bentuk sebaran yang diperoleh oleh LASSO yang cenderung menjulur ke kiri (median=7) dengan keragaman yang lebih kecil. Pendugaan galat prediksi dengan menggunakan nilai cv, kedua metode menghasilkan nilai dan keragaman yang relatif tidak berbeda seperti pada kondisi skenario simulasi pertama.

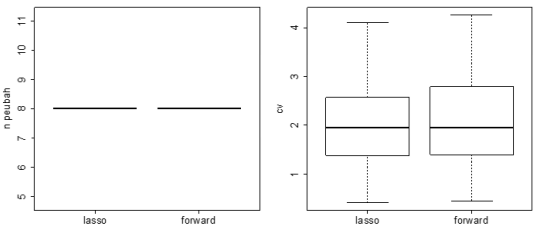


(a) (b)

Gambar 2 Sebaran (a) banyaknya peubah yang terseleksi, (b) nilai cv hasil simulasi untuk metode LASSO dan *forward selection* pada saat koefisien $\beta = (0.7, 0.7, 0.7, 0.7, 0.7, 0.7, 0.7, 0.7)^t$.

Skenario simulasi terakhir, di mana parameter koefisien β yang akan diduga adalah besar, metode gulud masih merupakan metode yang lebih unggul dibanding ketiga metode lainnya (Lampiran 3). Metode kuadrat terkecil, LASSO dan *forward selection* menghasilkan nilai pendugaan yang sama.

pada kondisi skenario simulasi ketiga ini, tidak ada peubah yang tidak terseleksi. Banyaknya peubah yang terseleksi adalah sama dengan peubah yang berpengaruh sebenarnya (Gambar 3). Demikian juga dengan penduga untuk galat prediksi menggunakan nilai cv yang tidak berbeda nyata, baik dalam hal bentuk maupun penyebarannya.



(a) (b)

Gambar 3 Sebaran (a) banyaknya peubah yang terseleksi, (b) nilai cv hasil simulasi untuk metode LASSO dan *forward selection* pada saat koefisien $\beta = (3, 3, 3, 3, 3, 3, 3, 3)^t$.

SIMPULAN

LASSO memberikan suatu alternatif bagi penyeleksian peubah dan pendugaan koefisien regresi pada kondisi peubah bebas tidak benar-benar saling bebas. Hasil simulasi dengan peubah-peubah bebas memiliki korelasi tinggi memberikan hasil yang cukup baik dibanding kuadrat terkecil yang memberikan ragam tinggi bagi penduganya. Metode ini tidak lebih baik dibanding *forward selection* (pada saat parameter sebenarnya banyak mengandung nilai nol), juga dibanding gulud (pada saat parameter semua mendekati nilai nol). Ketidaktahuan parameter sebenarnya dan kekonsistenan hasil pendugaan pada semua kondisi menempatkan LASSO lebih baik dibandingkan dengan gulud atau *forward selection*.

DAFTAR PUSTAKA

Breiman L. 1995. Better Subset Regression Using the Nonnegative Garrote. *J. Technometrics* 37(4): 373-384.

Draper NR, Smith H. 1998. *Applied Regression Analysis*. Ed. ke-3. John Wiley & Sons Inc.

Hastie T, Tibshirani R, Friedman J. 2008. *The Elements of Statistical Learning. Data mining, Inference, and Prediction*. Ed. ke-2. Springer.

Hesterberg T, Choi NH, Meier L, Fraley C. 2008. Least angle and l_1 penalized regression: A review. *Statistics Surveys* 2:61-93.

Izenman AJ. 2008. *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning.* Springer.

Mahmood Z, Khan S. 2009. On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression. *The International Journal of Biostatistics* 5(1), Article 25. http://www.bepress.com/ijb/vol5/iss1/25.

Miller A. 2002. *Subset Selection in Regression*. Ed. ke-2. Chapman & Hall/CRC.

Ryan TP. 1997. *Modern Regression Methods*. John Wiley & Sons, Inc.

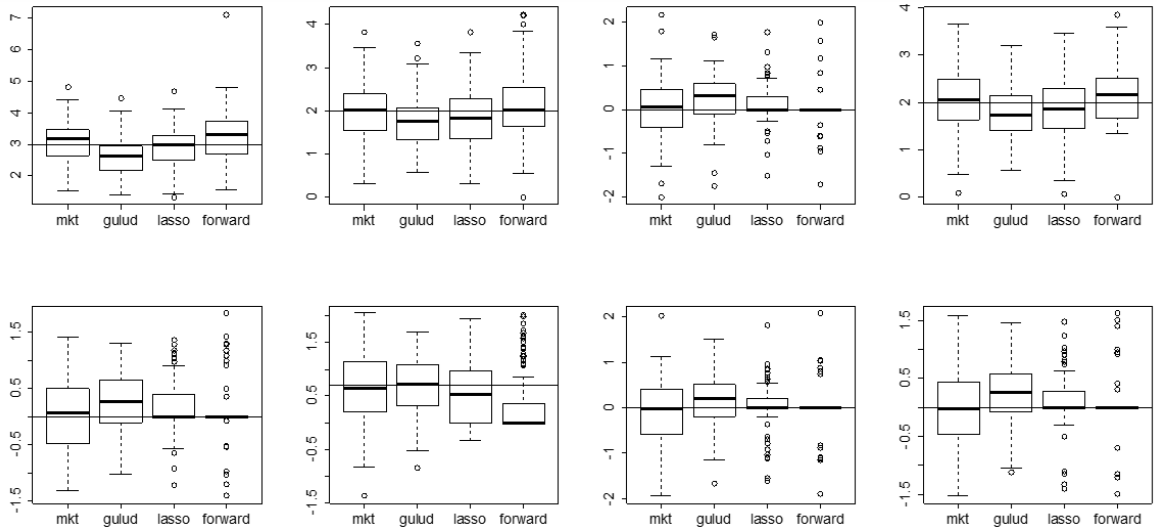
Tibshirani R. 1996. Regression Shrinkage and Selection via The LASSO. *J. R. Statist. Soc (B)* 58: 267-288.

Tibshirani R**.** 2011. Regression Shrinkage and Selection via The LASSO: a retrospective. *J. R. Statist. Soc (B)* 73:273-282.

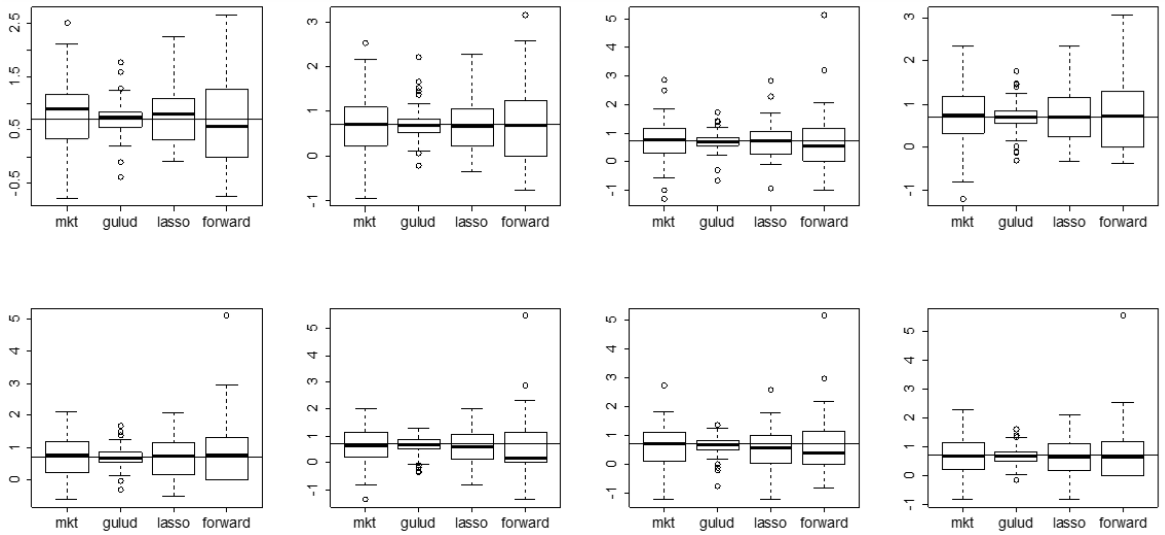
Venables, WN, Ripley BD. 2002. *Modern Applied Statistis with S*. Ed. Ke-3. Springer-Verlag.

.

Lampiran 1. Plot hasil 100 kali simulasi pendugaan untuk koefisien regresi $\beta = (3, 2, 0, 2, 0, 0.7, 0, 0)^t$.



Lampiran 2. Plot hasil 100 kali pendugaan untuk koefisien regresi $\beta = (0.7, 0.7, 0.7, 0.7, 0.7, 0.7, 0.7, 0.7)^t$.



Lampiran 3. Plot hasil 100 kali pendugaan untuk koefisien regresi $\beta = (3, 3, 3, 3, 3, 3, 3, 3)^t$ dari 100 simulasi.