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 know 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, result of this research are that for data with serious multicollinearity (there is high correlation between predictor variables), LASSO tend to have low bias average than PLS in prediction of response variable. OLS method has greatest variance of MSEP , that is most 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, 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.