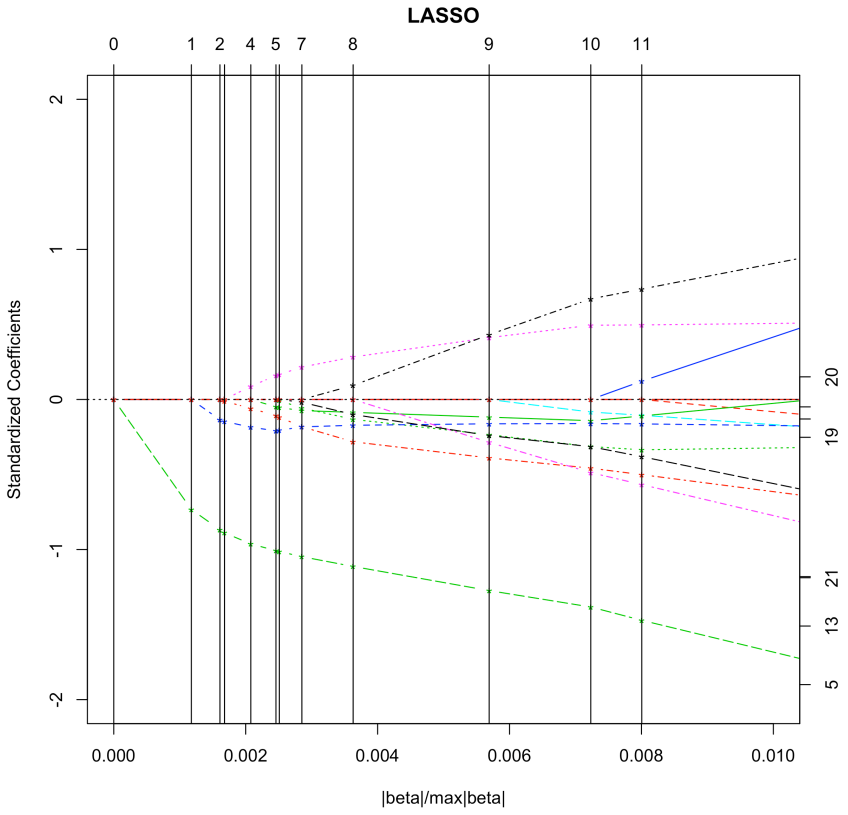
How to interpret the lasso selection plot [duplicate]

This question already has an answer here:
Interpretting LASSO variable trace plots *1 answer*

I did lasso selection using `lars::lars()`, then I got this plot. I have no idea how to interpret it:



Could anyone provide a brief explanation? Why does it plot `standardized coefficients` against `|beta|/max|beta|` ?

r | data-visualization | interpretation | lasso

edited Dec 5 '13 at 20:08
gung
64.2k 18 140 264

asked Dec 5 '13 at 20:02
David Z
586 4 14

marked as duplicate by gung, mpiktas, Scortchi ♦, Nick Cox, COOLSerdash Dec 6 '13 at 8:50

This question has been asked before and already has an answer. If those answers do not fully address your question, please ask a new question.

I am a bit fried to write a complete answer right now but for what's worth check the paper "Regression Shrinkage and Selection via the Lasso" by Tibshirani. You are practically looking at how many regression coefficients β you are using (the # of solid vertical lines). More β's enter your regression equation sequentially as your $|\beta|/max(|\beta|)$ increases. – user11852 Dec 6 '13 at 2:36

Thanks for your comments. What I am wondering most is whether this figure tells me the best selected model. In another word, which variables and how may variables should be used in the selected model? – David Z Dec 6 '13 at 3:09

No, it does not help with that. α/model selection is usually done via cross validation. – Affine Dec 6 '13 at 4:24

1 Answer

In regression, you're looking to find the β that minimizes:
$(Y - X_1\beta_1 - X_2\beta_2 - ...)^2$

LASSO applies a penalty term to the minimization problem:
$(Y - X_1\beta_1 - X_2\beta_2 - ...)^2 + \alpha \sum_i |\beta_i|$

So when α is zero, there is no penalization, and you have the OLS solution - this is max $|\beta|$ (or since I didn't write it as a vector, max $\sum |\beta_i|$).
As the penalization α increases, $\sum |\beta_i|$ is pulled towards zero, with the less important parameters being pulled to zero earlier. At some level of α, all the β_i have been pulled to zero.

This is the x-axis on the graph. Instead of presenting it as high α on the left decreasing to zero when moving right, it presents it as the ratio of the sum of the absolute current estimate over the sum of the absolute OLS estimates. The vertical bars indicate when a variable has been pulled to zero (and appear to be labeled with the number of variables remaining)

For the y-axis being standardized coefficients, generally when running LASSO, you standardize your X variables so that the penalization occurs equally over the variables. If they were measured on different scales, the penalization would be uneven (for example, consider multiplying all the values of one explanatory variable by 0.01 - then the coefficient of the OLS estimate would be 100x the size, and would be pulled harder when running LASSO).

answered Dec 6 '13 at 2:47

Affine
1,414 9 17