Stat 435 Project 1

Ramyasai Sanjita Bhavirisetty (011713976)

#Introduction

Linear model selection befits the data set containing multiple potential variables. Subsequently the performance of the model is evaluated by goodness of fit as well as complexity of the model. Subset of predictors from potential variables might be a good model for prediction. The model might not explain the variability compared to more complex model but its interpretable capturing important variables (as given by domain) (citation). That being said, in regression environment multiple linear model is given by

$$M0=\beta0+\beta1X1+\beta2X2+\beta3X3+\ldots+\beta p-qXp-q+\varepsilon$$

where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma2$, $\beta0$ is the intercept, and $\beta1,\ldots,\beta p$ are the regression coefficients and X1,X2…Xp are p predictions modeled for Y response. The two important parameters to evaluate models (citation G. James et al., An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 6,)conditioned when true relationship between predictors and response Y is linear.

a) Prediction Accuracy: Mean square error (MSE) when model is used to predict response based on test observations determines the prediction accuracy. This is also referred as test error of model. R squared can be calculated for prediction based on MSE to provide prediction accuracy as a comparative index.

b) Model Interpretability: Variables in model should have relevant meaning associated with response often domain originated. Irrelevant variables in model might increase prediction accuracy but fails in interpretability of model. Hence its overly important to choose the right bias-variance trade off by selecting right model with right predictors. There are classical/modern methods (citation chapter 6) for feature or variable selection

c) Subset selection: Identifying appropriate subset of variables by best subset, forward stepwise selection and backward stepwise selection. Least square estimates (LSE) of coefficients are used to model. LSE predicts well when number of observation is larger than number of predictors with low bias and variance.

ii) Shrinkage: All variables used to model and then their coefficients are shrunken towards zero. Shrinkage, also called regularization, tends to reduce variance. There are two prominent shrinkage techniques used namely ridge regression and lasso.

iii) Dimension reduction: projection of p variables into M-dimensional plane where M<p. This can be done by modeling M unique linear combinations or projections.

This paper panders multiple linear regression estimating “median value of owner-occupied homes in $1000” in “Boston” data set by taking into account best subset selection, ridge regression and lasso using R.

#Methods

##Dataset description

Data set `Boston` (which is contained in the library `MASS`). This data set contains 506 observations on 14 variables. The response variable is `medv`, and the rest are potential predictors. Namely, we are interested in predicting `medv` using a linear model.

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524
0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Model assumptions

a) True relation between “medv” and predictors is linear
b) Errors are normally distributed with mean zero and constant variance.
c) No interaction among predictors

Creating training and validation sets

Randomly splitting the observations into a training set and a validation set, so that the training set can be used to fit a linear model, and the validation set can be used to evaluate the prediction accuracy of the fitted model. The splitting has been done 50:50 into non intersecting training and validation set.

Best subset selection

Best subset selection on all potential predictors without interactions between them by applying regsubsets() on training set and based on Bayesian Information Criterion (BIC) a subset of predictors determined. A linear model created with training set and the predictors suggested by best subset selection. Diagnostics were plotted along with correlation matrix for predictors. The prediction accuracy of the fitted model on to validation set can be calculated as mean square error and thus R squared.

Lasso

The first step to get optimum tuning parameter (lamda) by implementing 10 fold cross validation on training set. Implementing LASSO with the optimal tuning parameter on all potential predictors in training set without interactions between them. Best model coefficients were determined.Hypothesis testing was conducted for fitted model coefficients using p values provided by either lasso.proj(). Prediction accuracy can be determined by calculating R squared when lasso model implemented to validation set.

Ridge regresion

The first step to get optimum tuning parameter (lamda) by implementing 10 fold cross validation on training set. Implementing ridge regression with the optimal tuning parameter on all potential predictors in training set without interactions between them. Best model were determined. Hypothesis testing was conducted for fitted model coefficients using p values provided by either ridge.proj(). Prediction accuracy can be determined by calculating R squared when ridge regression model implemented to validation set.

#Results and discussion

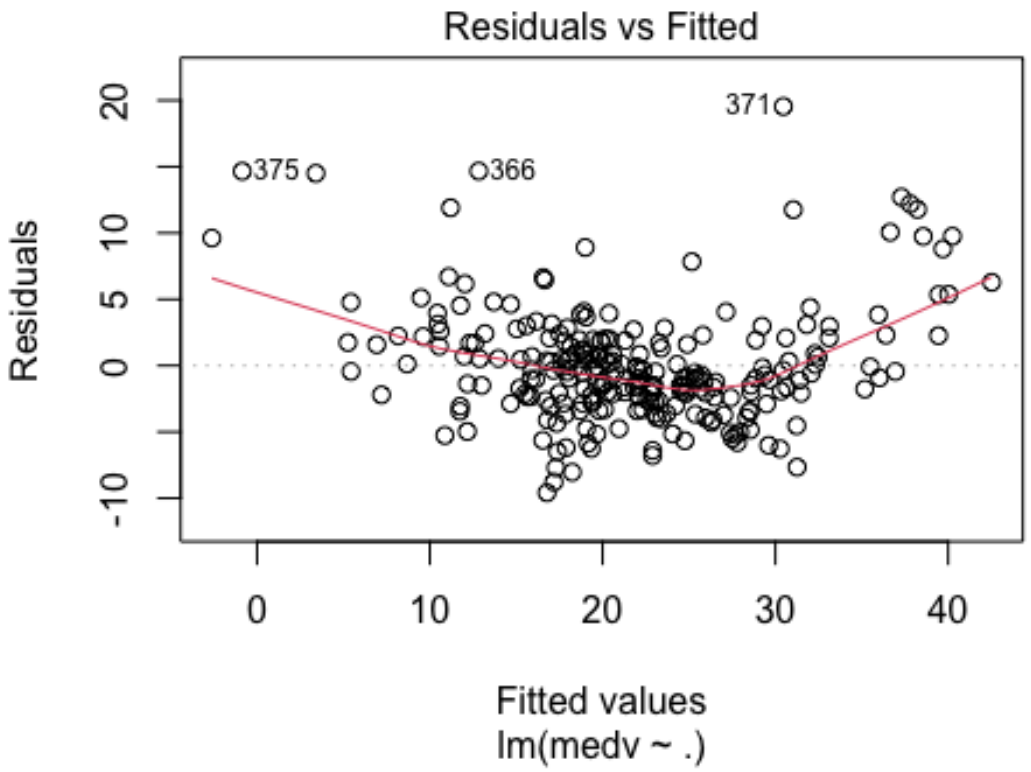
Least square estimates (by best subset selection)

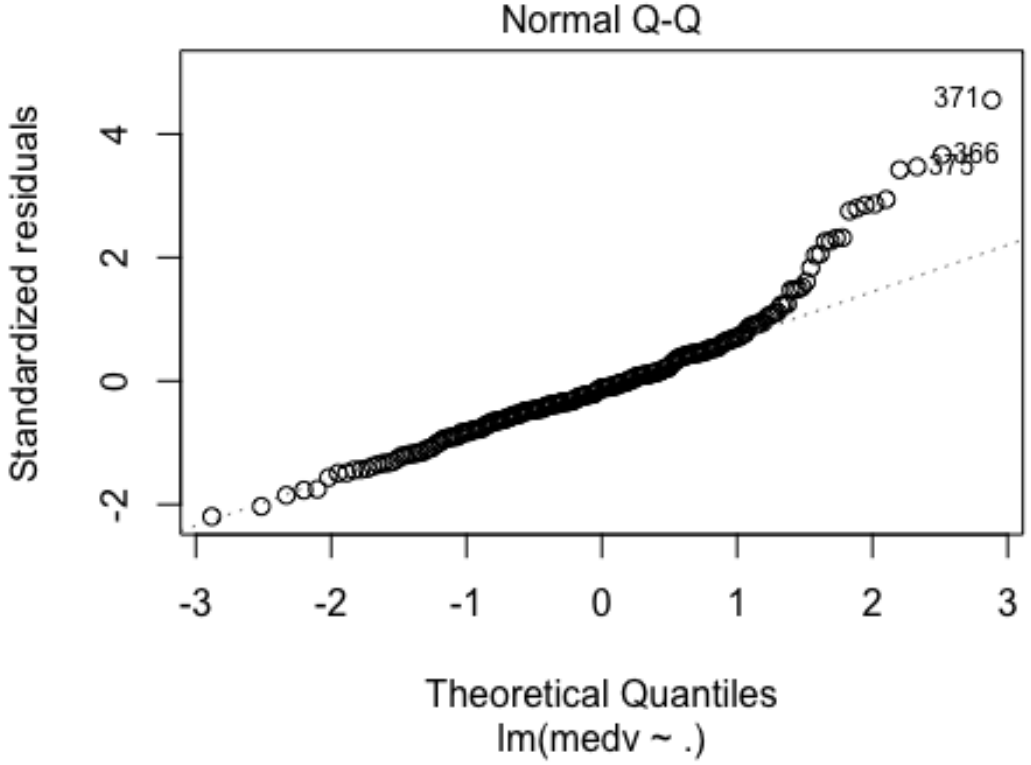
Linear model predictors and their coefficient estimates are as following

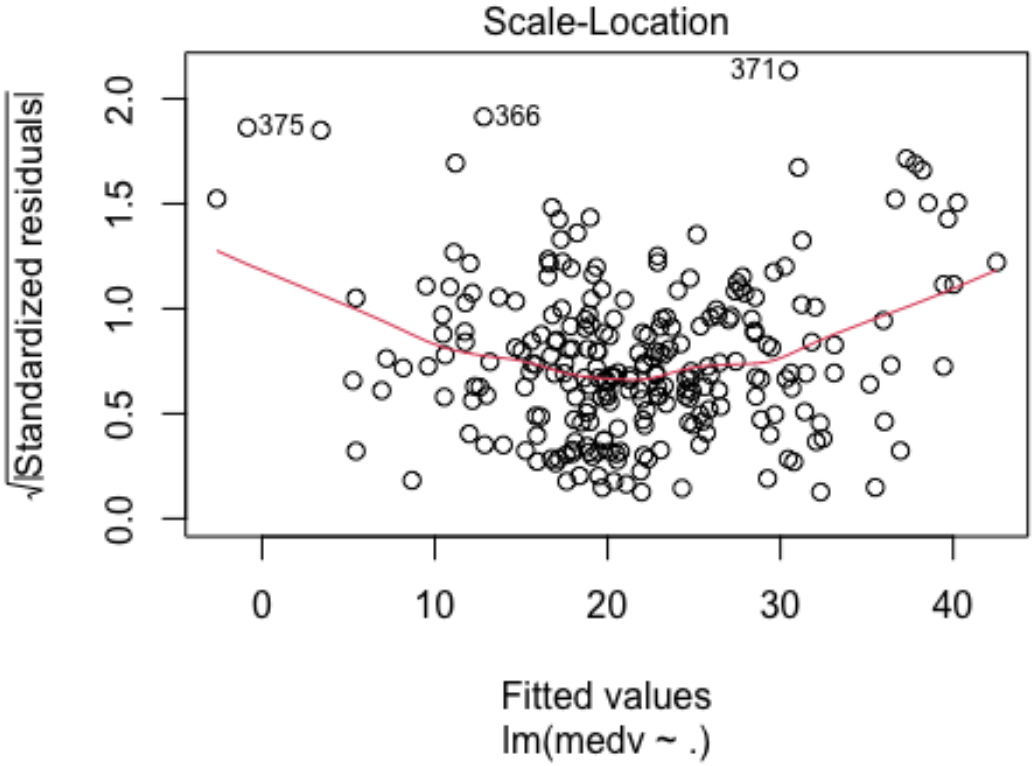
```
##
## Call:
## lm(formula = medv ~ ., data = B.train_LSE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5830 -2.4216 -0.5198  1.9033 19.5385
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.278895   6.547128   6.152 3.11e-09 ***
## crim         -0.106175   0.041483  -2.559   0.0111 *
## nox         -13.141484   4.848690  -2.710   0.0072 **
## rm            4.080221   0.522930   7.803 1.77e-13 ***
## dis          -1.150371   0.226065  -5.089 7.20e-07 ***
## rad           0.374693   0.084893   4.414 1.53e-05 ***
## tax          -0.019405   0.004448  -4.363 1.89e-05 ***
## ptratio      -1.120318   0.165843  -6.755 1.04e-10 ***
## lstat        -0.511106   0.062621  -8.162 1.76e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.403 on 244 degrees of freedom
```

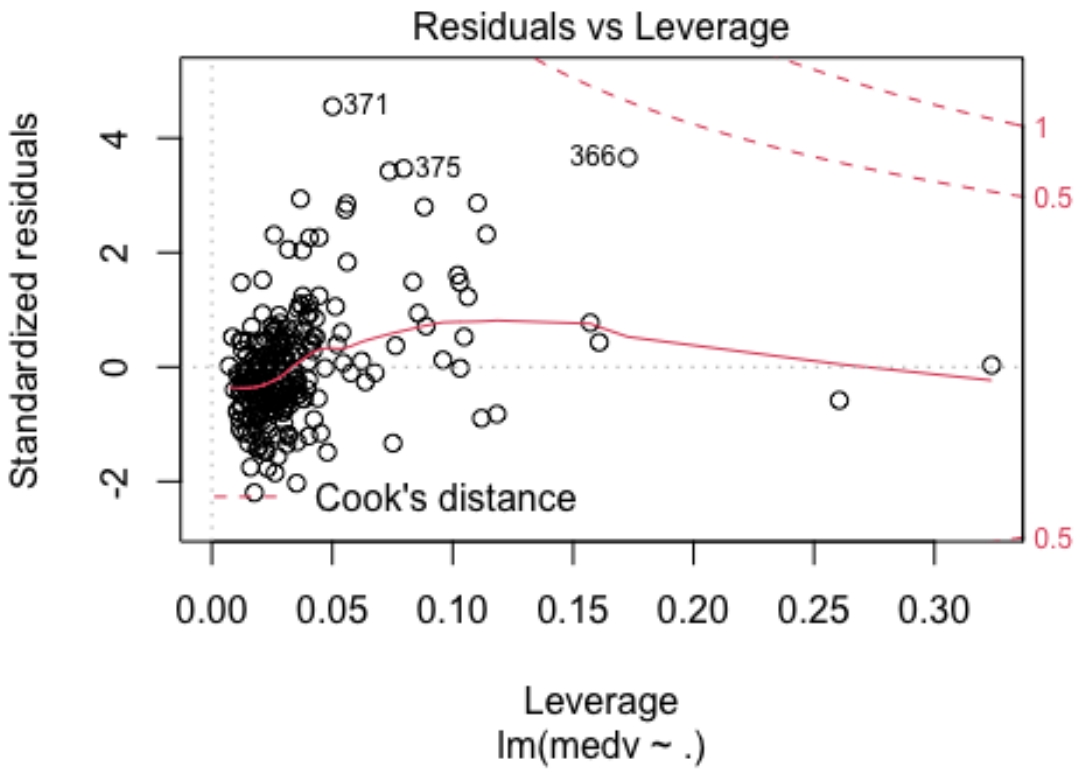
```
## Multiple R-squared:  0.7568, Adjusted R-squared:  0.7489
## F-statistic: 94.92 on 8 and 244 DF,  p-value: < 2.2e-16
```

The training error is given by considering R squared i.e. 0.7568. Hypothesis can be conducted and all the coefficients in the fitted model is significant (pval<0.05) as they were selected based on best subset.The predictors in the model by LSE are crim, nox, rm, dis, rad, tax, ptratio and lstat.


Residuals vs Fitted
Residuals
Fitted values
lm(medv ~ .)
371
375
366
20
10
5
0
-10
0
10
20
30
40

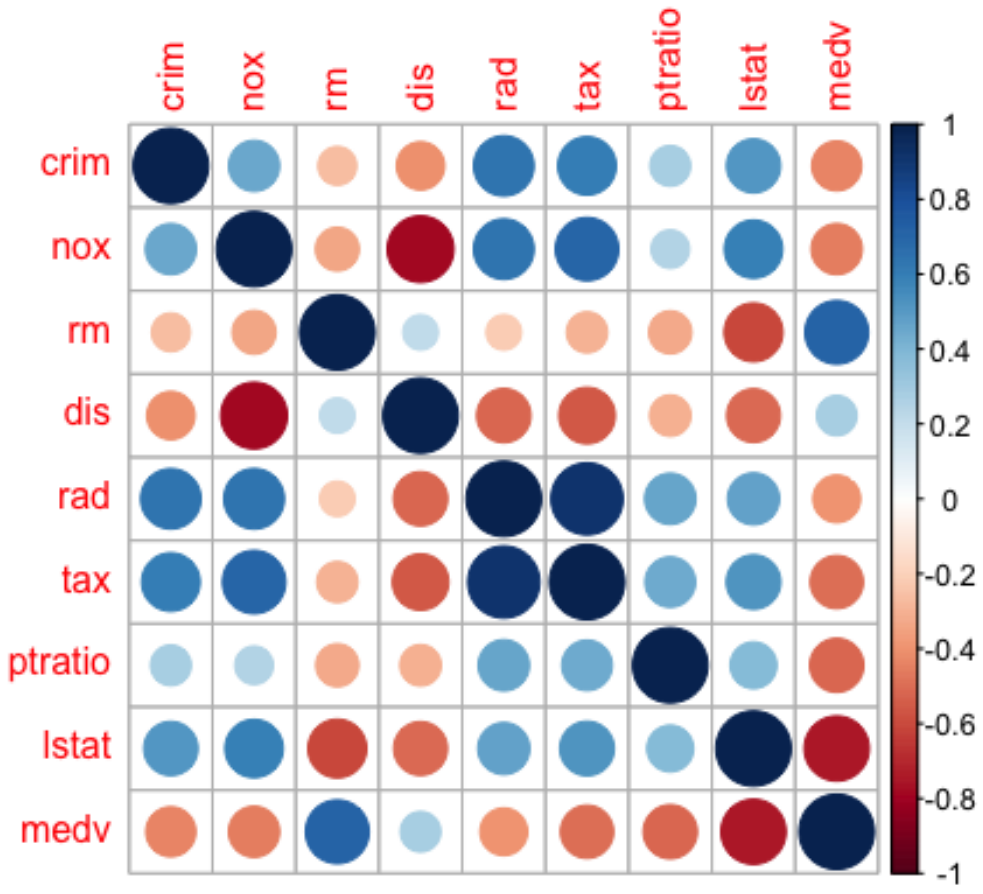

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(medv ~ .)
371
366
375
-2
0
2
4
-3
-2
-1
0
1
2
3


Scale-Location
371
375
366
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Fitted values
lm(medv ~ .)



The model diagnostics reflects violation of model assumptions. Residuals seems to have non-normal distribution with non zero mean.

```
## corrplot 0.92 loaded
```



Even though best subset selection picked the best model predictors based on goodness of fit, from above correlation matrix its evident that there are some variables not independent.R squared for validation set reflecting test error is 0.68. Correlation among predictors inflates variation increasing test error and thus reducing R squared.

##Lasso

Once optimum tuning parameter is obtained, lasso model was created with parameter and their coefficient estimates along with p values as following

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  34.338695732
## (Intercept)   .
## crim         -0.090767768
## zn            0.029263299
## indus         0.028216313
## chas1         2.312776641
## nox         -13.326603598
## rm            4.110453478
## age          -0.003314782
## dis          -1.250296622
## rad           0.371007029
## tax          -0.018725355
```

```
## ptratio      -0.976479058
## black         0.006750755
## lstat        -0.486508539

## Loading required package: scalreg

## Loading required package: lars

## Loaded lars 1.2

## Nodewise regressions will be computed as no argument Z was provided.

## You can store Z to avoid the majority of the computation next time around.

## Z only depends on the design matrix x.

##         crim           zn        indus        chas1          nox
rm
## 7.524148e-02 1.945356e-01 9.131452e-01 2.824889e-02 1.911982e-02
5.106918e-15
##          age          dis          rad          tax      ptratio
black
## 6.651561e-01 7.726536e-06 5.928523e-05 4.279804e-05 1.073130e-07
7.456036e-02
##        lstat
## 3.476851e-12
```

Predictors chas, nox, rm, dis, rad, tax, ptratio and lstat are significant (pval<0.05). Number of predictor is exactly same as best subset selection. R squared for validation set reflecting test error is 0.704.

##Ridge regression

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) 26.391709140
## (Intercept)  .
## crim        -0.074397774
## zn           0.020641787
## indus       -0.023218198
## chas1        2.695141957
## nox         -9.124017836
## rm           4.270708547
## age         -0.009041939
## dis         -0.977644401
## rad          0.178434823
## tax         -0.009530367
## ptratio     -0.864239839
## black        0.006525332
## lstat       -0.437223156
```

```
##         crim           zn        indus        chas1          nox
rm
## 5.927664e-02 1.620466e-01 7.313039e-01 7.709027e-02 2.167633e-02
2.841612e-11
##          age          dis          rad          tax      ptratio
black
## 8.473187e-01 5.212773e-05 1.519228e-04 4.967074e-04 1.289359e-06
9.028845e-02
##        lstat
## 8.446251e-10
```

Predictors nox, rm, dis, rad, tax, ptratio and lstat are significant (pval<0.05). Number of predictors are 7, less compared to both best subset selection and lasso. R squared for validation set reflecting test error is 0.700.

##Prediction accuracy and number of predictors

```
##            Methods Prediction.accuracy Significant.Predictors
optimum.lamda
## 1              LSE               0.680                      8
NA
## 2            LASSO               0.704                      8
0.004
## 3 Ridge regression               0.701                      7
0.483
```

Above table summarizes the linear model based on number of predictors and prediction accuracy. LSE has low accuracy with 8 predictors compared to lasso and ridge as LSE disregards correlation among predictors.

Shrinkage techniques incorporates colinearity in the model thus penalizing coefficients estimates to reduce variance inflated test error. Therefore, Lasso provides higher prediction accuracy of 0.704 even with 8 predictors. Ridge regression has 7 predictors with 0.700 prediction accuracy. Accuracy of ridge regression is similar lasso with a less predictor.

#Conclusion

The overall prediction accuracy is low as model assumptions were not met. Diagnostics plots suggested that error term is non normally distributed with mean non zero. Additionally, there is a possibility of variable interaction which was not considered in this paper. In order to further improve our model, interaction terms needs to be incorporated. As residual vs fitted values plot shows possibility of polynomial relation to response violating the assumption of true linear relation between predictors and response. With the above analysis, LSE lacks accuracy whereas ridge regression lack number of predictors. Based on the paradigm of this paper, a good trade off between bias and variance has been satisfied by lasso given the model assumptions are valid.