Project1-557

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

2/24/2017

Project 1: Techniques for Model Comparison

Introduction

The purpose of this project is to understand the different techniques used to compare models. Specifically, we will look at stepwise, ridge and lasso regression modeling techniques. To accomplish this task, we will utilize the diabetes data set provided by Efron et al. (2003) in the R package "lars." The first section of this project involves a short exploratory data analysis, the second section involves using stepwise, ridge and lasso on the "training" set of the data and on the "test" set. The final section compares the mean squared errors (MSE) of the training and test data on the 3 techniques and selects our best model from the MSE.

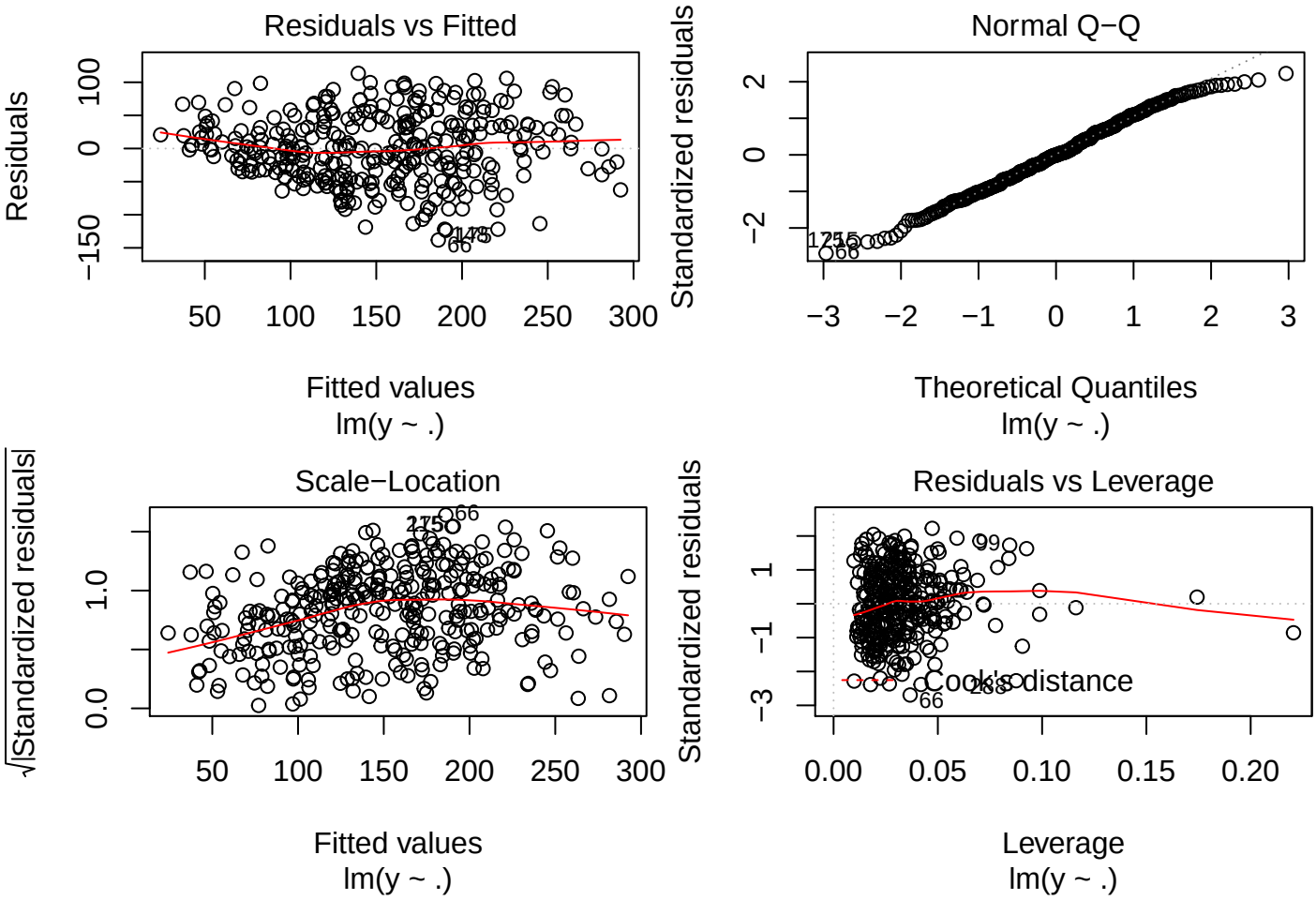
Exploratory Data Analysis

The diabetes sets contains 10 variables, age, sex, bmi, avg blood pressure, and six different blood serum measurements. It takes the 10 measurements on 442 diabetic patients. The repsonse of interest is a quantitative measure of disease progression one year after the baseline. We first take the diabetes data set and partition the data into a training and test data set. Twenty-five percent of the data will be kept as a test data set and the other seventy-five percent will be kept as our training data set. The next step involves building a model of all main effects using least squares on the training data set.

Table 1: Least Squares

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	152.75670	2.878364	53.0706622	0.0000000
age	44.34752	66.836855	0.6635190	0.5074745
sex	-215.18010	68.342766	-3.1485425	0.0017950
bmi	428.44236	74.075817	5.7838357	0.0000000
map	348.44749	76.449555	4.5578746	0.0000074
tc	-790.33439	534.420745	-1.4788617	0.1401579
ldl	545.22541	437.254165	1.2469302	0.2133322
hdl	-23.85126	264.833133	-0.0900615	0.9282945
tch	120.20077	182.367043	0.6591146	0.5102947
ltg	775.87396	207.777028	3.7341662	0.0002228
glu	65.31035	75.118855	0.8694268	0.3852634

Four predictors that are found to be statistically significant in the least squares model are sex, bmi, map and ltg. The predictor sex has a negative coefficient of -215.18, which we interpret as a -215.18 decrease in diabetes progression for a 1 unit increases in the sex coefficient. The coefficients tc and hdl also have negative coefficients, but are not significant to the model. The final step of our EDA will involve looking at the residual plots for the least squares model.



It appears that there are several minor violations of our key linear model assumptions. The data contains a few outliers, which exhibit some leverage on the model as seen in the "Residuals vs Leverage" plot. Also, in the Normal QQ-plot the model produces noticeable tails, which might indicate a violation of our model assumption of normally distributed data as well as a fanning effect in our Residuals vs Fitted values.. The response variable exhibits skewness in its histogram (not shown), which will also contribute to potential model violations.

Stepwise Regression

We perform two types of stepwise regression on the diabetes data set, forward and backwards. Forward selection begins with the null model, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model. Backwards begins with the full least squares model containing all of the predictors, and then removes the least useful predictor one-at-a-time.

Table 2: Forwards StepAIC

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	152.8568	2.867835	53.300401	0.0000000
ltg	507.5118	76.053949	6.673050	0.0000000
bmi	432.3686	72.984996	5.924075	0.0000000
map	368.7778	73.340871	5.028271	0.0000008

	Estimate	Std. Error	t value	Pr(>|t|)
hdl	-391.2643	74.272596	-5.267950	0.0000003
sex	-213.9445	67.623090	-3.163779	0.0017040

Table 3: Backwards StepAIC

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	152.6184	2.863244	53.302602	0.0000000
sex	-198.7019	66.878281	-2.971098	0.0031886
bmi	440.7110	73.075407	6.030907	0.0000000
map	367.3719	73.122053	5.024092	0.0000008
tc	-984.8258	183.238964	-5.374544	0.0000001
ldl	783.3168	171.237948	4.574435	0.0000068
ltg	900.6437	89.776664	10.032047	0.0000000

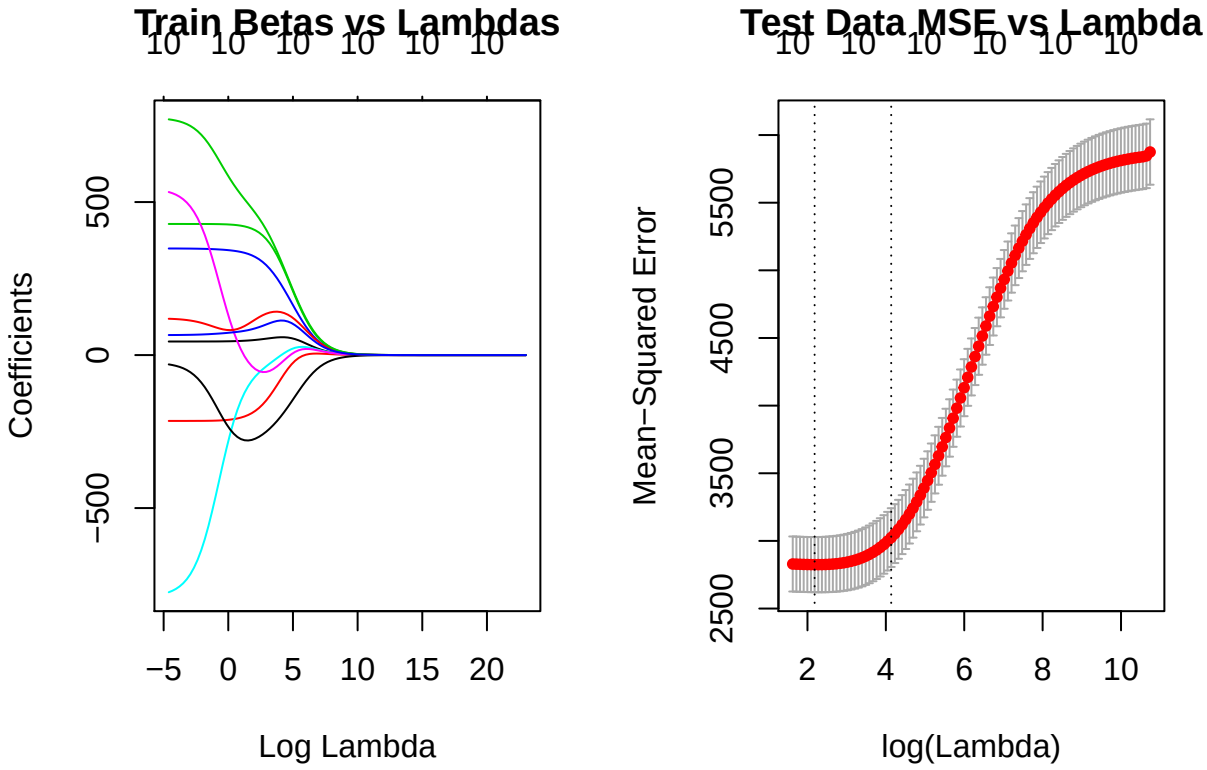
The final model for the forward stepwise contains 5 variables, ltg, bmi, map, hdl and sex, which differs slightly from our original least squares model. The final model for the backwards selection contain 6 variables, sex, bmi, map, tc, ldl and ltg Using stepAIC() in the MASS package, the forward model gave an AIC of 2632.17 while the backwards model gave an AIC of 2631.98. AIC is a measure of the relative quality of models for a set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models, which provides us with a form of model selection. Therefore, the backwards model will be selected as our final model for stepwise selection as it has the lowest AIC.

Ridge Regression

Ridge regression is very similar to least squares, except that the coefficients for ridge are estimated by minimizing a slightly different quantity. In particular, the regression ridge regression coefficient estimates β^R are the values that minimize the following equation:

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

The above equation has a tuning parameter, λ, which helps to address the the bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias. We chose our lambda using the cross validation technique, which is outlined in our textbook as follows: first, we partition the data into complementary subsets, performing the analysis on one subset(training set), and validating the analysis on the the other subset(testing set). Second, to reduce variability, multiple rounds of cross-validation are performed using different partitions, and lastly the validation results are averaged over the rounds. The cv-lambda for our ridge regression model will penalize the coefficients, such that those who are the least efficient in your estimation will "shrink." We can observe this shrinkage in the below plot titled, "Train Betas vs Lambda."



Ridge Model Coefficients

```
(Intercept)         age         sex         bmi         map          tc
   2.025843   49.831137 -179.441395  405.579522  318.237709  -64.397727
        ldl         hdl         tch         ltg         glu
 -49.788492 -271.345405  119.397736  458.556154   88.589295
```

Unlike the model chosen from stepwise selection, and the model chosen from LASSO (reported in the next section), the ridge model does not eliminate any covariates. The plot above for the MSE vs Lambda for cross-validation is shown with all 10 predictors being included and how the MSE changes with the change of lambda. This unfortunately makes model interpretation more difficult. Interestingly, some of the covariates removed from the model chosen by stepwise selection displayed coefficients of larger magnitude in the ridge model.

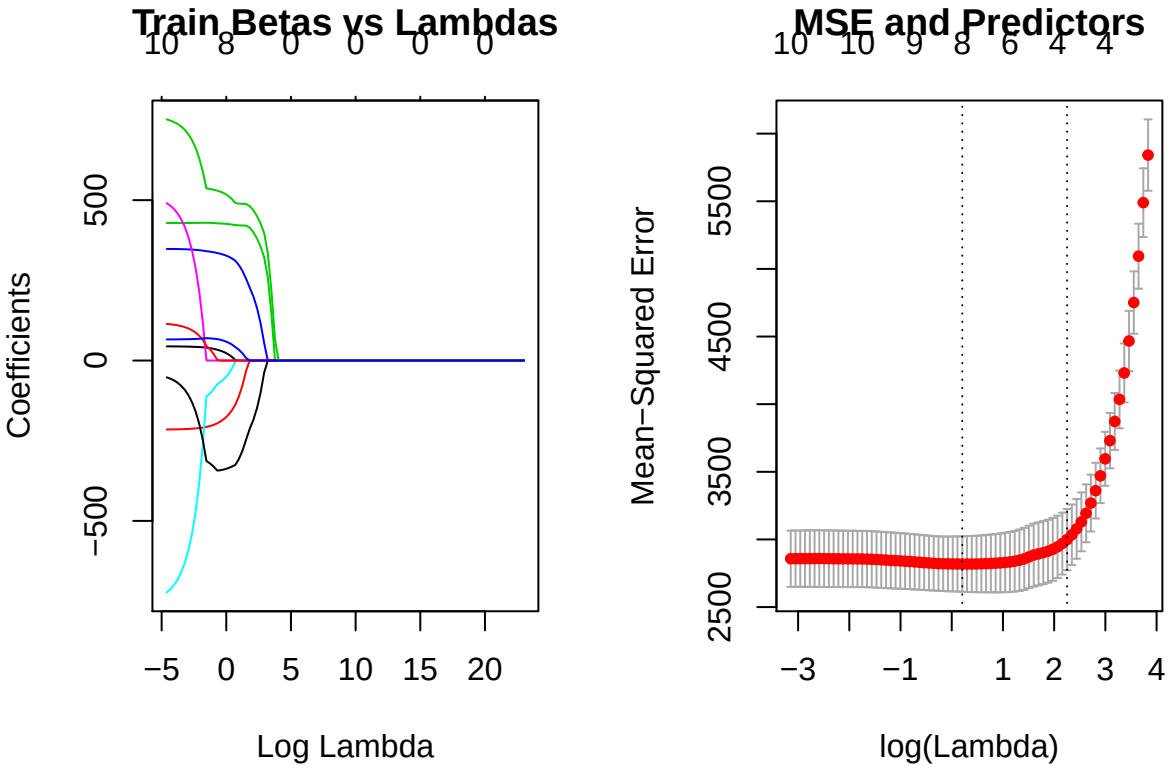
Lasso Regression

Lasso regression is similar to Ridge regression, but it allows for better model interpretability and variable selection. The lasso model coefficients, β_λ^L, seeks to minimize the following quantity:

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

The noticeable difference between the ridge and lasso model is that the β^2 term in the ridge regression penalty has been replaced by a $|\beta|$ term. This new penalty term has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Forcing some of

the coefficients to be zero can lead to better model interpretation and well as reducing the total number of predictors in the model. In the plot, "Train Betas vs Lambda," we can visually see some of the coefficients shrinking very fast to zero, while others linger on before going to zero as lambda becomes sufficiently large.



LASSO penalizes covariates more heavily than ridge, and effectively removed three covariates from the model. The plot "MSE and Predictors" gives us how many predictors we should use in our model. The numbers on top of the figure give the number of non-zero coefficients. Instead of using 10 predictors for the selected model if we would choose the one standard error estimate, then we should use four predictors in our final model. Below you will find the coefficients for the covariates for the lasso model. These are again not, however, the same covariates that were eliminated using stepwise selection.

Lasso Model Coefficients

(Intercept)	age	sex	bmi	map	tc
2.01	18.12	-167.41	425.11	323.36	-39.76
ldl	hdl	tch	ltg		
0.00	-334.64	0.00	511.67		

Mean-Square Error for all Models

Method	Test	Training
Lasso	3798.75	2666.69
Ridge-CV	3811.58	2669.78
Step	8839.62	2658.38

Stepwise AIC-based selection, as evidence by the mean squared prediction error both onto the training and test data sets, proved to be the worst fit. Alternatively, the LASSO and ridge models performed much better. LASSO displays slightly better performance than ridge for both the training and test data sets. Based on this criterion, and considering that a model with less variables is preferable, we select the fit produced by LASSO.

All members of our group contributed equally to this project.