CS498 Homework 5

Xinchen Pan, Fangxiaozhi Yu, Jiefei Shi

March 9, 2017

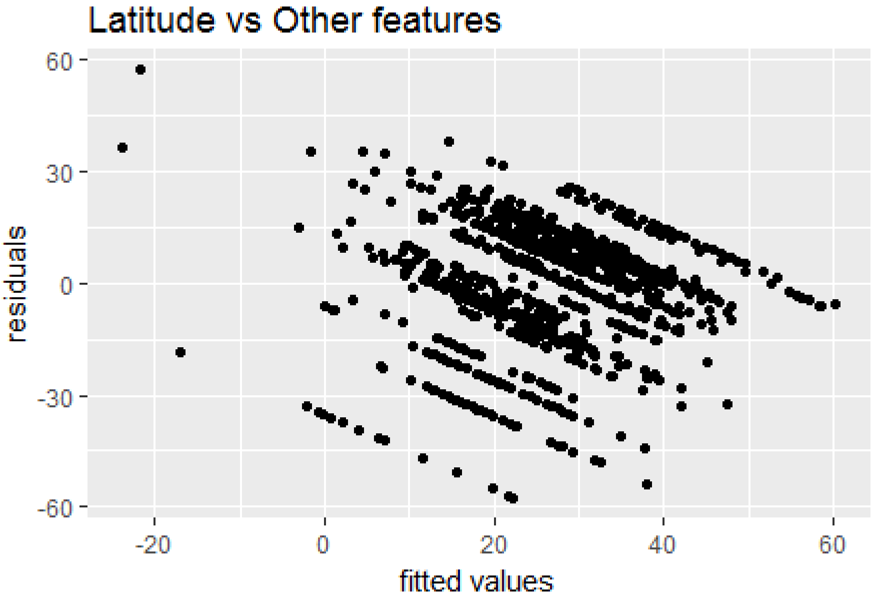
1. Linear regression with variaous regularizers

The dataset we used for this assigment was consisted of a bunch of features of music and we wanted to use these features to predict the latitude as well as the longtitude of the music origin. We chose the dataset with more features. There are 118 variables and 1059 obsevations in the dataset. The 117*th* variable is `Latitude` and the 118*th* variable is `Longtitude`. They are both response variables, the other are predictors.

We firstly built two simple linear regression models. The first regression model is `Latitude` against features. The second model is `Longtitude` against features.

Latitude

The R^2 value for the `Latitude` against features linear regression model is 0.2928092. It indicates our model does not fit the data well. Also from the **Fitted vs Residuals** plot below, we can see that it does not spread equally but having an obvious pattern.



Next, we tried to use Box-Cox transformation to see if it could improve the model performance.

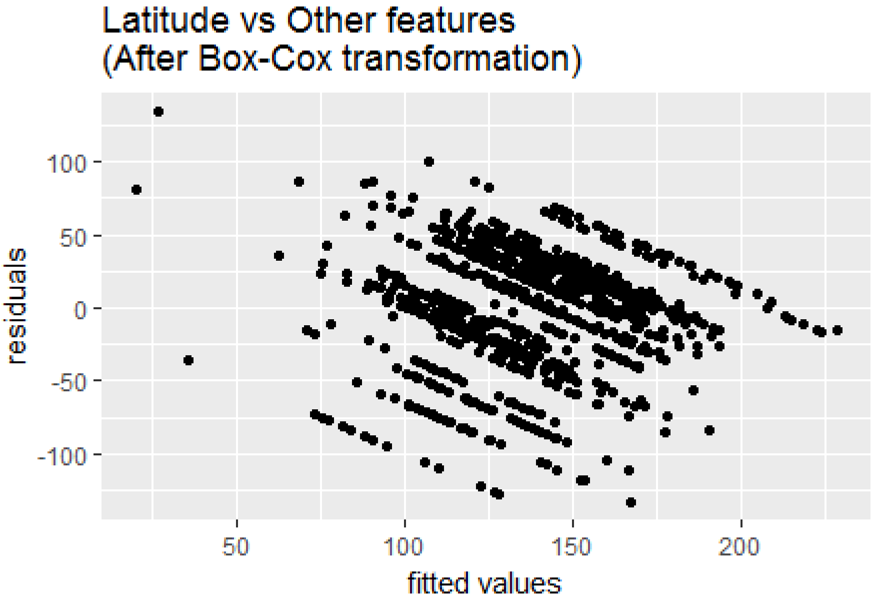
We noticed that there are some negatives values in our response variable which can cause problems when we do Box-Cox transformation. Thus, in order to deal with this problem, we did the following transformation

$$y_{\text{new}} = y - min(y) + \text{some very small number}$$

After the transformation, we guaranteed all values are positive. Then we did Box-Cox transformation using the `boxcox` function from `MASS` package.

We rebuilt the model by taking $y_{new}^{\lambda} \sim x$, $\lambda = 1.191919$ here. The R^2 value increased to 0.2977738. The 0.005 improvement is quite tiny thus we concluded that the Box-Cox transformation was not so helpful here.

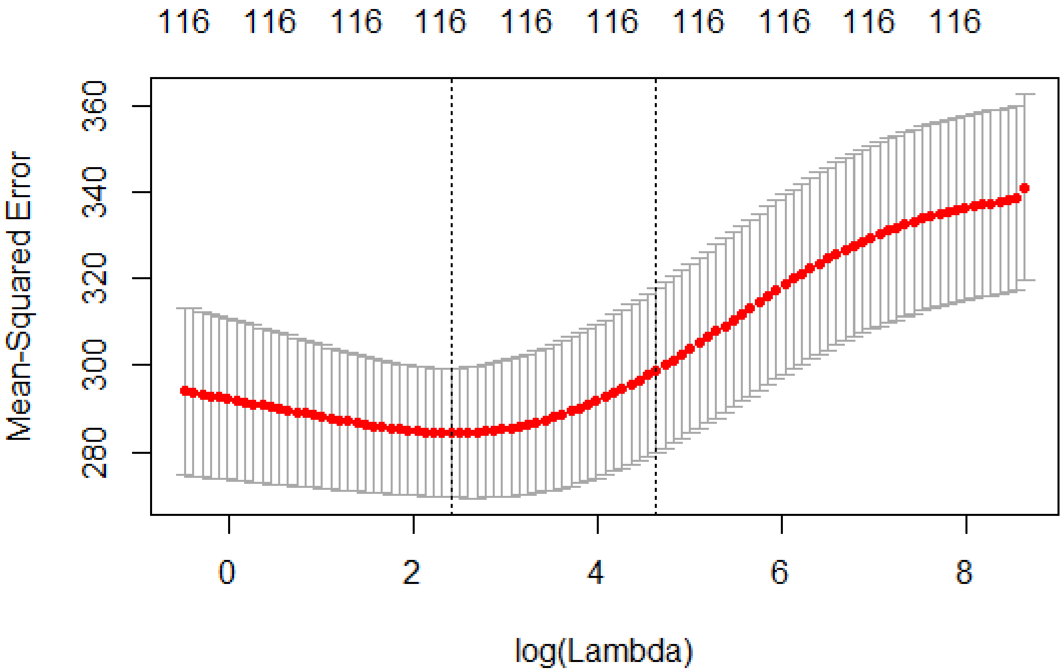
Besides that, the **Fitted vs Residuals** plots also very similar to the plot without transformation. Thus we decided to use raw dat a for the rest of the exercise.



Ridge regression

We used `glmnet` package to do the regularized regressions. We firstly tried a 10-fold cross-validated ridge regression.

The minimum error we got was 278.0356. The mean squared error before was 240.7481. The regularized regression did not behave better than the unregularized one.



Lasso regression

After doing the 10-fold lasso regression, we got a error of 277.7604. It did not beat the unregularized regression either. Only **20** variables were used by this regression.

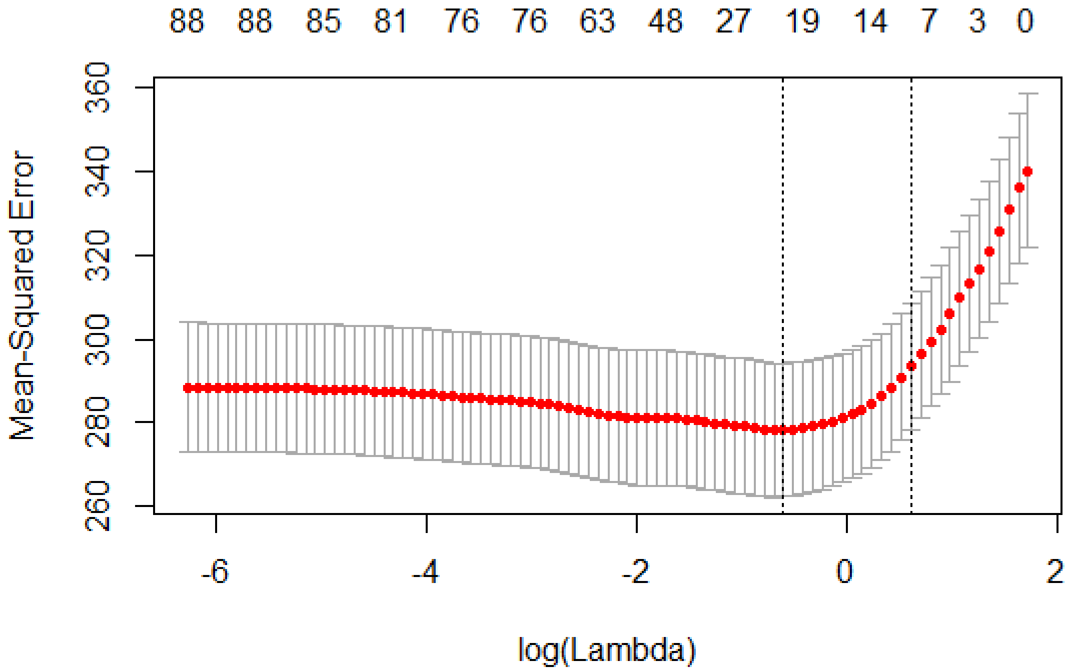
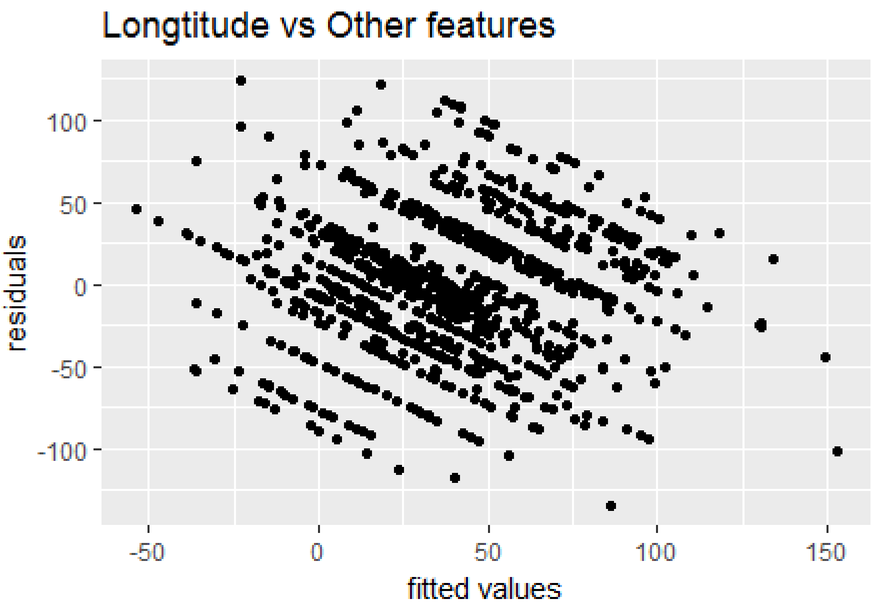


Table 1: Summary Table for Longtitude models

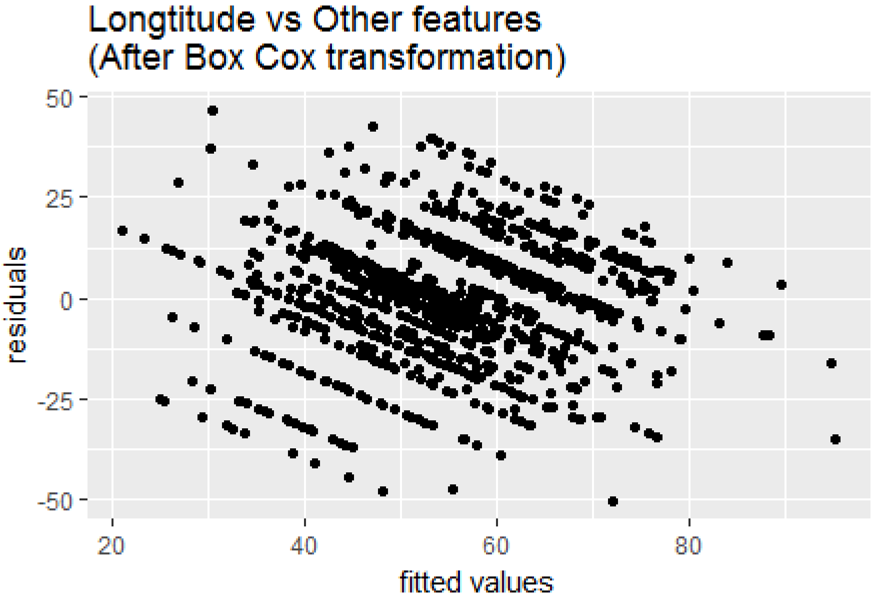
	Unregularized	Ridge	Lasso
MSE	240.7481	280.6861	279.2262
Predictors used	116.0000	116.0000	20.0000

Longtitude

The R^2 value for the `Latitude` against features linear regression model is 0.3645767. It also a fairly small number meaning that the model does not fit the model well. The **Fitted vs Residuals** has an obvious trend too.

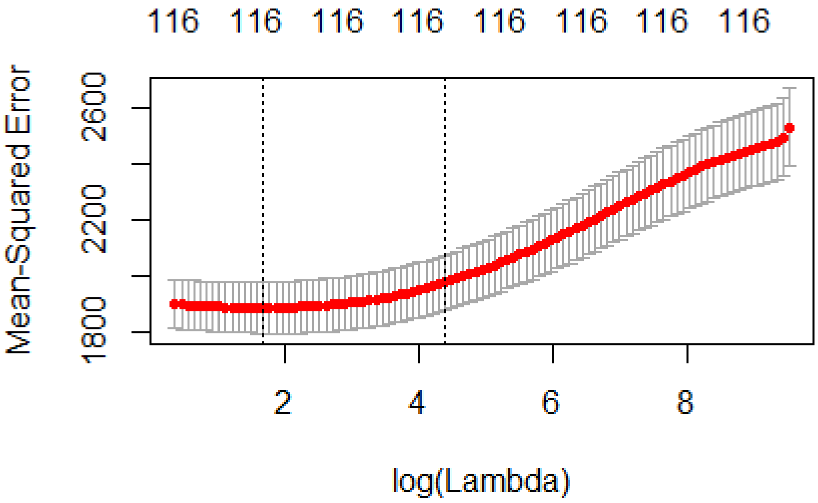


Then we used Box-Cox transformation following the same procedure like above. We transformed y and then rebuilt the model by taking $y_{\text{new}}^{\lambda} \sim x$, $\lambda = 0.8282828$. The R^2 decreased to 0.3616467. The **Fitted vs Residual** plot does not change much. We would again only use the raw data for the regularization exercise.



Ridge regression

We again used 10-fold cross-validation to do the ridge regression. We got a minimum error of 1886.567. Compared to the unregularized model, it hasn't been improved.



Lasso regression

We got a minimum cv error of 1884.248 by using **45** variables. The regularized model is not better than the unregularized one.

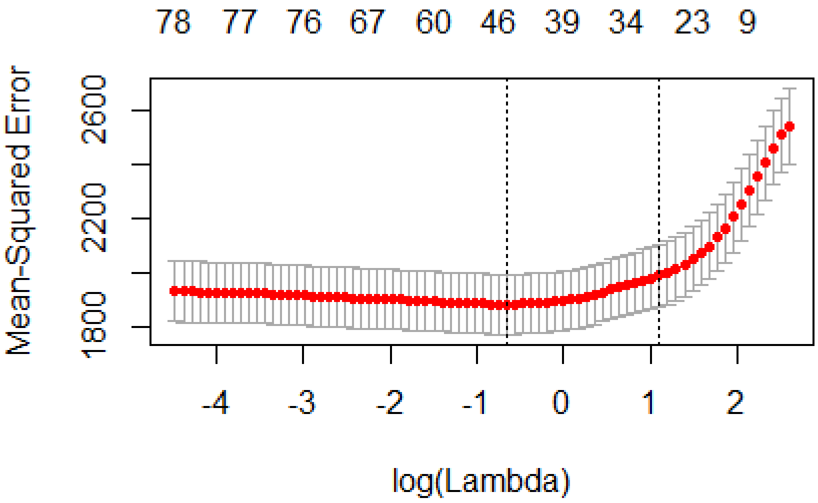


Table 2: Summary Table for Longtitude models

	Unregularized	Ridge	Lasso
MSE	1613.819	1842.147	1884.248
Predictors used	116.000	116.000	45.000

2. Logistic Regression

The default of credit card clients dataset has 30000 observations and 24 variables. Our goal is to predict whether the client will default the credit card payment.

After transformed all the categorical variables by using `factor` function, we built the simple logistic regression model. We used a threshold of 0.5. Any predicited probabilities that is greater than 0.5 will be classified to 1, vice versa.

We got an accuracy of 0.8218333. If we randomly guess a client will default the payment, the probability is 0.7788. Thus our model is a better model.

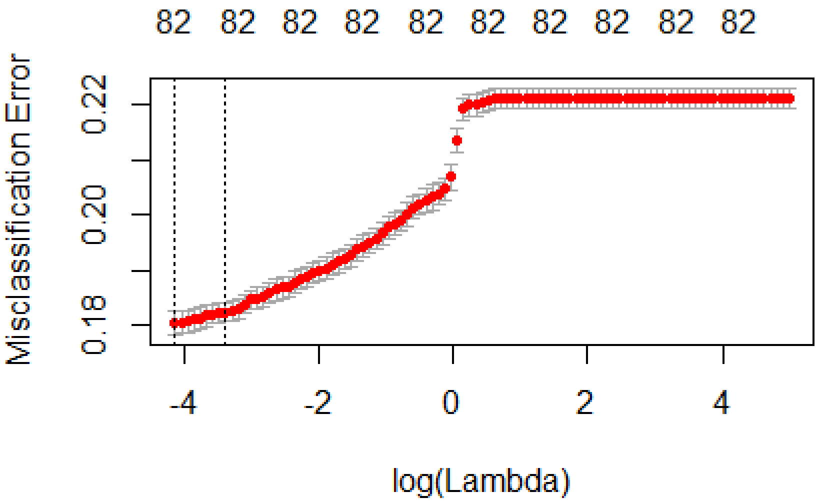
	0	1
0	22280	4261
1	1084	2375

Regularization Method

We then tried regularized method to see if it can improve our model's performances.

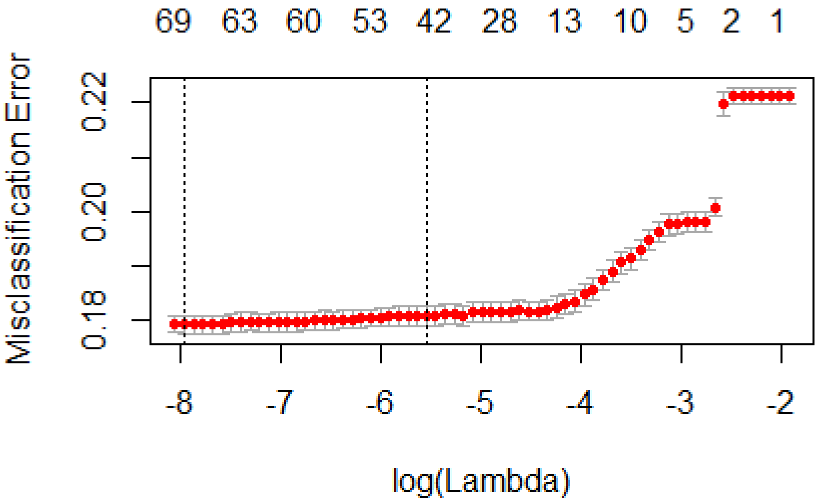
Ridge regression

We used a 10-fold ridge regression and got an accuracy of 0.82 which is lower than unregularized one.



Lasso regression

We fitted a logistic Lasso regression with 10-fold cross-validations. The accuracy is about 0.8208.



Elastic net We fiited a logistic Elastic net regression with 10-fold cross-validations. The accuracy is about 0.8204667.

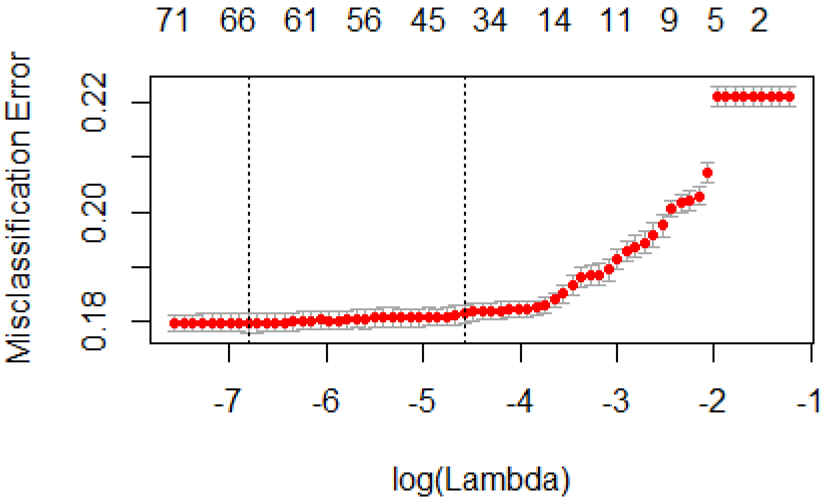


Table 3: Summary Table for Logistic Regression Models

	Random Guess	Unregularized	Ridge	Lasso	Elastic net
Accuracy	0.7788	0.8218333	0.8195333	0.8208	0.8204667

Unregularized method seems to give the best prediction.

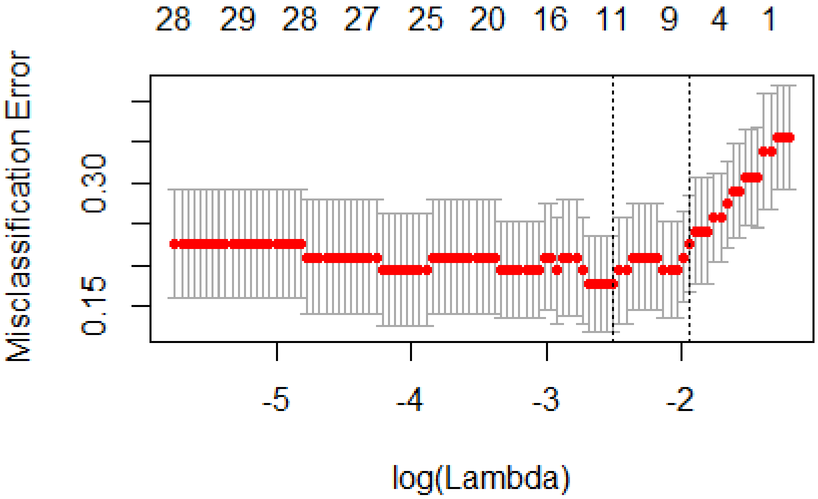
3. A wide dataset, from cancer genetics

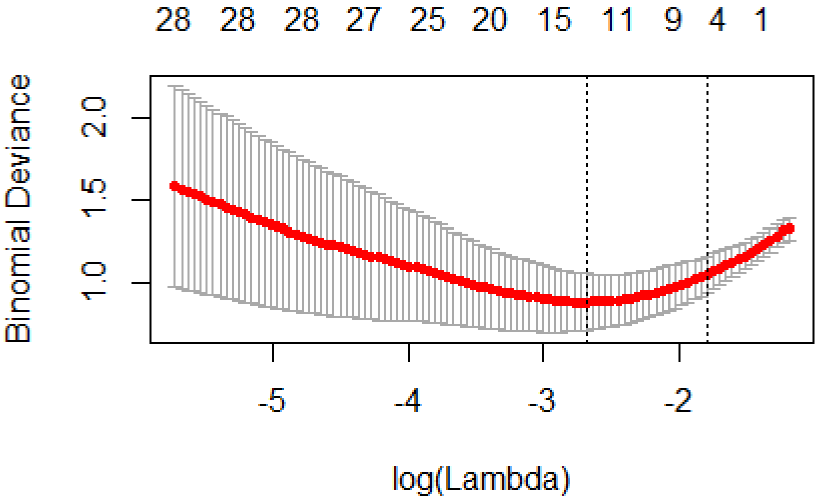
We copied the dataset from http://genomics-pubs.princeton.edu/oncology/affydata/I2000.html to excel and loaded the data into R. Since the 2000 genes are actually the indepenent variables and the 62 tissue samples are observations, we did the transposition to the data. Then we loaded another data which is the response variable. Based on the description, the negative number indicates a tumor tissue and the positive means a normal tissue. We converted the data to 1 and 0 indicating tumor and normal tissues.

Normal	22
Tumor	40

We used a binomial regression model with lasso to do the prediction for the tissue. We used 10-fold cross-validation and got an accuracy of 0.8225806. The deviance is 0.8808075. The AUC is 0.8748976.

A total of 12 genes were used.





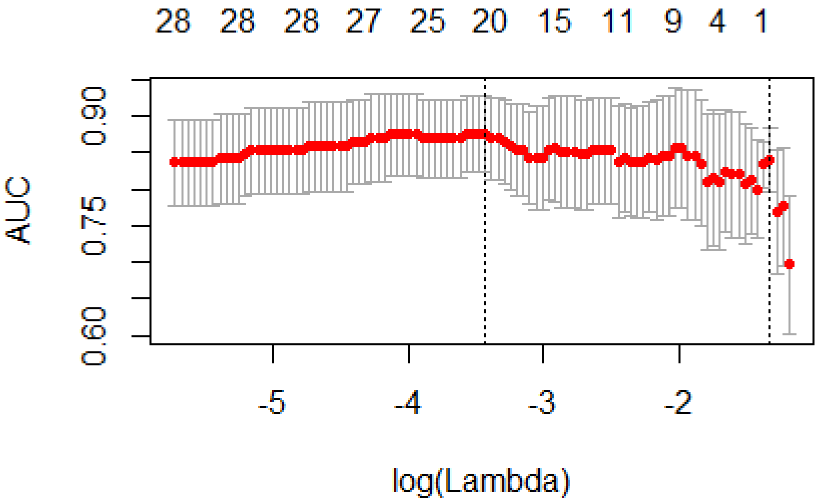


Table 4: Summary Table for Logistic Lasso Regression Model

Measures	Lasso
Random Guess	0.6451613
Accuracy	0.8225806
Deviance	0.8808075
AUC	0.8748976