# HOMEWORK6

**Problem 1, Part 1**

The dataset we used for performing this exercise is about the origination of music based upon a set of features. There are two output variable that we are planning to use latitude and longitude.

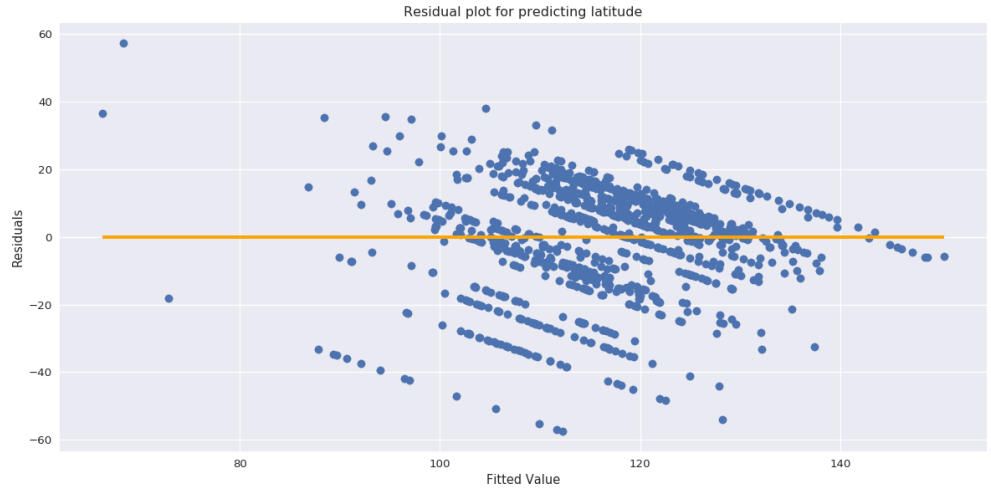
In the first part we create 2 models to perform linear regression one predicting the latitude and other predicting the longitude. Following were the findings:

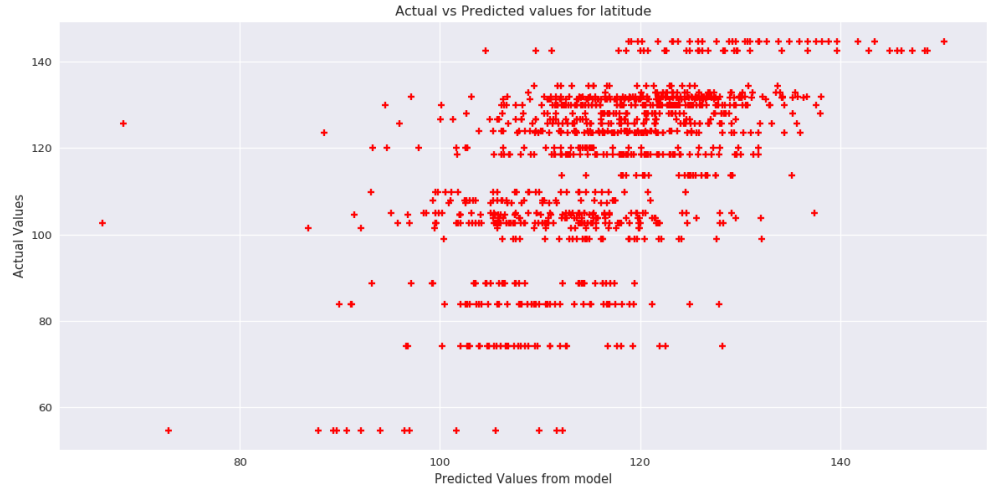
i) Latitude:

***Mean square error for regression model to predict latitude is*:**
**240.747785129**

***R squared value for regression model to predict latitude is:***
**0.292810268928**

***For latitude for regression the aic is = 207.102339607 and bic is = 6615.24041204***




The above plots are for residual which shows a lot of variance around the 0 reference line and the actual vs predicted plot does no justice as well as the value does not seem to be aligned with the prediction as for different predicted values we get the same actual values. The r2 reported above also adheres to the low strength generated by model.

ii) Longitude:

***Mean square error for regression model to predict longitude is:*** **1613.81962346**

***R squared value for regression model to predict longitude is:*** **0.364576573072**

***For longitude for regression the aic is = 203.297121139 and bic is = 8630.10359124***


The above plots are for residual which shows a lot of variance around the 0 reference line and the actual vs predicted plot also shows no strength in predictions. It's very much similar to the one we got for latitude but slightly better than that. We have the r2 reported above which is also low but better than the model for predicting latitude.

**Problem 1, Part 2**

We then went ahead trying correcting the skewness in the dependent variable by applying a box-cox transformation to both latitude and longitude. Following are the results:

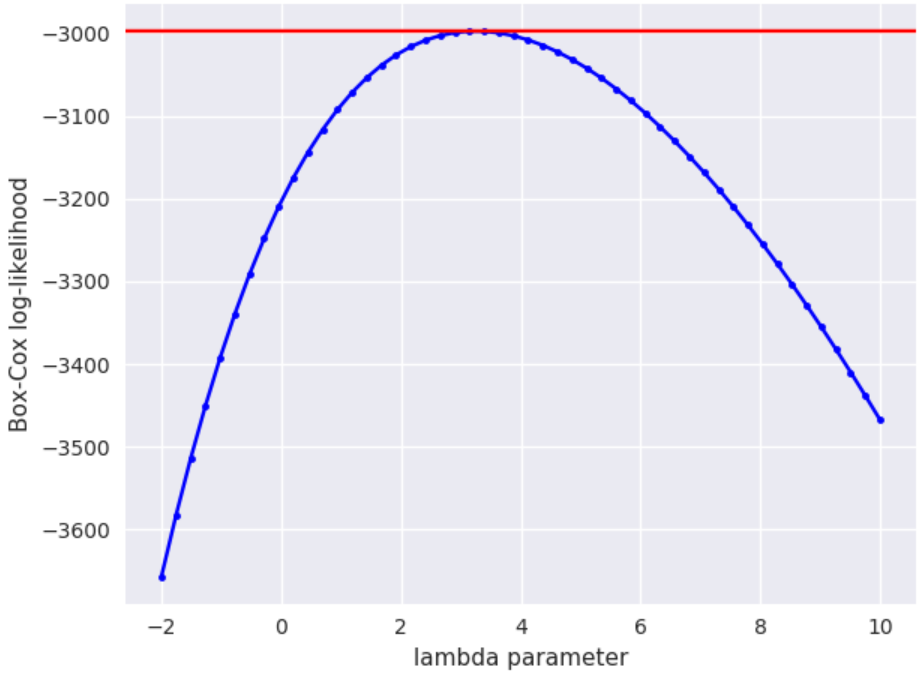
i.) Latitude:

***Lambda values that maximizes the log likelihood of latitude is:*** **3.24706370962**

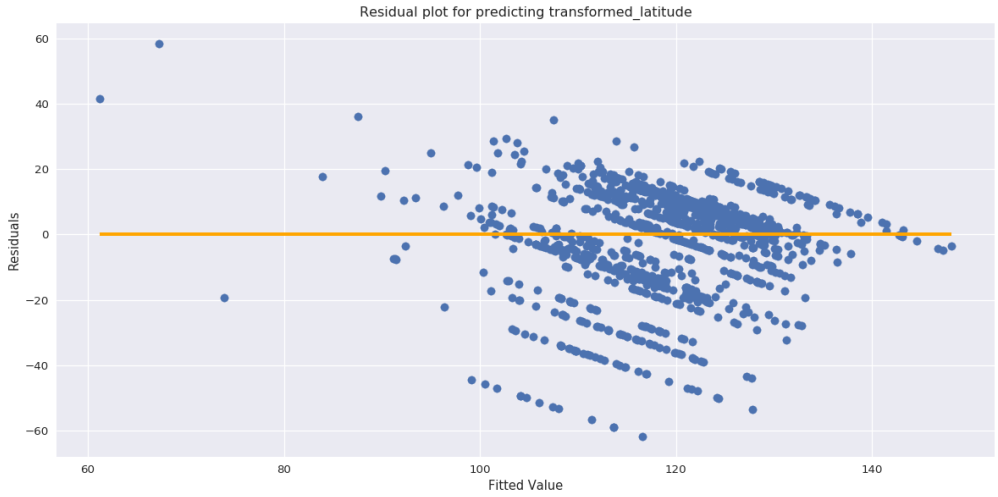
***Mean square error for regression model to predict transformed latitude is:*** **253.536061753**

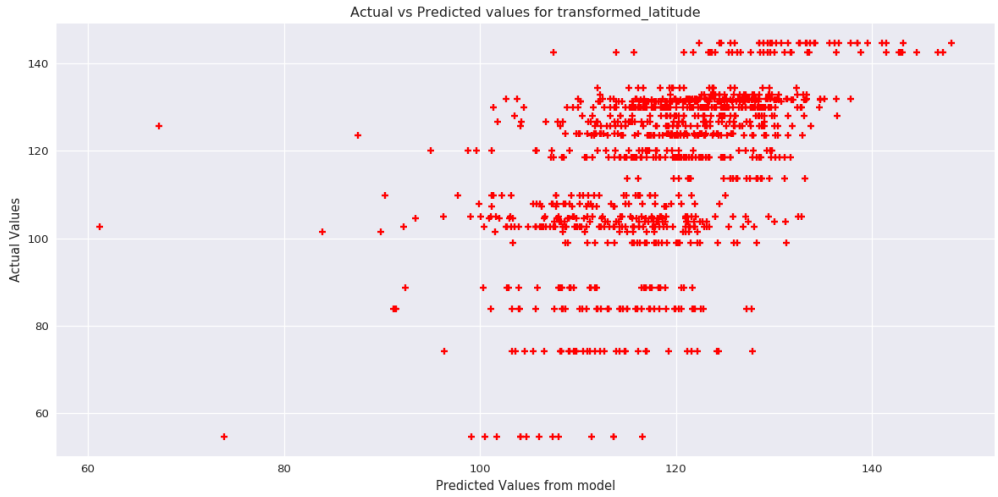
***R squared value for regression model to predict transformed latitude is:*** **0.255245072216**

***For transformed latitude for regression the aic is = 206.998827172 and bic is = 6670.05024647***




The above plot shows the best value to lambda picked by box-cox.




The above residual plot shows a lot of variance and is not completely randomly scattered, also the predicted vs actual plot does not align well. The r2 score reported above decreases. The r2 is calculated by retransforming the prediction back to original space by doing inverse of box-cox transformation so that we have a common ground to compare r2 values among different models**.** Comparing AIC and BIC also shows that this transformation does not help as there is not much change in AIC but BIC is smaller for original model suggesting its superiority. **My takeaway is that box-cox does not help for latitude.**

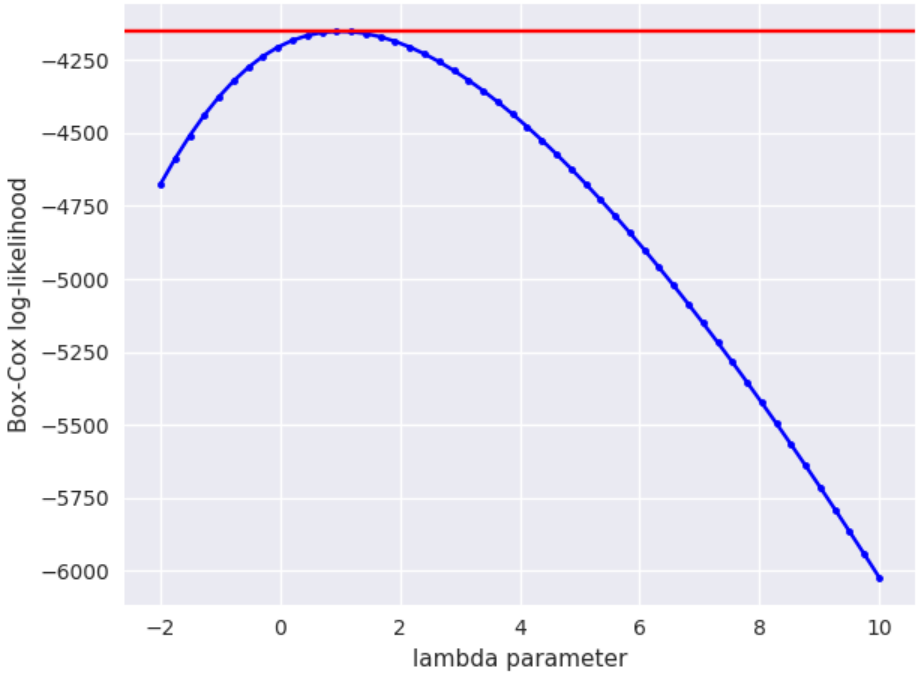
ii.) Longitude

***Lambda values that maximizes the log likelihood of longitude is:*** **1.00076481994**

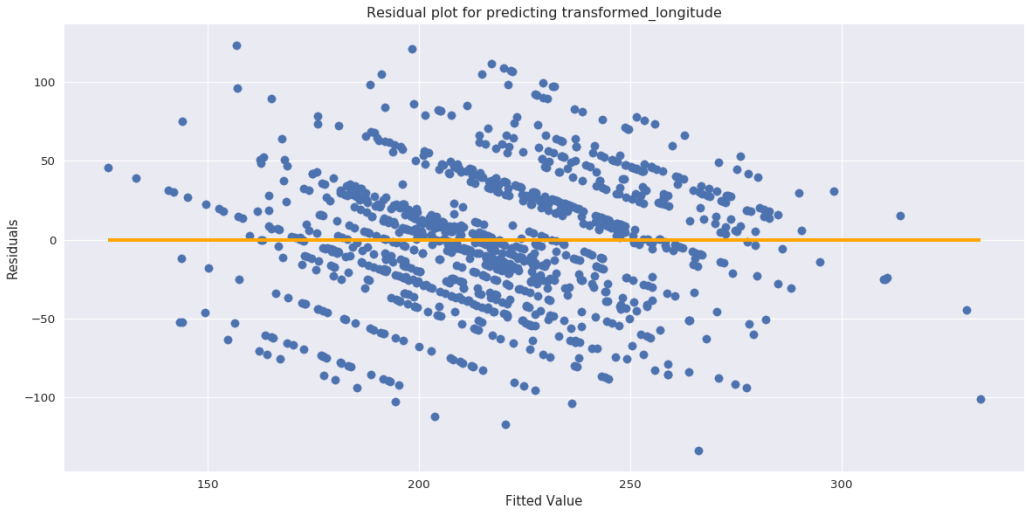
***Mean square error for regression model to predict transformed longitude is:*** **1613.81629695**

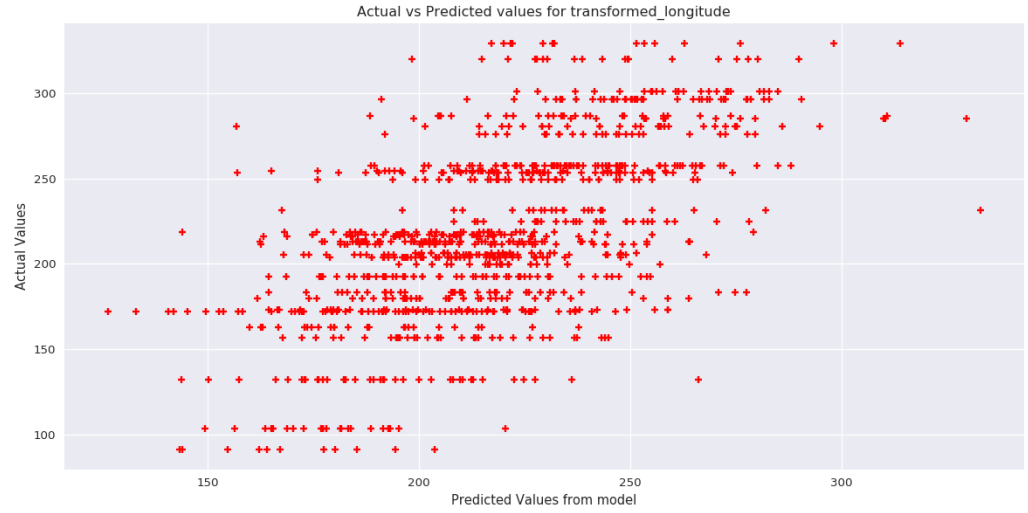
***R squared value for regression model to predict transformed longitude is:*** **0.36457788285**

***For transformed longitude for regression the aic is = 203.297125261 and bic is = 8630.10140835***




The above plot shows the best value to lambda picked by box-cox.




The analysis shows that there is exactly no change in the residual plot as well as predicted vs actual plot, also the r2 values reported above are exactly same as that of original model. The r2 is calculated by retransforming the prediction back to original space by doing inverse of box-cox transformation so that we have a common ground to compare r2 values among different models. The lambda of lambda =1 also means that exactly nothing happens after doing box-cox transformation. The AIC and BIC values are same for original and transformed model. **My takeaway is that box-cox does not help for longitude.**

**Thus neither latitude nor longitude changes after box-cox transformation.**

## Problem 1, Part 3a

We applied the regularization to our simple liner model (without box-cox) to see if something improves, we started by doing ridge regression and following are the results:

## For latitude:

***The value of lambda giving minimum mse for ridge for latitude is [ 4.79698913]***

***The value of lambda giving 1se mse for ridge for latitude is [ 71.23376611]***

***The cross validated error for ridge for latitude is 299.856064553***

***r2 score for ridge regularization for latitude is:* 0.255467573458.**

***Mean square error for regression model to predict latitude with ridge regularization is:* 253.460315912.**

***The no of explanatory variables used for ridge regularization for latitude is:* 116**

## For longitude:

***The value of lambda giving minimum mse for ridge for longitude is [ 3.4016071]***

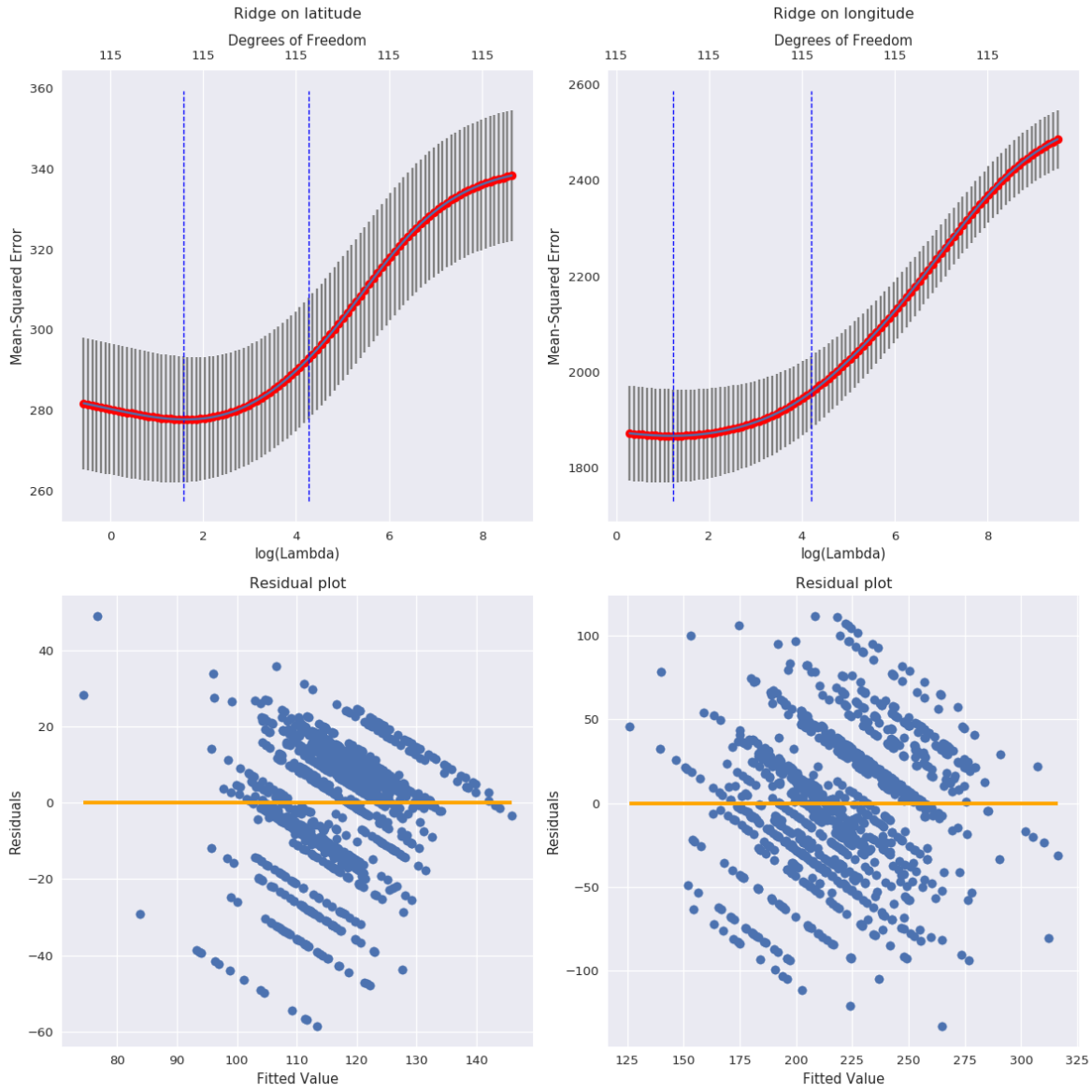
***The value of lambda giving 1se mse for ridge for longitude is [ 66.77493021]***

***The cross validated error for ridge for longitude is 2085.44316473***

***r2 score for ridge regularization for longitude is*: 0.353066999255.**

***Mean square error for regression model to predict longitude with ridge regularization is:* 1643.05111745.**

***The no of explanatory variables used for ridge regularization for longitude is:***
**116**




The above plots show the best value of lambda and the residual plot for both longitude and latitude after applying ridge regression. The residual plot has the same high variance and the r2 score for latitude decreases and for longitude its approx. the same. The mean square error increases which align with the fact about the model not getting improved.

**So this regularized regression is not better than the original one.**

## Problem 1, Part 3b

We applied the regularization to our simple liner model (without box-cox) to see if something improves, by doing lasso regression following are the results:

## Latitude:

***The value of lambda giving minimum mse for lasso for latitude is*: [ 0.45789586].**

***The value of lambda giving 1se mse for lasso for latitude is [ 1.68431486]***

***The cross validated error for lasso for latitude is 286.940190421***

***r2 score for lasso regularization for latitude is*: 0.228992430643.**

***Mean square error for regression model to predict latitude with lasso regularization is*: 262.473218268.**

**The no of explanatory variables used for lasso regularization for latitude is: 14**

## Longitude:

***The value of lambda giving minimum mse for lasso for longitude is*: [ 0.29585444].**

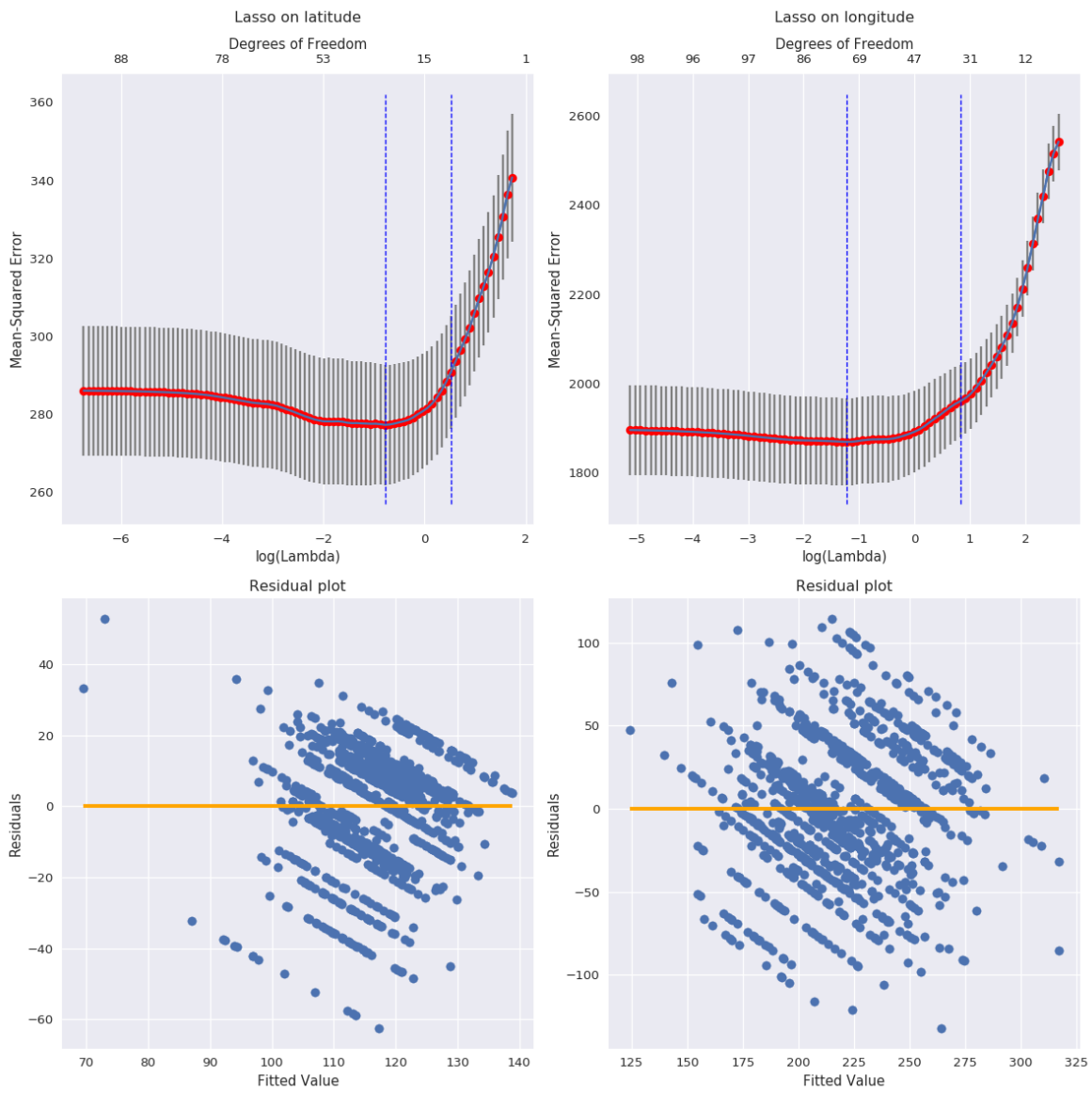
***The value of lambda giving 1se mse for lasso for longitude is [ 2.29069347]***

***The cross validated error for lasso for longitude is 1957.51997032***

***r2 score for lasso regularization for longitude is*: 0.353269940277.**

***Mean square error for regression model to predict longitude with lasso regularization is*: 1642.53569704.**

***The no of explanatory variables used for lasso regularization for longitude is*: 38**




The above plots show the best value of lambda and the residual plot for both longitude and latitude after applying lasso regression. The residual plot has the same high variance and the r2 score for latitude further decreases and for longitude its approx. the same. The mean square error increases which align with the fact about the model not getting improved.

**So this regularized regression is not better than the original one.**

## **Problem 1, Part 3c**

We applied the regularization to our simple liner model (without box-cox) to see if something improves. We used 3 different values of alpha 0.25,0.5,0.75, by doing elastic net regression following are the results:

## **Latitude**:

The value of lambda giving minimum mse for elastic net for latitude at alpha= 0.25 is: **[ 1.6688706].**

The value of lambda giving 1se mse for elastic net for latitude at alpha= 0.25 is **[ 6.13873987]**

The cross validated error for elastic net for latitude at alpha= 0.25 is **287.020932252**

r2 score for elastic net regularization for latitude at alpha= 0.25 is **0.227901033137**.

mse for elastic net regularization for latitude at alpha= 0.25 is: **262.844761463.**

The no of explanatory variables used for elastic-net regularization for latitude at alpha= 0.25 is: **15**.

---

The value of lambda giving minimum mse for elastic net for latitude at alpha= 0.5 is: **[ 0.91579171].**

The value of lambda giving 1se mse for elastic net for latitude at alpha= 0.5 is **[ 3.36862972]**

The cross validated error for elastic net for latitude at alpha= 0.5 is **286.929378951**

r2 score for elastic net regularization for latitude at alpha= 0.5 is: **0.227931063395**.

mse for elastic net regularization for latitude at alpha= 0.5 is: **262.834538297**.

The no of explanatory variables used for elastic-net regularization for latitude at alpha= 0.5 is: **15**.

---

The value of lambda giving minimum mse for elastic net for latitude at alpha= 0.75 is**: [ 0.61052781].**

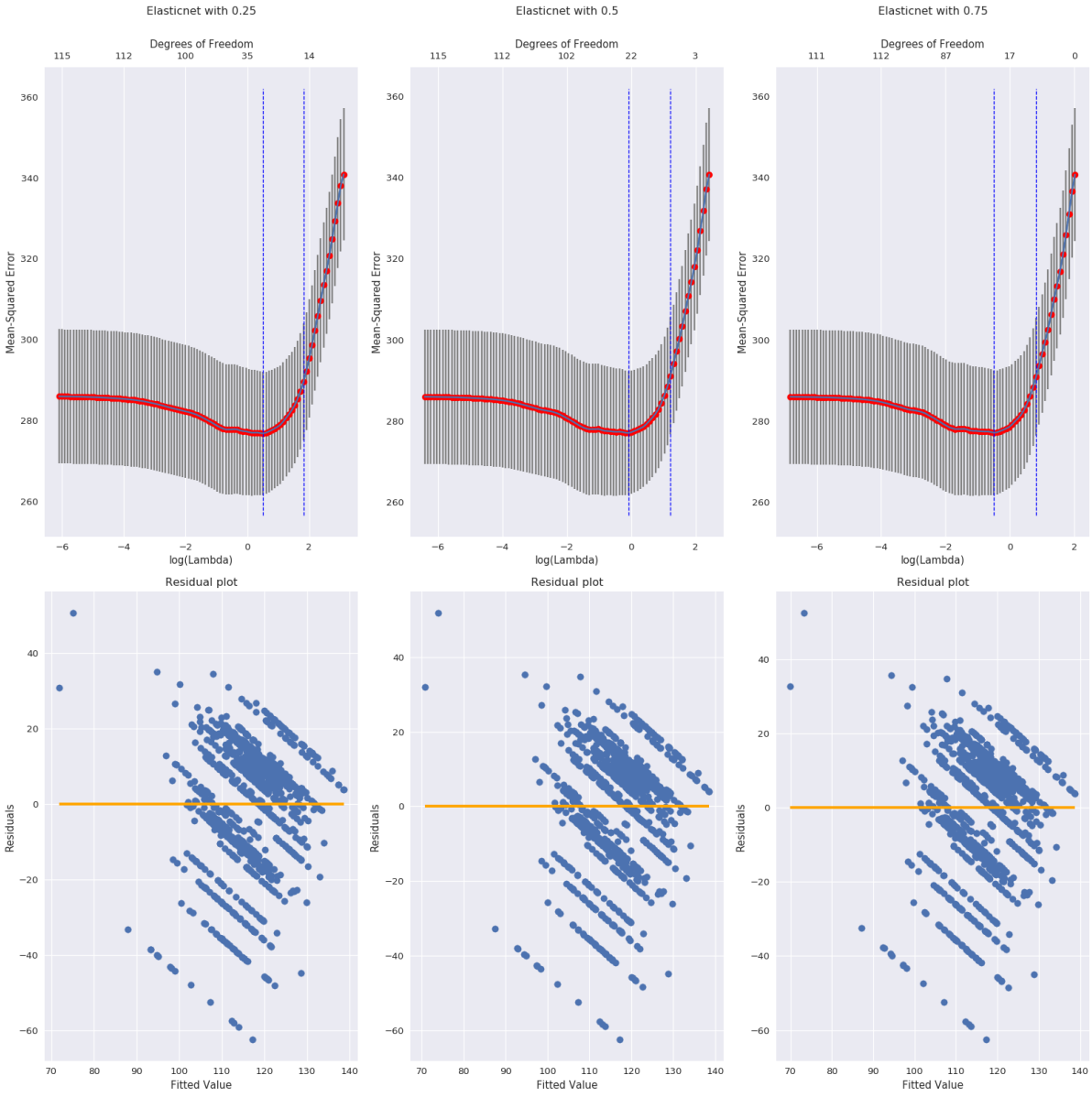
The value of lambda giving 1se mse for elastic net for latitude at alpha= 0.75 is **[ 2.24575314]**

The cross validated error for elastic net for latitude at alpha= 0.75 is **286.903768754**

r2 score for elastic net regularization for latitude at alpha= 0.75 is: **0.228638974691**.

mse for elastic net regularization for latitude at alpha= 0.75 is: **262.593544871**.

The no of explanatory variables used for elastic-net regularization for latitude at alpha= 0.75 is: **14**




## Longitude:

The value of lambda giving minimum mse for elastic net for longitude at alpha= 0.25 is**: [ 0.74322084].**

The value of lambda giving 1se mse for elastic net for longitude at alpha= 0.25 is **[ 8.34877831**]

The cross validated error for elastic net for longitude at alpha= 0.25 is **1960.78148512**

r2 score for elastic net regularization for longitude at alpha= 0.25 is: **0.356685072255**.

mse for elastic net regularization for longitude at alpha= 0.25 is: **1633.86209961**.

The no of explanatory variables used for elastic-net regularization for longitude at alpha= 0.25 is: **74**.

---

The value of lambda giving minimum mse for elastic net for longitude at alpha= 0.5 is**: [ 0.53914309].**

The value of lambda giving 1se mse for elastic net for longitude at alpha= 0.5 is **[ 4.58138694]**

The cross validated error for elastic net for longitude at alpha= 0.5 is **1958.34517342**

r2 score for elastic net regularization for longitude at alpha= 0.5 is: **0.353399390465**.

mse for elastic net regularization for longitude at alpha= 0.5 is: **1642.20692532**.

The no of explanatory variables used for elastic-net regularization for longitude at alpha= 0.5 is: **56**

---

The value of lambda giving minimum mse for elastic net for longitude at alpha= 0.75 is**: [ 0.39447258].**

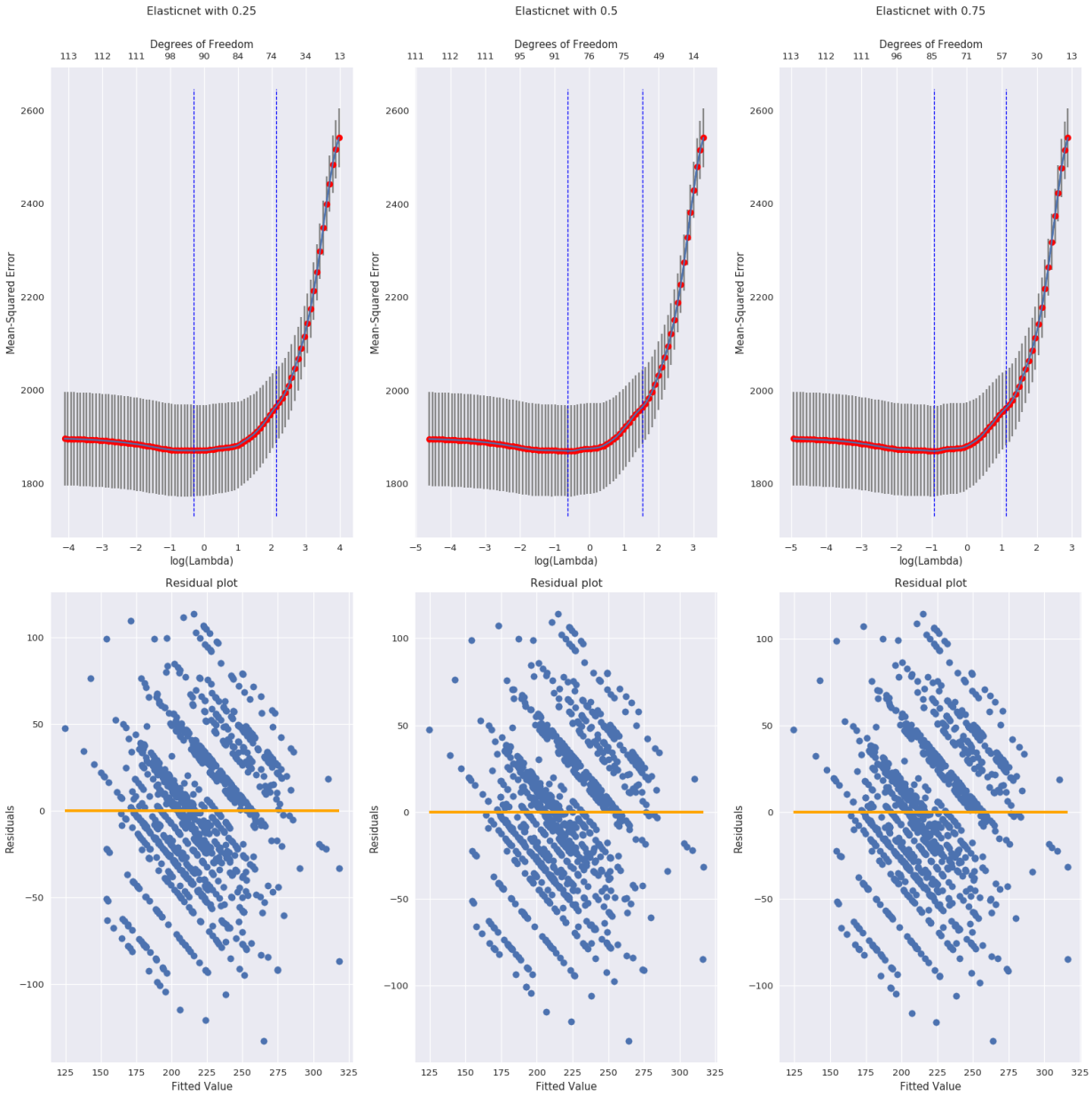
The value of lambda giving 1se mse for elastic net for longitude at alpha= 0.75 is **[ 3.05425796]**

The cross validated error for elastic net for longitude at alpha= 0.75 is **1957.56210874**

r2 score for elastic net regularization for longitude at alpha= 0.75 is: **0.352691021927**.

mse for elastic net regularization for longitude at alpha= 0.75 is: **1644.00600763**.

The no of explanatory variables used for elastic-net regularization for longitude at alpha= 0.75 is: **55**.



**Problem 2**

We used the credit card data set to perform a classification task using logistic regression. We started of first by building a simple logistic regression model by doing a train/test split and we got the following result:

Accuracy score for logistic regression model is: **0.783444444444**

Cross validated error for logistic regression model is **0.221233333333**

We then planned to perform regularization on different values of alpha to see if the result improves or not and we got the following results:

Accuracy score for logistic regression model with regularization at alpha= 0 is: **0.807133333333**

The value of lambda giving minimum mse for elastic net at alpha= 0 is: **[ 0.01348072]**

The value of lambda giving 1se mse for elastic net at alpha= 0 is 0 is **[ 0.01955822]**

The cross validated error for elastic net regularization at alpha= 0 is 0 is **0.214194666667**

---

Accuracy score for logistic regression model with regularization at alpha= 0.2 is: **0.810766666667**

The value of lambda giving minimum mse for elastic net at alpha= 0.2 is: **[ 0.0006899]**

The value of lambda giving 1se mse for elastic net at alpha= 0.2 is 0.2 is **[ 0.006434]**

The cross validated error for elastic net regularization at alpha= 0.2 is 0.2 is **0.202515555556**

---

Accuracy score for logistic regression model with regularization at alpha= 0.4 is: **0.810733333333**

The value of lambda giving minimum mse for elastic net at alpha= 0.4 is: **[ 0.00050046]**

The value of lambda giving 1se mse for elastic net at alpha= 0.4 is 0.4 is **[ 0.00512237]**

The cross validated error for elastic net regularisation at alpha= 0.4 is 0.4 is **0.201471759259**

---

Accuracy score for logistic regression model with regularization at alpha= 0.6 is: **0.8107**

The value of lambda giving minimum mse for elastic net at alpha= 0.6 is: **[ 0.00033364]**

The value of lambda giving 1se mse for elastic net at alpha= 0.6 is 0.6 is **[ 0.00451432]**

The cross validated error for elastic net regularization at alpha= 0.6 is 0.6 is **0.200819248826**

---

Accuracy score for logistic regression model with regularization at alpha= 0.8 is: **0.810766666667**

The value of lambda giving minimum mse for elastic net at alpha= 0.8 is" **[ 0.00027463**

The value of lambda giving 1se mse for elastic net at alpha= 0.8 is 0.8 is **[ 0.00371584]**

The cross validated error for elastic net regularization at alpha= 0.8 is 0.8 is **0.200231924883**

---

Accuracy score for logistic regression model with regularization at alpha= 1.0 is: **0.810833333333**

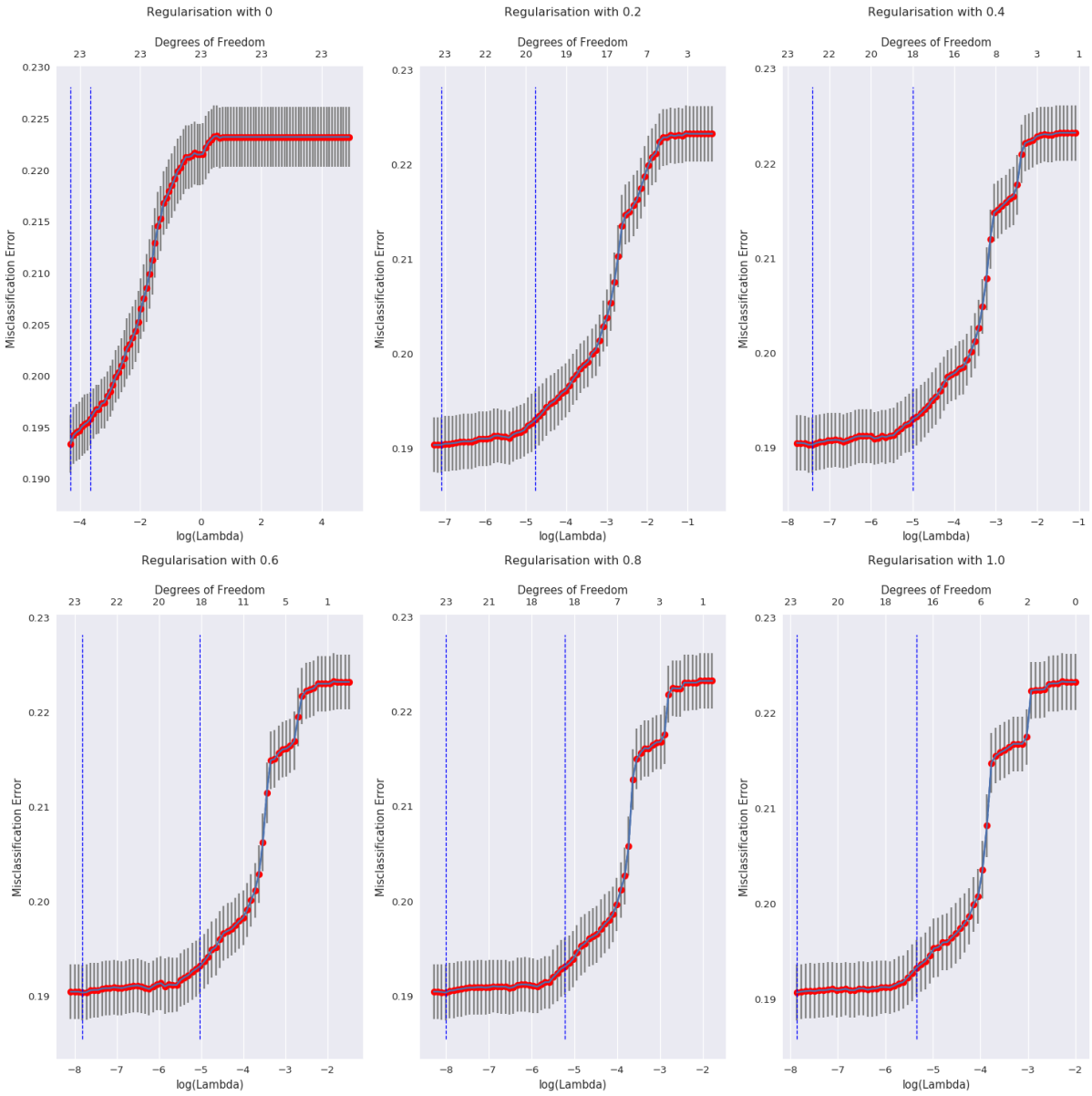
The value of lambda giving minimum mse for elastic net at alpha= 1.0 is: **[ 0.00024112]**

The value of lambda giving 1se mse for elastic net at alpha= 1.0 is 1.0 is **[ 0.00326251]**

The cross validated error for elastic net regularization at alpha= 1.0 is 1.0 is **0.199982857143**

---

We have also created glmnet plot for each alpha values and following are the results:




Looking at the above results we see a slight **increase in the accuracy score after applying regularization**, to be specific at alpha=1, I see the maximum accuracy.

**Note**: The following code uses glmnet package which runs on Linux machine and not macOS or windows as the package has been precompiled for red-hat distributions.

**Citations**: We looked at some stack overflow links for some helper functions, we also went through piazza posts and slack conversations.

*https://medium.com/@rob3hr/lost-when-it-comes-to-multi-linear-regression-988785c3fa55*

*https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html*