Comment and Visualization:

I. Linear Regression

1.1 Straightforward Linear Regression

First, we conducted the straightforward linear regression on Latitude and Longitude against all the features respectively and the R-Square is as below:

Latitude against features	Longitude against features
0.2928	0.3464

Comment:

One issue we notice with the result is that there are many NAs value in the coefficients. This could be because of collinearity of high dimensional independent features. Among the features, we find there are duplication variables which add no additional information to the model. Hence, we try to remove these variables and refit the model again.

To evaluate on the regression, the fitted(predicted) result against the observed values is plotted as below:

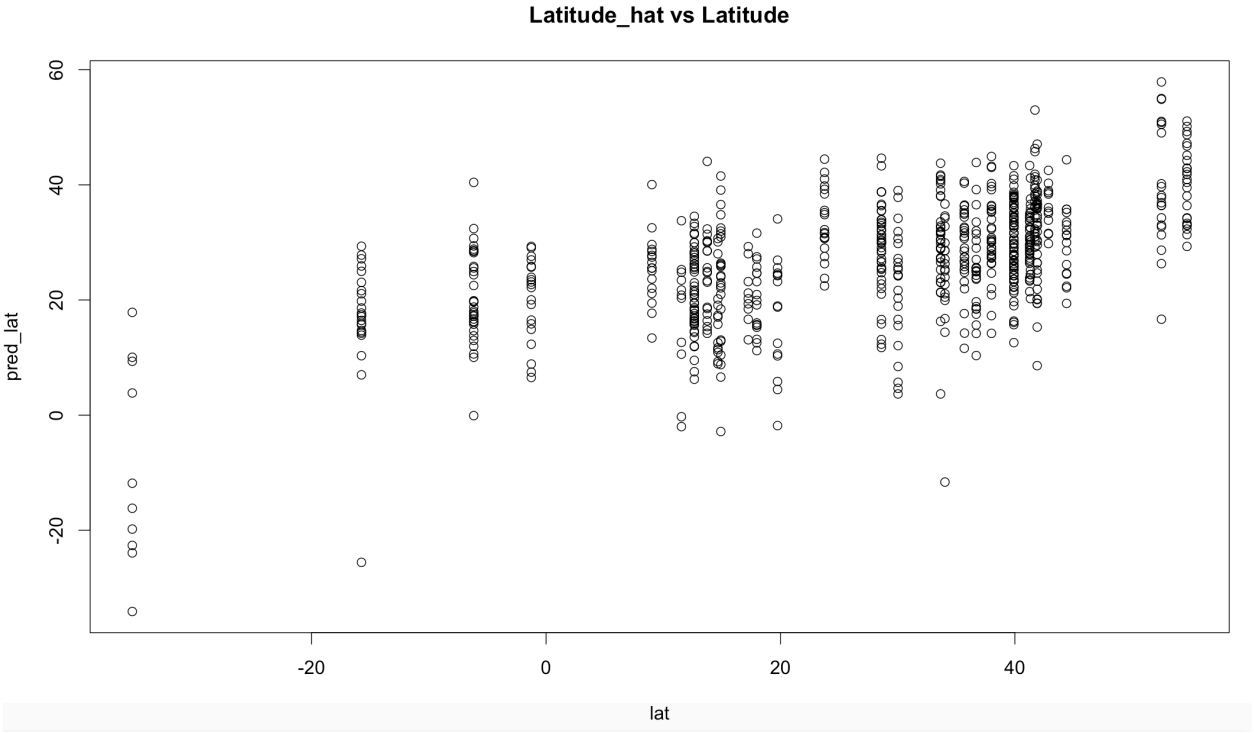


Figure 1 Plot of predicted Latitude vs observed Latitude

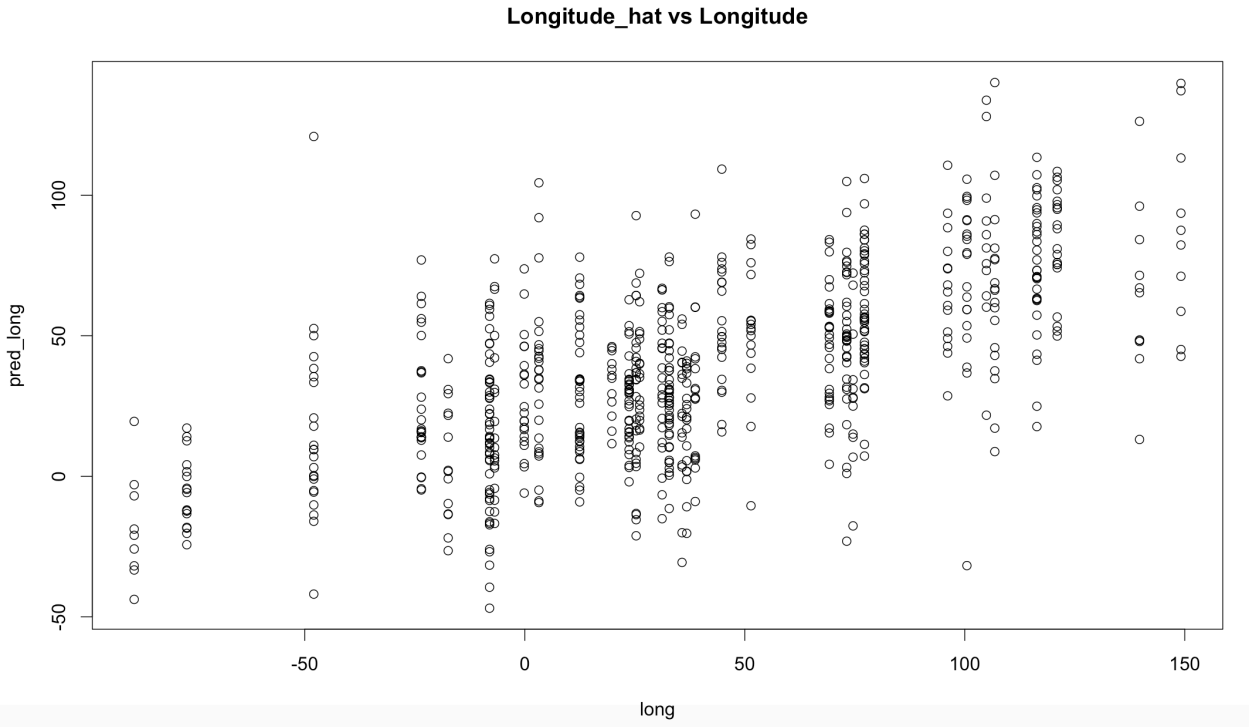


Figure 2 Plot of predicted Longitude vs observed Longitude

Comment:
The above plot implies how good the model is. The y-values are latitude and longitude respectively. In both plots, we can see that the predicted y follows some vertical line patterns. This makes sense because the observed y values are the coordinates of certain cities, which is not completely continuous. On the other hand, the predicted y come from a minimum error model. Overall, the group of lines follows a trend of y = x which explain the result from the linear regression model. However, the line pattern also implies that the there are different predictions for same y, hence Linear regression may not be the best model considering the nature of the observed y-values.

After removing the duplicated columns, The R-square for the linear regression model of latitude and longitude did not change, but the NA coefficients has been removed. This make sense because same information has been explained by the model before and after removing those redundant columns. Since the new data set has less number of features, it would be used for the analysis ahead.

Box-Cox transformation and Regularization

Since there are negative values in the latitude and longitude, a constant has been added (which is the minimum value among all the samples) to all the samples to conduct the box cox transformation. The result would still make sense because the dependent variables are just angle.

Validation and Use of Performance Metric

The data has been split into training set and testing set. The training set is use for building various models. The model is then fitted into the testing set and the Mean square error (MSE) of the predicted y over observed y is used for accessing performance:

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

The result of box-cox transformation and regularization are computed and summarized in table

1. Latitude

Table 1 Box-Cox transformation on Latitude

	No Box Cox	Box-Cox (best lambda = 1.79798)
R-Squared Value	0.296	0.3243
MSE (test)	317.18	324.86

Comment:

Comparison between box-cox and un-box-cox transformation

There is just very small increase in R-squared value after box-cox transformation (with optimal value of lambda = 1.79798). Besides, the MSE on test data is even larger than model without box-cox transformation. This implies that box-cox transformation could barely give improvement for the regression model.

In regard of this, the regularization is continued without box-cox transformation, and the result is as below:

Table 2 Regularization on Latitude

	No regularization	L1(Lasso)	L2(Ridge)	ElasticNet		
Alpha		1	0	0.25	0.5	0.75
Lambda		0.822063	11.38464	1.562186	1.032559	1.026084
MSE (test)	317.18	300.29	303.21	296.43	296.62	300.61
Number of Variables	72	16	72	27	22	16

Comparison between L1, L2 and Elastic Net

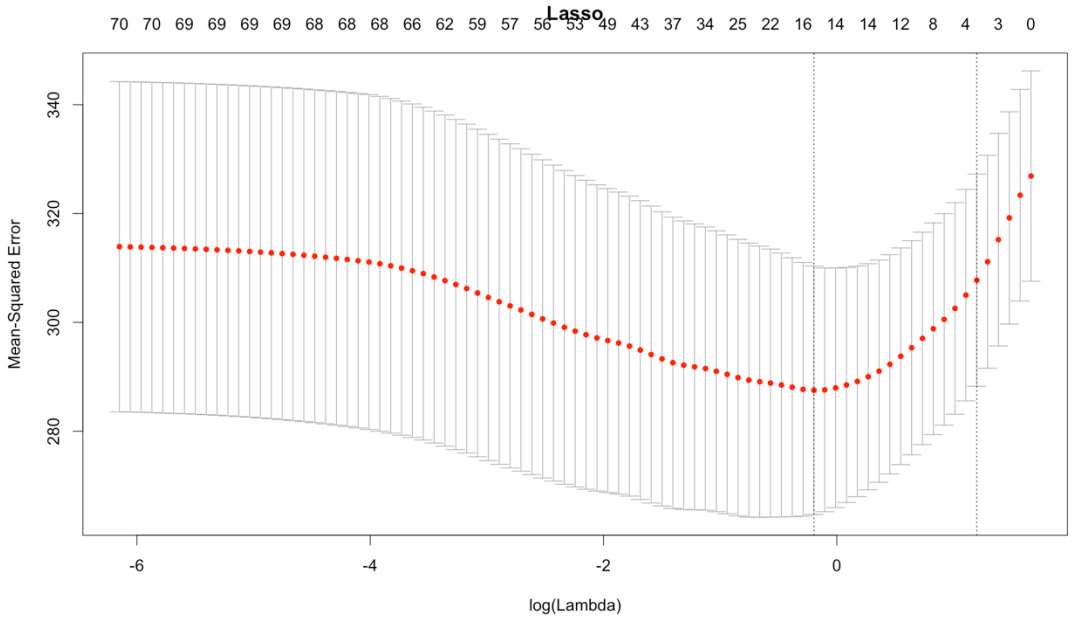
For L2, the regularization coefficient that produce minimum error is 11.38464, which used 72 variables in the model. This model has a lower value of MSE on test data than the unregularized regression. Hence, the L2 regularized regression model is better than then unregularized one.

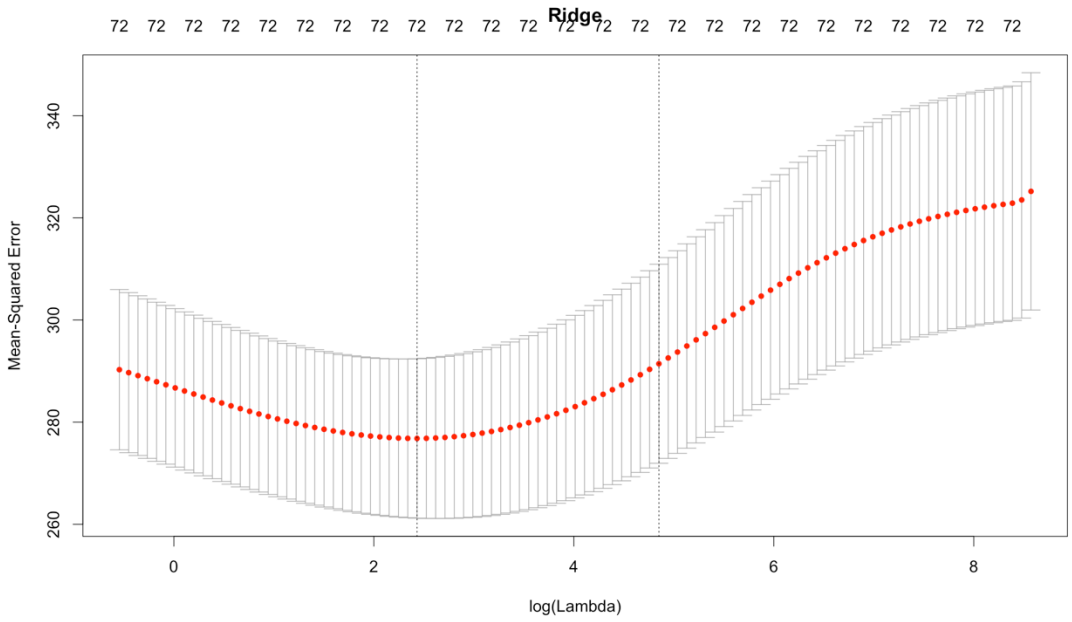
For L1, the regularization coefficient that produce minimum error is 0.822063, which use 16 variables in the regression. This model has a lower value of MSE on test data than the unregularized regression. Hence, the L1 regularized regression model is better than then unregularized one.

For elastic net, three values of alpha have been tried: 0.25, 0.5 and 0.75. The regularization coefficient that produce minimum error is 1.562186 with the alpha value of 0.25, which use 27 variables in the regression. This model has a lower value of MSE on test data than the unregularized regression. Hence, this regularized regression performs better than then unregularized one.

Overall, elastic net model with alpha = 0.25 give the lowest MSE = of 296.43 perform best among all other regularized models as well as the unregularized model.

The 3 plots of MSE over log(lambda) for lasso, ridge and elastic net and best alpha is as below:


Lasso
70 70 69 69 69 69 68 68 68 66 62 59 57 56 55 49 43 37 34 25 22 16 14 14 12 8 4 3 0
Mean-Squared Error
340
320
300
280
-6
-4
-2
0
log(Lambda)


Ridge
72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72 72
Mean-Squared Error
340
320
300
280
260
0
2
4
6
8
log(Lambda)

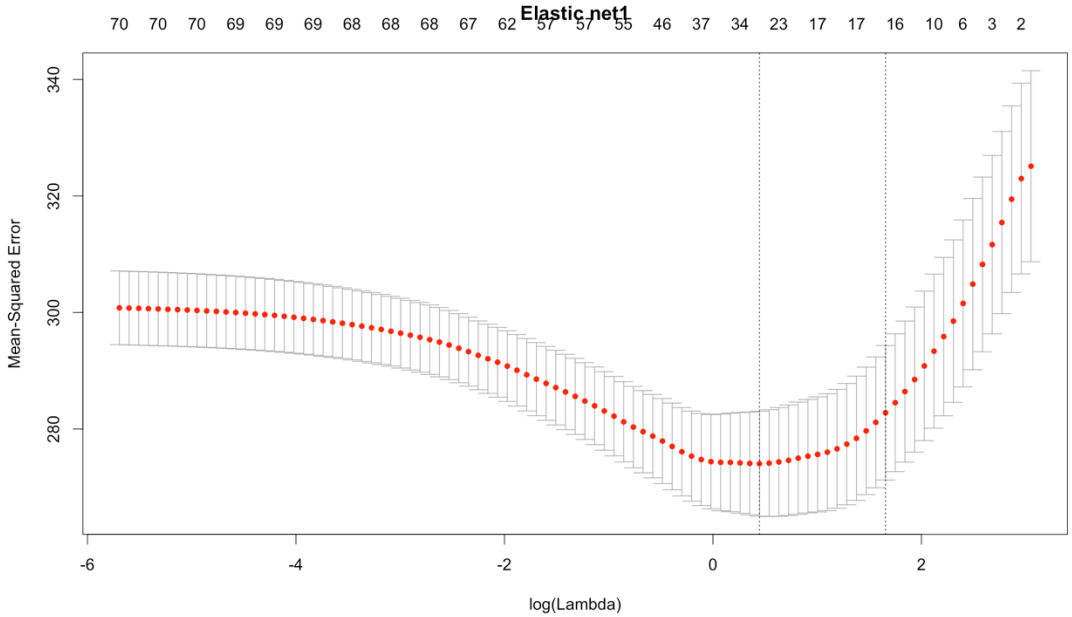


Figure 3: Plots of MSE over log(lambda) for Lasso, Ridge and r Elastic Net with best Alpha = 0.25

Analysis: The variance of MSE of elastic net is much smaller than that of Lasso and Ridge. The best value of log(lambda) is reported in Table 2.

2. Longitude

Table 3: Box-Cox transformation on Longitude

	No Box Cox	Box-Cox (best lambda = 0.989899)
R-Squared Value	0.4127	0.4126

Comment:

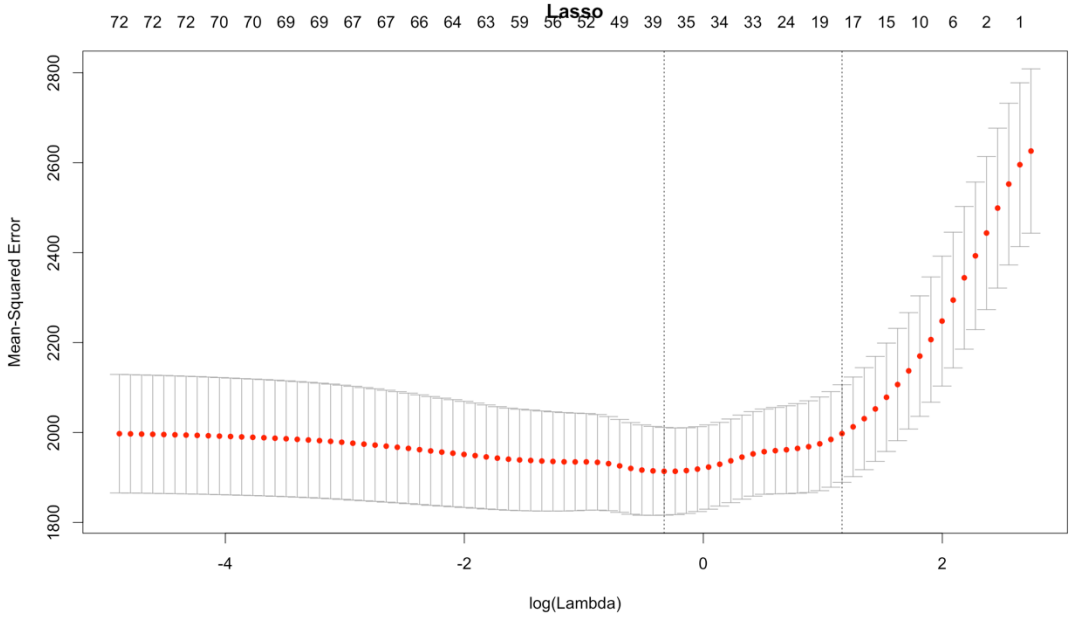
Comparison between box-cox and unbox-cox transformation

The box-cox transformation work best at lambda = 0.989899 which is very close to 1 which corresponding to the untransformed model. The R-squared values of the box-cox and unbox-cox model are very close too. This implies that the box-cox transformation give not much improvement to the overall regression. Hence, it is determined that the regularization is continued without box-cox transformation, and the result is as below:

Table 4: Regularization on Longitude

	No Regularization	L1 (Lasso)	L2(Ridge)	Elastic Net		
Alpha		1	0	0.25	0.5	0.75
Lambda		0.7202707	13.18627	3.470275	1.904311	0.7973086
MSE (test)	3379.74	3050.94	3078.85	3049.13	3047.34	3063.76
Number of Variables	72	39	72	37	35	43

Comparison between L1, L2 and Elastic Net



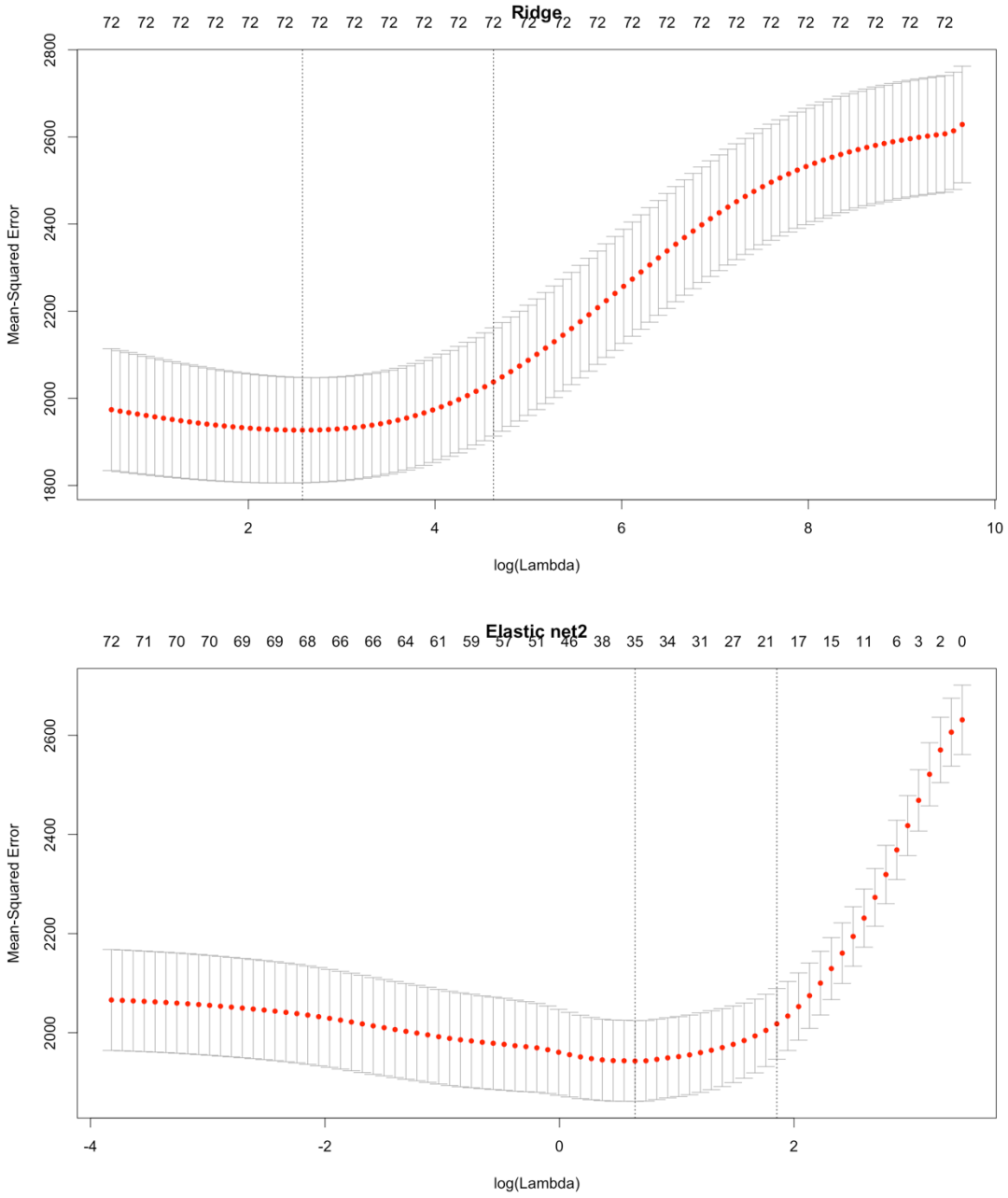


Figure 4 Plot for Regularization on Longitude

For L2, the regularization coefficient that produce minimum error is 13.18627, which used 72 variables in the model. This model has a lower value of MSE on test data than the unregularized regression. Hence, the L2 regularized regression model is better than then unregularized one.

For L1, the regularization coefficient that produce minimum error is 0.7202707, which use 39 variables in the regression. This model has a lower value of MSE on test data than the

unregularized regression. Hence, the L1 regularized regression model is better than then unregularized one.

For elastic net, three values of alpha have been tried: 0.25, 0.5 and 0.75. The regularization coefficient that produce minimum error is 1.904311 with the alpha value of 0.75, which use 35 variables in the regression. This model has a lower value of MSE on test data than the unregularized regression. Hence, this regularized regression performs better than then unregularized one.

Overall, elastic net model with alpha = 0.5 give the lowest MSE = of 3047.34. It performs best among all other regularized models as well as the unregularized model.

II. Logistic Regression

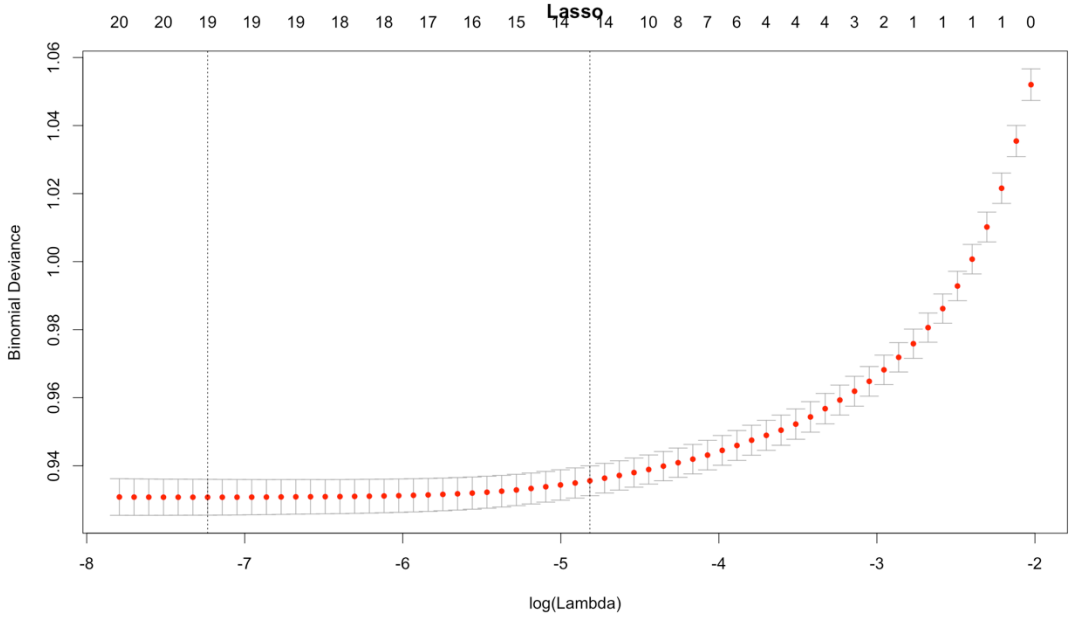
Performance Metric:

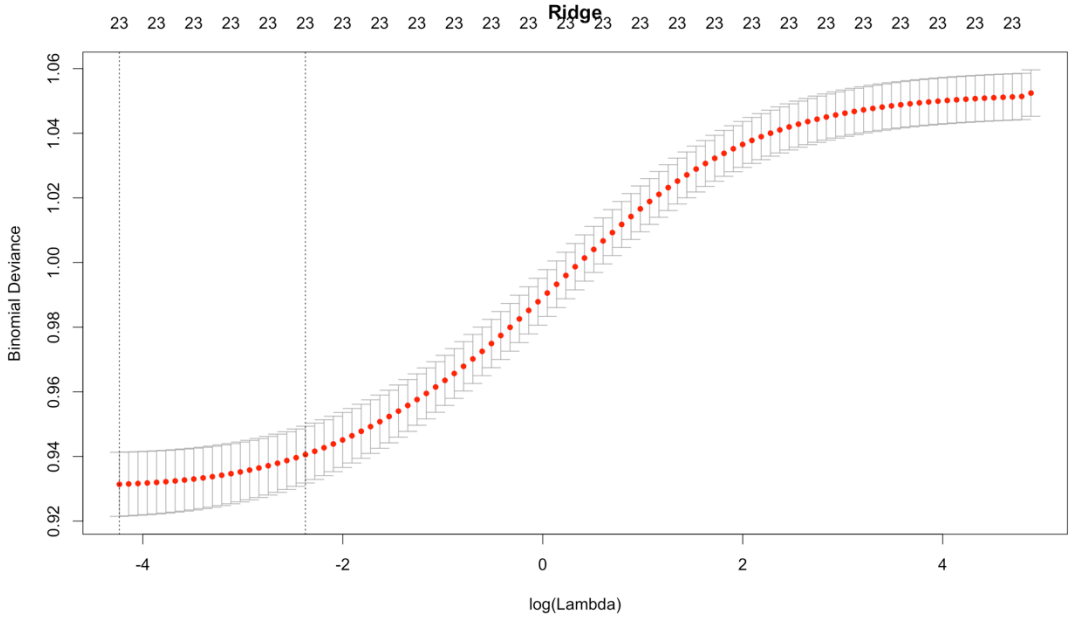
Accuracy = (# of predicted y = observed y)/(total # of observed y) in the test set

Result and Plots:

Table 5: Result of

Regularization Scheme	No regularization	L1	L2	Elastic Net
Alpha		0	1	0.5
Lambda		0.00072	0.01449	0.001197
Number of Variables		19	23	19
Accuracy (on test data)	0.795	0.805	0.803	0.806


Lasso
20 20 19 19 19 18 18 17 16 15 14 14 10 8 7 6 4 4 4 3 2 1 1 1 1 0
Binomial Deviance
0.94 0.96 0.98 1.00 1.02 1.04 1.06
-8 -7 -6 -5 -4 -3 -2
log(Lambda)


Ridge
23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23
Binomial Deviance
0.92 0.94 0.96 0.98 1.00 1.02 1.04 1.06
-4 -2 0 2 4
log(Lambda)

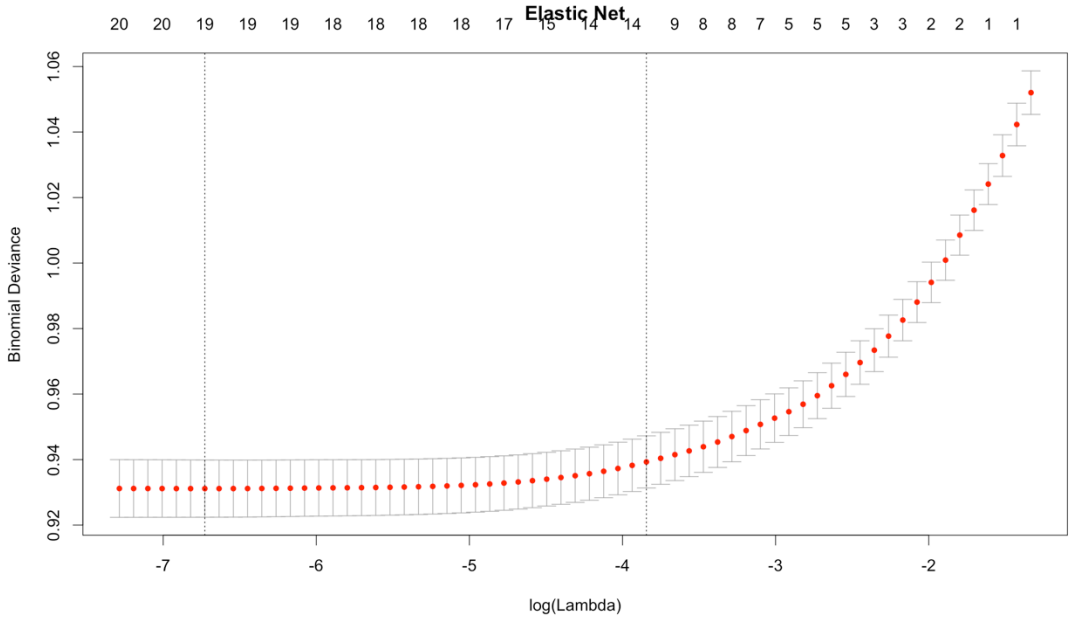


Figure 5 : Plot of Regularization on

Comparison between L1, L2 and Elastic Net

For L2, the regularization coefficient that produce minimum error is 0.01449, which used 23 variables in the model. This model has higher prediction accuracy on test data than the unregularized regression. Hence, the L2 regularized regression model is better than then unregularized one.

For L1, the regularization coefficient that produce minimum error is 0.00072, which use 19 variables in the regression. This model has a higher prediction accuracy on test data than the unregularized regression. Hence, the L1 regularized regression model is better than then unregularized one.

For elastic net with an alpha value of 0.5. The regularization coefficient that produce minimum error is 0.001197 which use 19 variables in the regression. This model has higher prediction accuracy on test data than the unregularized regression. Hence, this regularized regression performs better than then unregularized one.

Best Performance:

Overall, Elastic Net with alpha = 0.5 give the higher prediction accuracy on test data, it performs best among all other regularized models as well as the unregularized model.

General Conclusion and Discussion

Box-Cox Transformation for Linear Regression

For both latitude and longitude, the box-cox transformation does not show significant improvement in regression. This could be because for high-dimensional data, power transformation is not the most optimal tool for producing reliable linear regression models with high-dimensional data.

Regularization

In general, the regularized regression (both L1 and L2) perform better than unregularized regression. The advantages and disadvantages of the two regularization will be discussed as below:

Lasso Regression

In general, Lasso regression to perform better than ridge in this data set. Advantage of Lasso is that it performs well on high-dimensional data. Lasso induce sparsity in the model by shrinking non-significant variables to zero. As a result, the model complexity has been reduced and results in higher generalization power on the test set. Lasso regression should be used for variable selection.

However, the disadvantage of lasso is that it's performance will be lowered if there are independent variables of high collinearity. It will only select one variables of the variables with high pairwise correlations and shrinking the other one to zero. This could possibly lead to poorer predictive power than ridge regression.

Ridge Regression

One advantages of ridge regression is that it strikes a better balance between bias and variance by retaining all the variables in the model and prevent the coefficients of the model from going too large or too small, hence preventing overfitting.

Elastic Net

It is a convex combination of L1 and L2 penalty to penalizing regression model. In some case it would give better prediction accuracy than purely Lasso or Ridge regression.

Issues of Outliers

Regularized is highly-sensitive to outliers in the model, they could be removed from the data to further improve the comparisons among different models.