**Module 3: Week 3 Project Intermediate Analytics**

**Author’s Name:** **(Group A)** Shivani Sharma and Rupali Somani

**Course Number:** ALY6015

**Course Title:** Intermediate Analytics

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Joe Manseau

**Assignment Completion Date:** 11-20-2019



**Introduction**

This project is based on the concept of regularization and I am using Ridge regression model for handling my data set. The data set is about economic, banking and systematic crisis in 13 African countries from 1860 to 2014. I wanted to use Ridge model to figure to find out what are the factors which are associated with the Systematic crisis in Africa. This is a real time data gathered from actual results. In order to use the model, I have updated the data set and converted categorical data into numerical data. In statistics we have two critical characteristics of estimators to be considered i.e bias and variance. Bias is the difference between true population parameter and expected estimator which is measuring the accuracy of estimates and variance is measuring the spread and randomness. From the below graph we can infer that lower is the bias higher will be the variance and vice versa.

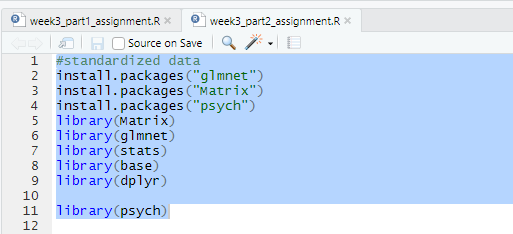
Here are the formulas:

https://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1543418450/eq3_inauam.png

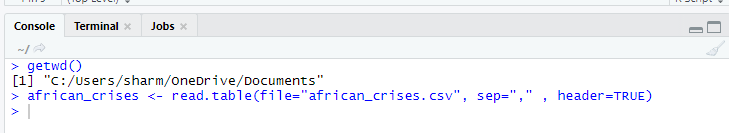
https://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1543418450/eq4_hwqu5v.png

**Analysis**

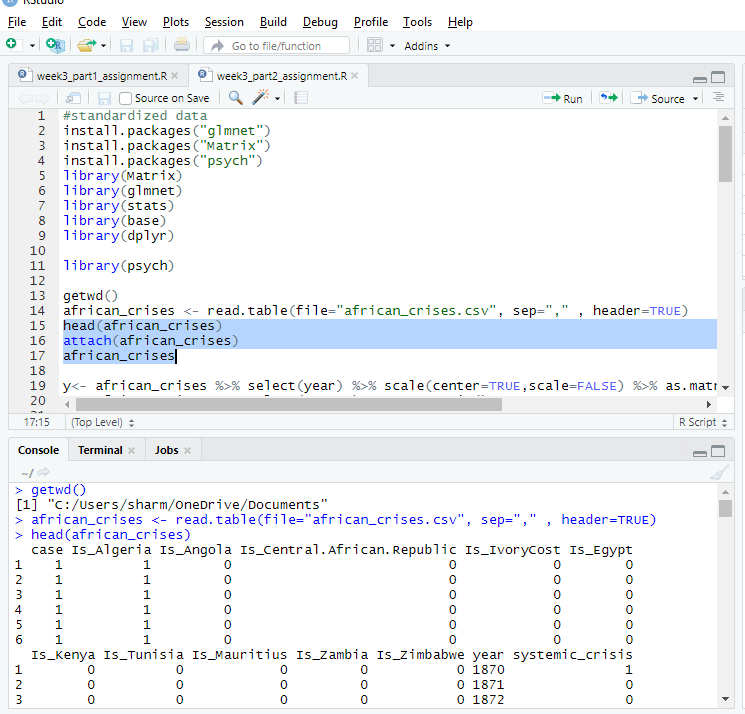
The data set focuses on the Banking, Debt, Financial and inflation and systematic crisis which happened from 1860 to 2014. I want to figure out the factors linked to Systematic crisis with the help of Ridge model in 13 different African countries. So before starting we have to process the data and convert into numeric data so as to apply the regularization model which we can achieve with the help of dummy data or pivot tables.



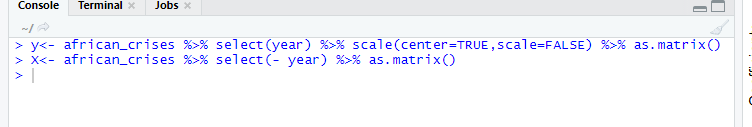
In order to use this model we need to download few packages in R i.e glmnet, Matrix, psych, dplyr, stats, base. Then we have to load the libraries and finally we are all set to use the model on data set. First read the data set file using following logic:

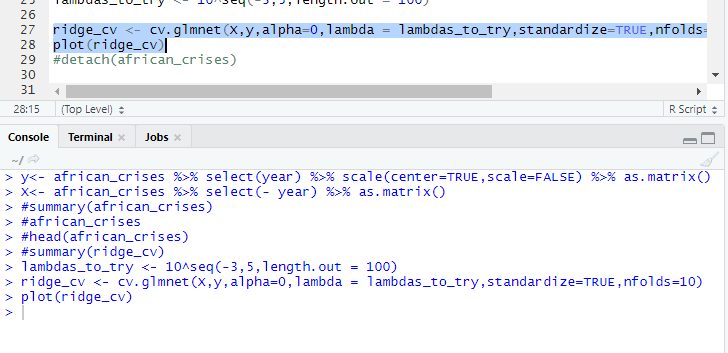


Here is the output for the same after the head i.e columns of the data set.



Now I am plotting the data year wise so I will put the values in the X and Y axis for plotting using the Ridge model and now using the glmnet function we will plot the data.



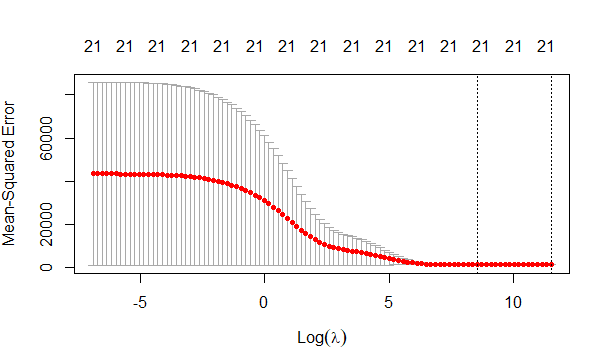


In Ridge model, lambda is the regularization penalty parameter. And we have to following the logic of ridge model which is stated below:

As λ→0, β^ridge→β^OLS

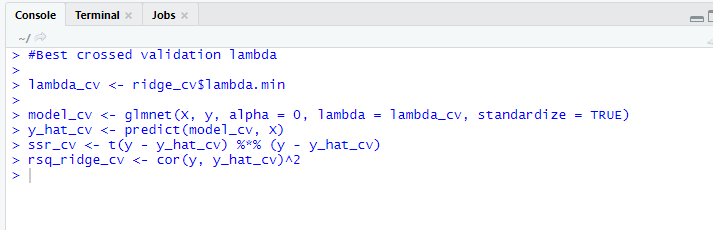
As λ→∞,β^ridge→0

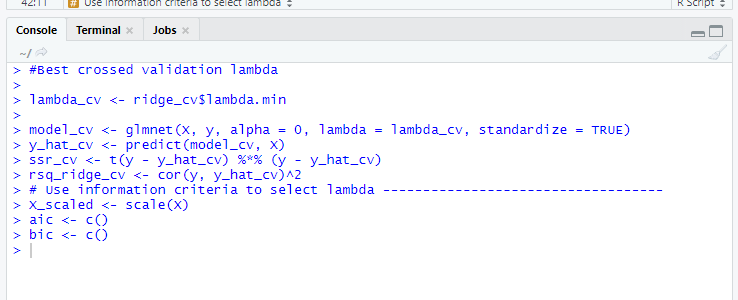
Then setting lambda to zero is same as using the OLS and larger the value of this the more strongly the coefficients size will be getting penalized. So, as the value of lambda is becoming more the respective variance will be decreased and the bias will be increased. From the graph we can infer that the mean-squared error is showing a non-linear relation with the lambda. It is becoming a constant once it reaches to a certain limit.

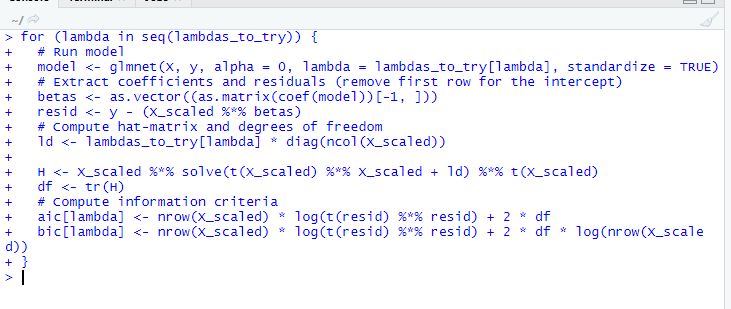


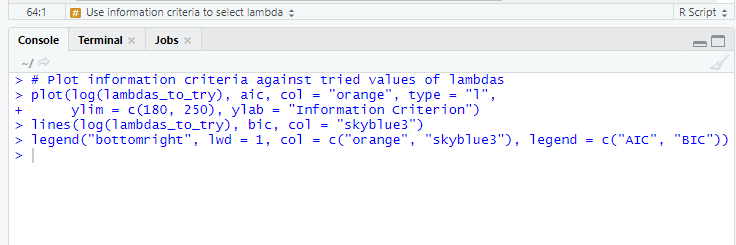
This graph is showing this behavior as the values are very small in the data set for inflation annual CPI and exch\_usd columns.

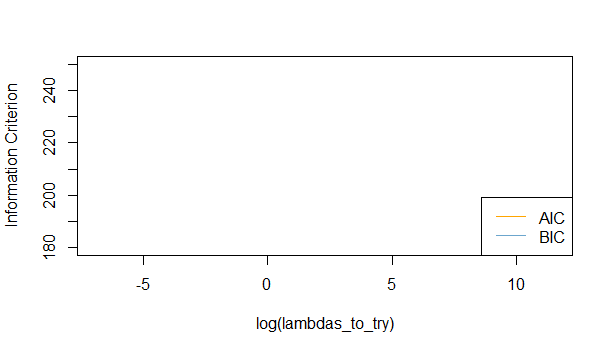
I wanted to have a clearer insight of the data so I decided to apply a crossed validation and more and then applying the model again. So, following is the logic:



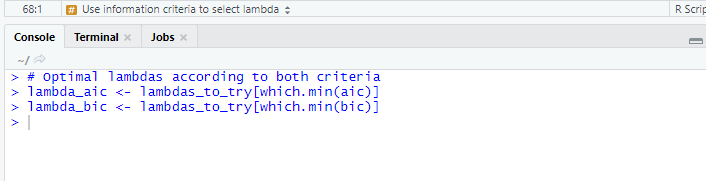




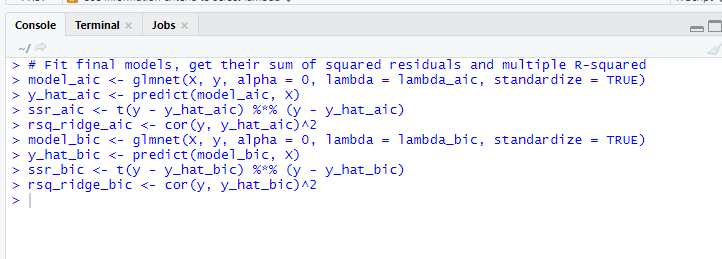




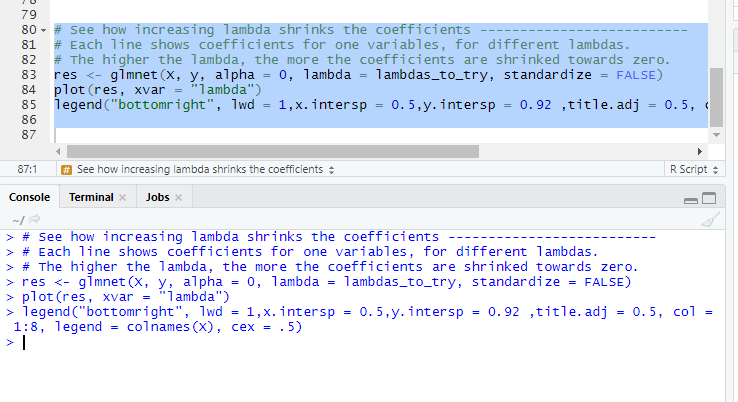
From the above graph we can observe that there is no relationship obtained between AIC and BIC parameters against information criterion and lambdas\_to\_try. So, we are not able to conclude about the behavior and this is very odd and the reason behind this could be negative values in the inflation annual cpi column. So, we are unable to provide information for AIC and BIC parameters. Now we need to optimize the values of lambdas for AIC and BIC.



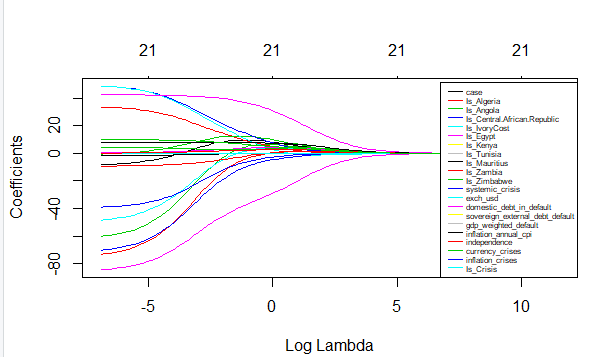
Now I want to get the sum of squared residuals and multiple R-squared and applying the model.



I wanted to show the increasing lambas values which is inversely related to coefficients i.e shrinking of coefficients which concludes that more is the value of lambas , the more coefficients are shrinking towards zero.



So following is the optimized results for all the parameters with more clear information of all the tuples in the data set.



The values for each coefficient is showing up a clear relation with lambda.

**Conclusion**

# At the end I would like to conclude that in order to reduce the complexity of model i.e number of predictors, we can use the forward or backward selection for this. Removing the predictors from model is done by setting the coefficients to zero. However here is a point i.e rather that forcing the predictors accurately to 0 ,we can put penalty if they are very far from 0. With this process we will reduce the complexity of model and keep all the variables of model. So from the final graph we can conclude that Algeria is showing more crisis as compared to other countries.

# References

1. “(Tutorial) Regularization: Ridge, Lasso and Elastic Net.” *DataCamp Community*, <https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net>.
2. Chiri. “Africa Economic, Banking and Systemic Crisis Data.” *Kaggle*, 21 July 2019, https://www.kaggle.com/chirin/africa-economic-banking-and-systemic-crisis-data.
3. “Your Home for Data Science.” *Kaggle*, <https://www.kaggle.com/>.
4. *YouTube*, YouTube, <https://www.youtube.com/watch?v=Q81RR3yKn30>.