COMPUTATIONAL METHODS IN STATISTICS
Final Project- PART I
REGRESSION ANALYSIS

February 2, 2020

Author: Bita Nezamdoust
Student ID: 002332793
Instructor: Dr. Xin Qi

Georgia State University
Department of Mathematics and Statistics

INTRODUCTION

The present project deals with regression analysis of *Parkinsons Telemonitoring Data Set* aimed at predicting two desired response variables, based on sixteen predictor variables that have been collected. In this study, two linear regression models are made for two separate response variable, namely *motor UPDRS* and *total UPDRS*. The goal is to find the best fitting model that provides the lowest prediction error and hence, the highest rate of prediction accuracy. The regression models obtained are tested against appropriate testing data and the best models with highest rate of accuracy are selected, among several model-fitting techniques that have been employed.

Data Description

The dataset was created at the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. The dataset is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring[1].

The data consists of twenty-two variables, among which variables *motor_UPDRS* and *total_UPDRS* are the two response variables. The first four variables represent subject ID, age, gender, and test time, and the last sixteen variables are biomedical voice measures which are those of interest to form the predictor variables of the regression model that is going to be built. The data is split into a training set with 1000 observations and a test set with 4875 observations for further investigation of the model's accuracy.

The predictor variables appear to be strongly correlated, as seen in the provided figures. Fig. 1 presents the correlation between every two response variables. The collection shown by the pie charts represents the strength of the linear relationship between the variables. The associated correlation coefficients are provided on the left side figure. The strong upward and downward correlation of some of the predictors is apparent in the scatter plots provided in Fig. 2. The multicollinearity that exists between the predictors could impact the prediction based on these variables negatively, which calls for appropriate measures to be taken to tackle the issue. I will address the issue in the model-fitting section.

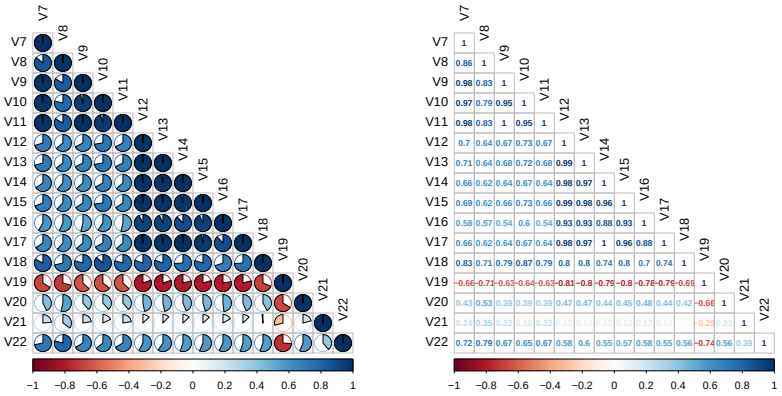

Figure 1: Multicollinearity between the predictor variables

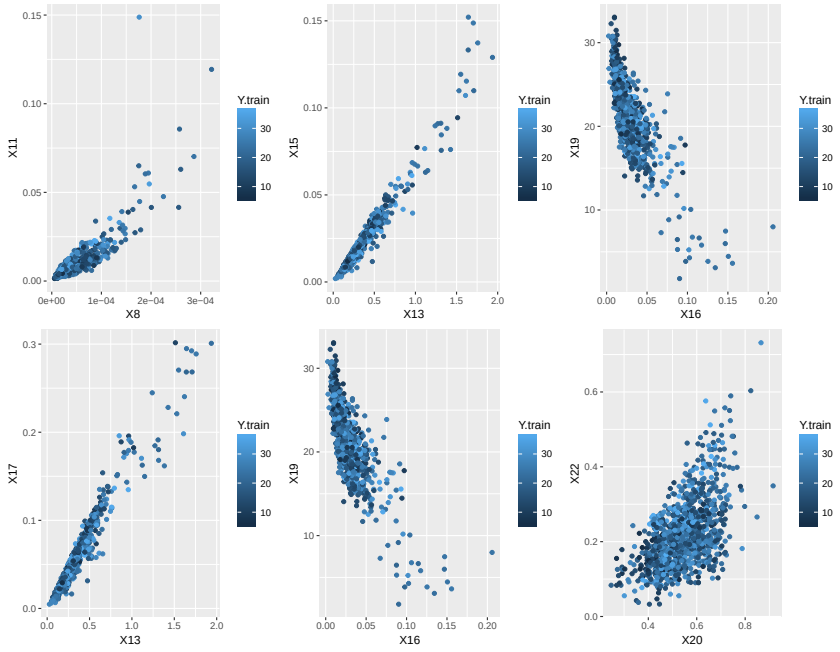

Figure 2: The scatter plot of some of the correlated predictor variables

Regression Models

The first response variable to consider is *motor_UPDRS*. Initially a *Usual Linear Regression Model* is employed to fit a regression line to the data and create the prediction line. The

obtained model's performance is tested on the testing data set. As the first column in the summary table below shows, the mean square error, a.k.a the prediction error, from the linear regression fit is equal to 0.069[1], which means the model is able to predict the *motor_UPDRS* variable with 93% accuracy. While the obtained accuracy level is relatively good, further models are employed to reach higher accuracy, especially as the predictor variables are quite strongly correlated and can impact the prediction results, as has been discussed.

Linear Regression Prediction Error (motor UPDRS)			
Linear Regression	Ridge Regression	Lasso Regression	Ridge Sqrt transform
0.069	0.055	0.056	0.055

To tackle the issue of multicollinearity in multiple regression analysis, two of the common regularity methods to utilize are *Ridge Regression* and *Lasso Regression*. Strong correlation between predictor variables can lead to large variation of the estimated parameters in the model. By adding a degree of bias as penalty to the regression estimates, the Ridge and Lasso models attempt to restrict the range of the parameters and so reduce the standard error. As the table indicates, the two models have improved the accuracy by about 2% as the Ridge penalty results in the an error of 0.055 and Lasso penalty 0.056.

Another way to deal with large number of variables and possible multicollinearity is to use dimension reduction methods that reduce the number of parameters and can improve the model accuracy. *Principal Component Analysis (PCA)* is applied as the variable selection method. 3 The new resulting variables (a.k.a PC's) were used to refit the previous models. The accuracy improved only slightly, as the mean square errors were 0.0546 and 0.0540, respectively for Ridge and Lasso fit. Additionally, A square root data transformation was also applied to the analysis as shown in the table, but it was observed that it did not make a significant difference in the accuracy of prediction.

Linear Regression Prediction Error (motor UPDRS)- PCA applied	
Ridge Regression	Lasso Regression
0.0546	0.0540

[1]The data set of the present section has been *normalized* for more precision and the ease of interpretation of the obtained mean square errors

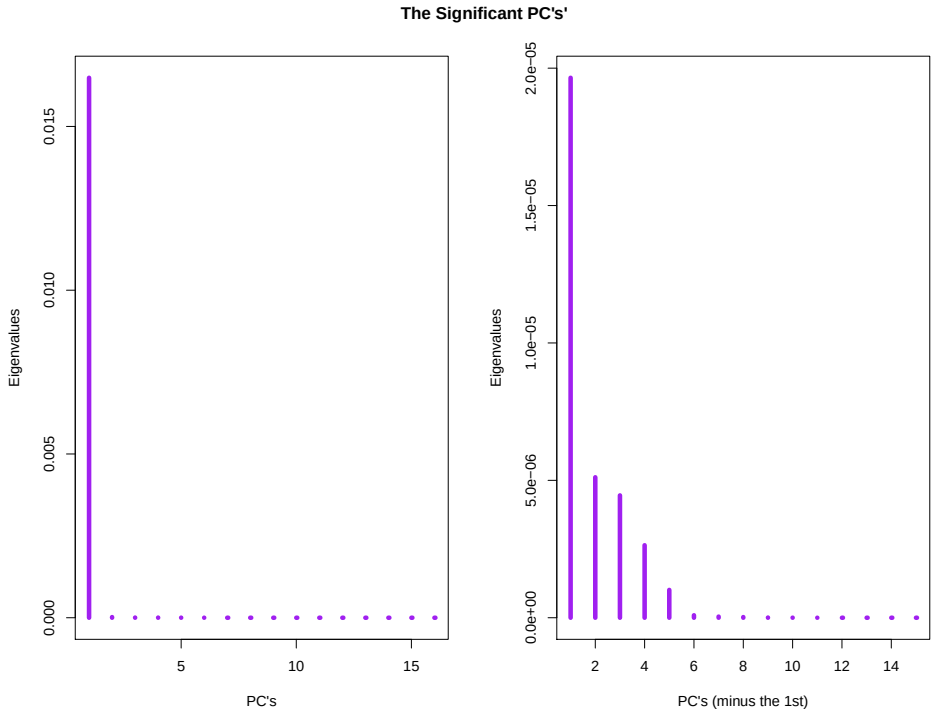

Figure 3: The amount of variability in the data explained by the PC's

Therefore, the best model for the prediction of motor UPDRS is **Linear Regression with Lasso Penalty - PCA applied** with the prediction error of 0.0540 and %95 level of accuracy. The tuning parameter of lambda is "lambda.min = 0.00187" chosen by 10-fold cross-validation and the regression coefficients are presented in the table below. It is evident from the table that further variable selection has also been accomplished by the Lasso method, as the coefficients corresponding to the second and third PC's are zero.

Intercept	Coefficients				
0.67986900	-0.04557849	0	0	5.58207043	9.26479508

The second response variable for which a linear model is built here is *total_UPDRS*. The same procedure as above is employed with the same predictor variables and a different response, and the results are summarized in the two tables below. Relatively similar to the results for *motor_UPDRS*, the prediction accuracy for *total_UPDRS* improved when the Ridge and Lasso penalties were added to the usual linear regression model, as the mean square error improved from 0.058 to 0.055 for both methods. Another observation is that the prediction improved slightly, i.e. by 0.001, when the square root transformation was applied, unlike for

the previous response variable where it remained unchanged.

Linear Regression Prediction Error (total UPDRS)			
Linear Regression	Ridge Regression	Lasso Regression	Ridge Sqrt transform
0.058	0.055	0.055	0.054

Linear Regression Prediction Error (total UPDRS) - PCA applied	
Ridge Regression	Lasso Regression
0.0477	0.0477

And finally, also similar to the first desired response variable, the best model for the prediction of total UPDRS is obtained by **Linear Regression with Lasso Penalty - PCA applied** with the prediction error of 0.0477 and the accuracy level of %95. The tuning parameter of lambda is "lambda.min = 0.001178" chosen by 10-fold cross-validation and the regression coefficients are presented in the table below. Variable selection is also accomplished by the Lasso method as the coefficients corresponding to the second and third PC's are zero.

Intercept	Coefficients				
0.71780993	-0.06315695	0	0	6.65172809	10.65443339

CONCLUSION

I built two linear regression models using sixteen variables involving biomedical voice measure relevant to the prediction of Parkinson's disease, to predict two response variables representing records of remote symptom progression monitoring of the disease on a telemonitoring device. The best final models to predict the symptom progression was obtained to be the Linear Regression model reduced in dimentionality by PCA and under Lasso penalty to tackle multicollinearity, with improved accuracy of %95, up from %93 from the initial model. Therefore, it could be stated that in this study, both regularity methods such as Ridge and Lasso penalties and dimension reduction methods such as PCA have been proven to make noteworthy improvements in the accuracy levels of the prediction models, and best possible accuracy was achieved with their addition to the model.

REFERENCES

[1] A Tsanas, MA Little, PE McSharry, LO Ramig (2009) 'Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests', IEEE Transactions on Biomedical Engineering [online] available at:
http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring