PARKINSON'S DISEASE TELEMONITORING ANALYSIS

Project Motivation

This analysis focuses on the development of an objective, automated method to extract clinically useful information from in the context of Parkinson's disease (PD). The main aim of the data is to predict the motor and total UPDRS scores ('motor_UPDRS' and 'total_UPDRS') from the 16 voice measures.

Parkinson's disease is a progressive disorder of the nervous system that affects movement. The person's motor skill is affected and their speech may become soft or slurred and this feature is used to determine the condition as this can be collected easily.

Dataset Overview

This 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. The recordings were automatically captured in the patient's homes. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

Dataset Source

The dataset was created by Athanasios Tsanas (tsanasthanasis@gmail.com) and Max Little (littlem '@' physics.ox.ac.uk) of 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 original study used a range of linear and nonlinear regression methods to predict the clinician's Parkinson's disease symptom score on the UPDRS scale.

Citation: 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

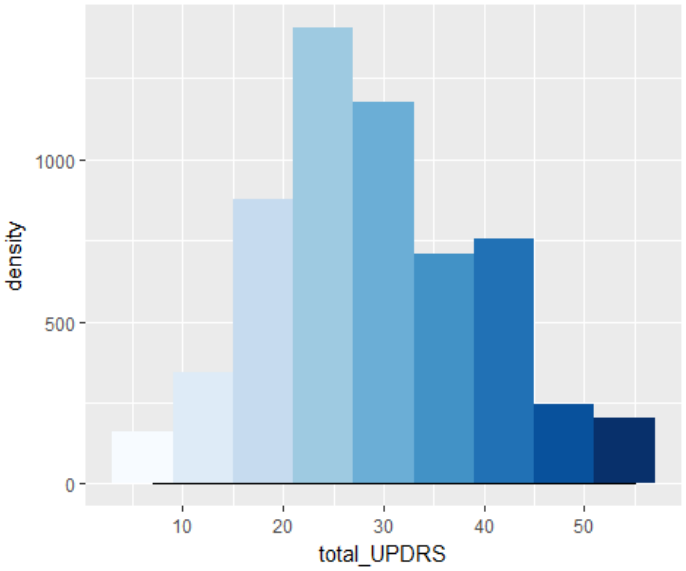
Approach

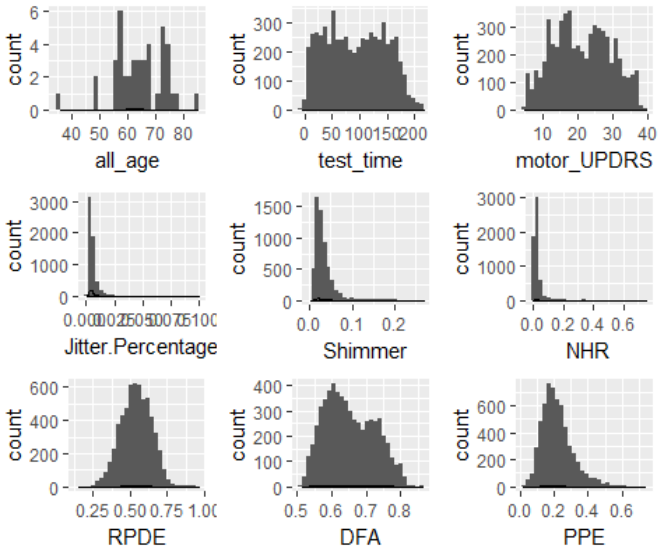
We started with finding the correlation between various parameters using different visualizations followed by the linear regression model that identifies the significant parameters for predicting the total_UPDRS score. Later time series analysis is done based on the test time.

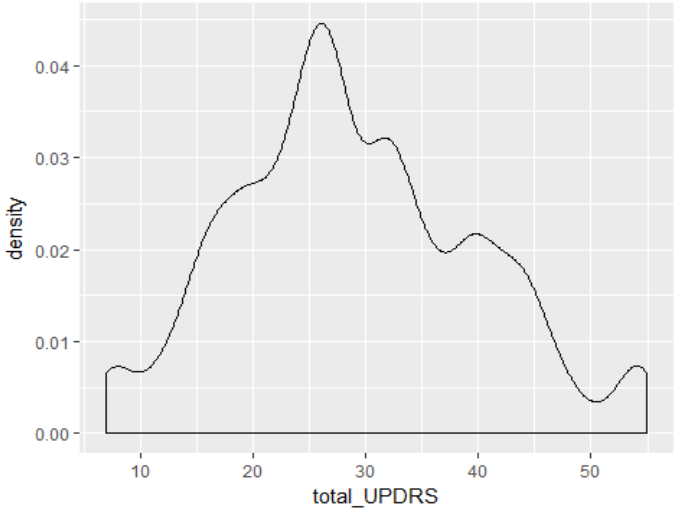
Basic Visualizations

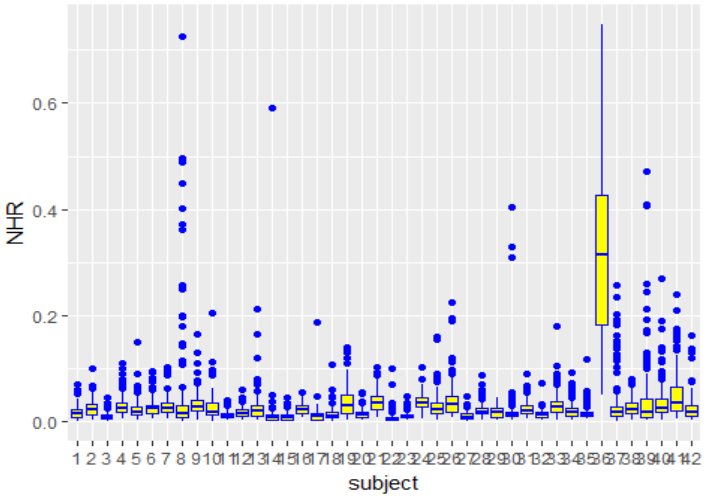
First the collected data is plotted to see the general spread of data and to get an idea of how the density of the data is with respect to various parameters. As we can see, age selected is mostly over 55 with 1 person under 40. There is a good distribution of the test time. The jitter and shimmer percentage doesn't seem to

vary much and is concentrated to the lower side. The distribution of RPDE, DFA and PPE are more evenly distributed in the range.









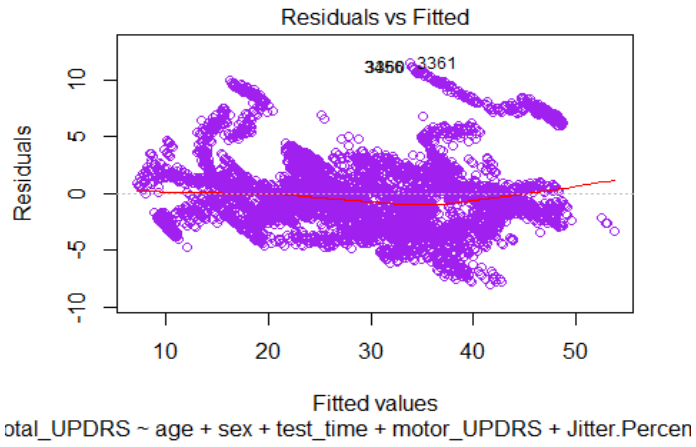
Linear Regression:

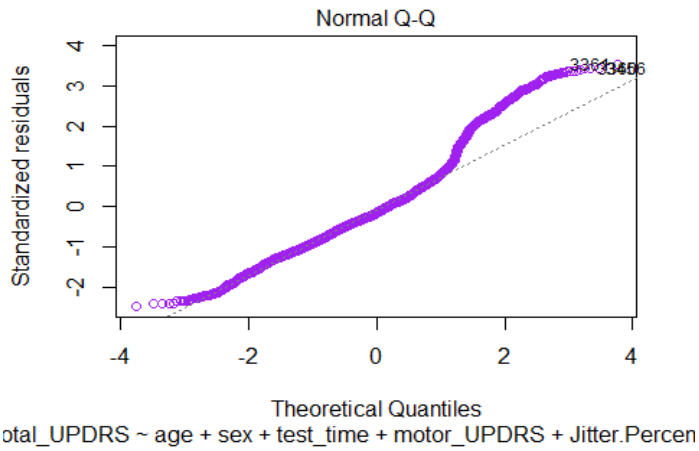
We plot the linear regression model for total_UPDRS against all other parameters. Initially on all the parameters and improve the model by removing the insignificant parameters from the model and analyze.

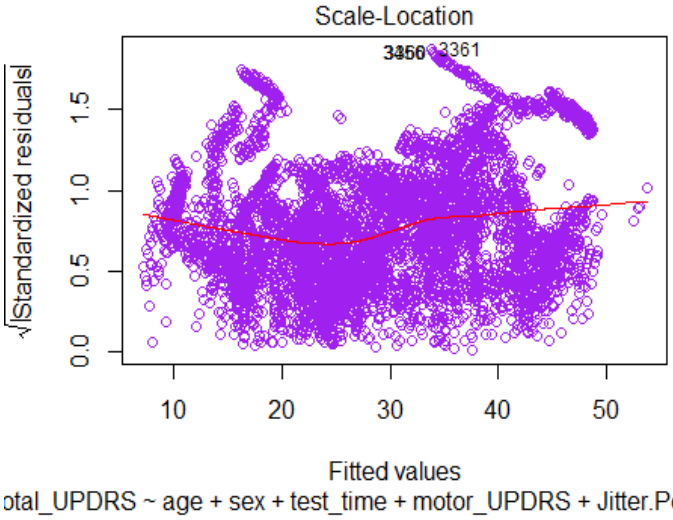
Model before optimization:

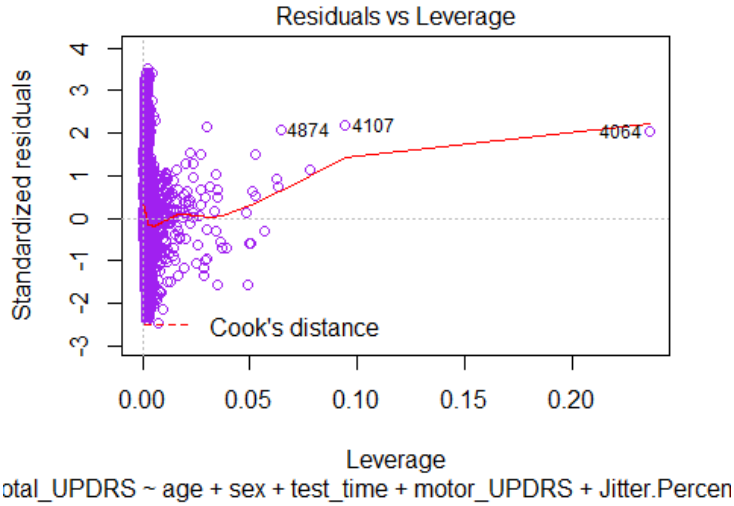
R-square of the above model is 90.74% and adjusted R-square is 90.71% which shows good correlation in the model.

Using step AIC we get the subset of the model we will evaluate the model by checking AIC values for each variable and try to reduce AIC of the model by removing the variables of high AIC. Once we do this, step$anova gives us the best model among others.









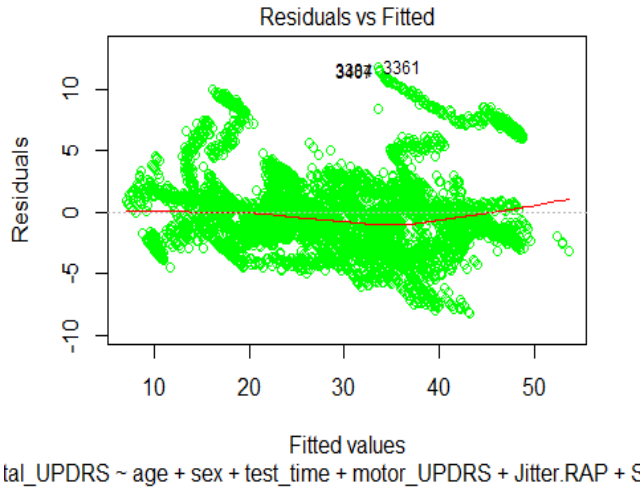
Model after optimization:

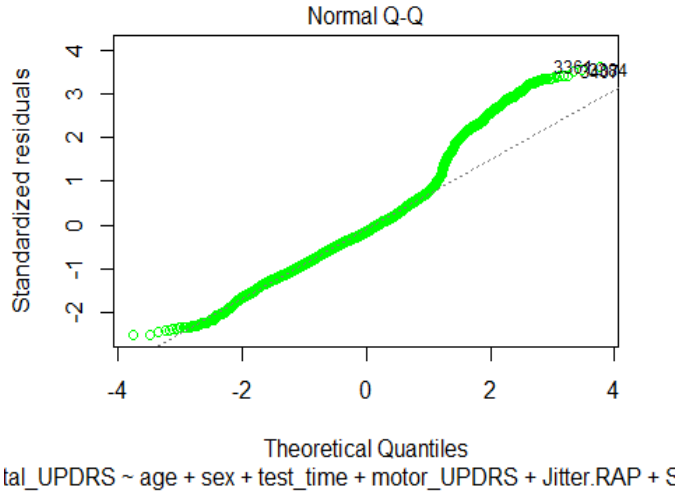
The parameters: Jitter.PPQ5, Jitter.Percentage, Jitter.DDP, Shimmer.dB, Shimmer.DDA, NHR are insignificant and hence removed from the final model. This model seems to be fittest model as R-square is 90.7% and Adjusted R-square is also high equal to 90.68% which show a good correlation.

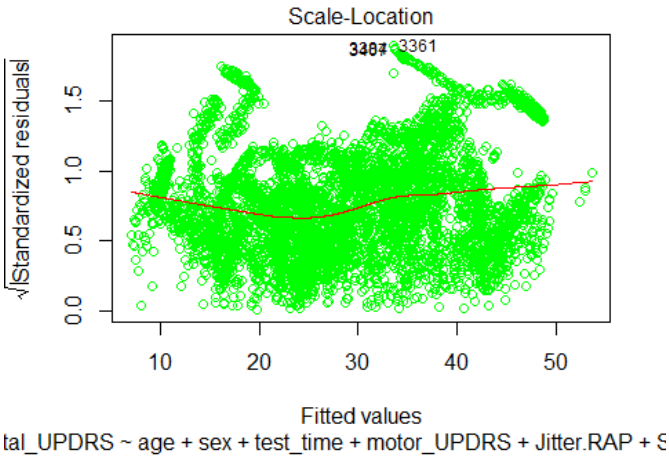
The linear regression models a range of -2.079 as a first quantile and 1.4599 as a third quantile with median at -0.05 seem to be symmetric and low enough.

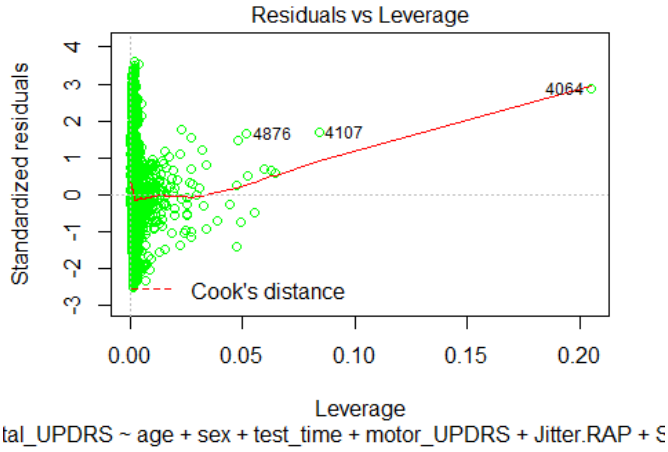
The coefficient for example age, shows that there is a good dependency on the parameter. There seem to be a change of 6.843e-02 for a unit increase in age with an error of 5.167e-03. Based on the t value of 13.243 and the probability of finding a value greater than t at < 2e-16, we can state that there is a clear relation with the parameter selected. As can be seen, not all the parameters considered seem to have an effect on the output and there is some null hypothesis present in that regard.

The R-square error above shows a 90 % dependency showing a good correlation of the model with the actual data.





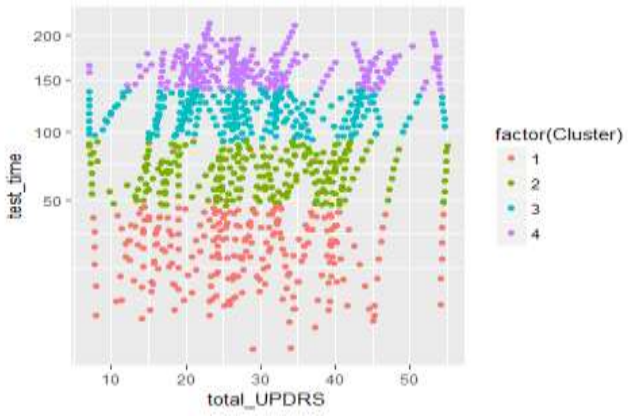


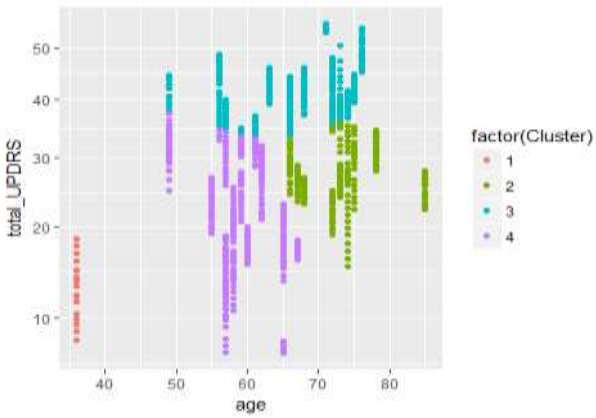


Our model seems to be linearly fit as seen from the residual vs fitted values plot. Norm QQ plot also telling that quantiles and standardized residual are equally normally distributed. Residual vs Leverage plot is also uniform.

Clustering

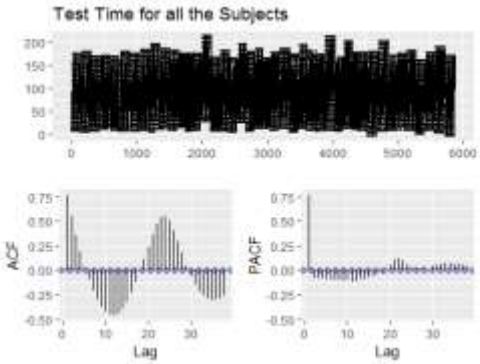
Clustering is performed based on total_UPDRS against test_time to give 4 clusters. Clustering on multiple variables with age, total_UPDRS and sex is also done which can be shown in the below figures.

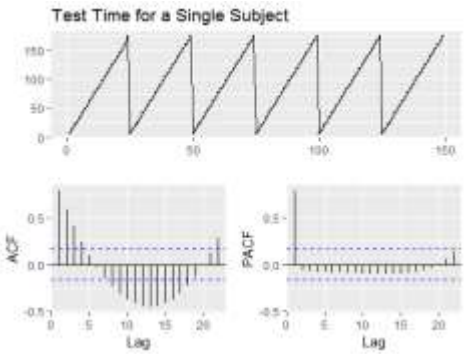


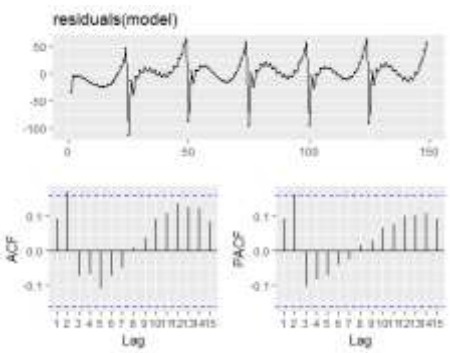


Timeseries

Time series is done on test_time attribute to predict the test time for single subject and all the subjects. Then ARIMA model is applied, its summary is verified, ACF and PACF graphs are plotted, prediction is made.







Conclusion:

This project had a lot of learning for us as this was a full project and gave us an idea as to how analysis is done over a data set using R.

We initially selected a data which had a binary output with mostly categorical input parameter and so demonstrating regression over it was challenging and we couldn't get a satisfactory classification. We then selected the current data set as this has both a continuous output and also a timeline information which would enable us to do a timeline analysis also on the same data set. Some things we tried which was in addition to the features covered in the lecture was a great learning opportunity. A possible improvement is to better do the time series analysis where we wanted to predict the values based on the initial values to see how the future is going to be based on the parameters and couldn't conclude.