Final Project *Biostatistics 2016*

Parkinsons Telemonitoring

The dataset was created by Athanasios Tsanas and Max Little 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.

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.

Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals. 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.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around 200 recordings per patient, the subject number of the patient is identified in the first column.

Attribute Information:

1. subject# - Integer that uniquely identifies each subject
2. age: Subject age
3. sex: Subject gender '0' - male, '1' - female
4. test_time: Time since recruitment into the trial. The integer part is the number of days since recruitment.
5. motor_UPDRS: Clinician's motor UPDRS score, linearly interpolated
6. total_UPDRS: Clinician's total UPDRS score, linearly interpolated
7. Jitter(%),Jitter(Abs),Jitter:RAP,Jitter:PPQ5,Jitter:DDP - Several measures of variation in fundamental frequency
8. Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,Shimmer:APQ11,Shimmer:DDA - Several measures of variation in amplitude
9. NHR,HNR - Two measures of ratio of noise to tonal components in the voice
10. RPDE - A nonlinear dynamical complexity measure
11. DFA - Signal fractal scaling exponent
12. PPE - A nonlinear measure of fundamental frequency variation

Aim of the Project: In this project, you are required to build one or more regression models to predict the variables "motor_UPDRS" and "total_UPDRS". Note that each patient has 200 recordings, and thus observations within a patient would be correlated. Explain why you think your model is good for describing this data set. After you build your model, you should use various statistical analysis techniques to show that your model is appropriate and correctly accounts for the potential structure in the data. You will also need to evaluate how good your model can predict the two response variables. To avoid the overfitting problem, you should only use a part of your data as your training data and use the other data as your testing data for model evaluation.