Predicting Parkinson Disease from Voice Recordings

Abstract

The diagnosis of Parkinson's disease has always been a major challenge for the medical community. In this study, we aim to construct a statistical model based on voice information from Parkinson's disease patients to predict the progression of the disease. By analyzing and classifying 5875 audio data samples, we discovered that different vocal methods require the establishment of distinct multivariate regression models in order to accurately predict the patients' conditions. And the project also suggests that it is necessary to first classify the vocal methos and then make predictions using the designated model to achieve a more precise Unified Parkinson's Disease Rating Scale.

I Introduction

Parkinson's Disease (PD) is a globally prevalent neurodegenerative disorder impacting an estimated 10 million individuals. Characterized by a reduction in dopamine levels due to degenerating dopaminergic neurons, PD manifests a range of symptoms including tremor, rigidity, slow movement, and postural instability. These symptoms vary in severity and combination but are invariably chronic and degenerative. Notably, approximately 90% of PD cases display vocal impairments, making voice recordings a potentially invaluable, non-invasive diagnostic tool. This insight presents a compelling opportunity for Machine Learning (ML) algorithms to utilize unique vocal features extracted from these recordings to predict or diagnose PD. Such models, leveraging signal processing algorithms, could predict scores on the Unified Parkinson's Disease Rating Scale (UPDRS)—the most commonly used scale for assessing PD progression. By successfully predicting UPDRS scores, we could streamline PD diagnosis, foster early intervention, and ultimately, enhance patient outcomes in the battle against this debilitating disease.

II Data and Methods

The dataset used in this study comes from a Parkinson disease study conducted by six U.S. medical centers. All institute made use of the *at-home testing device* (AHTD) to record patient behaviors remotely. A total of 52 patients were put on a six-month tracking period where the patients were asked to complete automatically prompted vocal protocols, and a total of 5875 voice recordings were recorded. These protocols consist of six phonation assignments, requiring the patient to pronounce the constant vowel sound "ahhhhh" at their comfortable pitch for four times and two times with twice the initial loudness. The term MDVP stands for the often-used KayPentax multidimensional voice program.

This study consists of four main parts: 1) EDA of the data, 2) A Full Regression Model of the dataset, 3) ANOVA blocked by different phonation task, 4) LSD analysis blocked by patient gender.

1) EDA
   During the Exploratory Data Analysis, we double checked the missing values of each feature, and drew up a correlation matrix of all the features used presented in the dataset.
2) Full Regression model
   Following the EDA, a full regression model of the dataset was performed. Based on the results gotten from the regression model, we were able to extract the significant features using the mallows' Cp criterion. Then, we performed an ANOVA test on the full regression model and the reduced regression model with only the extracted significant features.
3) ANOVA blocked by different phonation task
   As introduced in the dataset overview, the data acquiring consists of 6 tasks. Presented with such characteristic, we chose to block this factor. By assigning each phonation task with their corresponding number, we were able to 1) run regression on all six methods, 2) extract significant features of each method, 3) run ANOVA for each method of the full and the reduced regression model.
4) LSD Analysis blocked by patient gender

Lastly, within each analysis of the different phonation task, we also performed a LSD analysis on the two patient genders. By blocking patient's gender, we hope to investigate the difference between male patient Parkinson disease and female patient Parkinson disease.

III Analysis

Total UPDRS is a subjective assessment of Parkinson's severity by doctors in clinical practice. Researchers hope to achieve real-time assessment of Parkinson's disease by measuring the patient's voice. So, they try to analyze its relationship with total UPDRS. Therefore, in this study, we build basic a linear regression model with total UPDRS as the response variable and data of patient voice features as the independent variables.

We first establish linear regression models for different genders. Mallows' Cp criterion are used to select the variables and variable 'sex' are regarded as factors. It is tested that there is no difference between the reduced model composed of selected variables and the full model formed by adding all variables. All models were diagnosed after establishment, and the significance of the variables was tested.

For males:

$$\begin{aligned} total\_UPDRS = {} & 39.61 + 0.304age + 0.01667test\_time - 63380Jitter(Abs) \\ & - 388.1Jitter{:}\,PPQ5 + 321Jitter{:}\,DDP + 131.1Shimmer \\ & - 8.44Shimmer(dB) - 247Shimmer{:}\,APQ3 + 41.96Shimmer{:}\,APQ11 \\ & - 16.49NHR - 0.6234HNR + 4.12RPDE - 31.62DFA + 17.98PPE \end{aligned}$$

For females:

$$\begin{aligned} total\_UPDRS = {} & 36.81 + 0.304age + 0.01667test\_time - 63380Jitter(Abs) \\ & - 388.1Jitter{:}\,PPQ5 + 321Jitter{:}\,DDP + 131.1Shimmer \\ & - 8.44Shimmer(dB) - 247Shimmer{:}\,APQ3 + 41.96Shimmer{:}\,APQ11 \\ & - 16.49NHR - 0.6234HNR + 4.12RPDE - 31.62DFA + 17.98PPE \end{aligned}$$

In the regression model for males and females, the coefficients of all independent variables are the same, but the intercepts are different. This indicates that there is no interaction between gender and sound related variables. Therefore, in subsequent research, gender will be added to the model as a blocking factor.

During the test, patients are required to emit six different types of sounds during each sound collection, so the entire experimental data can be divided into six blocks based on the different types of sounds. However, if we take sound type (named 'method' in the following text) as a block variable, it will not be significant in the ANOVA test. This is because in each collection, the six sounds are emitted almost the same time and the total UPDRS marks for them are the same which seems that method does not have impact on total UPDRS. Actually, 'method' influences all variables about sound characteristics which means it influence the model. Therefore, six basic linear models are built for each 'method'. The six models are as follows:

Model1 (method=1):

$$\begin{aligned} total\_UPDRS = {} & 39.09 + 0.2912age - 2.287sex + 0.01476test\_time - 53510Jitter(Abs) \\ & + 200.6Jitter{:}\,DDP - 209.3Shimmer{:}\,APQ5 + 101.1Shimmer{:}\,APQ11 \\ & - 0.496HNR - 34.96DFA + 25.38PPE \end{aligned}$$

Model2 (method=2):

$$total\_UPDRS = 45.65 + 0.3038age - 2.866sex + 0.01285test\_time - 65860Jitter(Abs) + 177300Jitter:RAP - 59010Jitter:DDP - 55.82Shimmer:DDA - 0.7507HNR - 30.21DFA + 21.65PPE$$

Model3 (method=3):

$$total\_UPDRS = 24.94 + 0.3031age - 3.52sex + 0.01459test\_time - 126400Jitter(Abs) + 285.6Jitter:DDP - 13.86Shimmer(dB) + 93.37Shimmer:APQ11 - 0.2736HNR + 9.064RPDE - 24.76DFA + 29.98PPE$$

Model4 (method=4):

$$total\_UPDRS = 47.77 + 0.293age - 3.327sex + 0.01858test\_time - 72680Jitter(Abs) - 193300Jitter:RAP - 1336Jitter:PPQ5 + 65050Jitter:DDP + 324.7Shimmer - 214600Shimmer:APQ3 + 71360Shimmer:DDA - 34.37NHR - 0.7504HNR - 34.9DFA + 9.831PPE$$

Model5 (method=5):

$$total\_UPDRS = 52.55 + 0.2976age - 3.077sex + 0.01911test\_time + 942.2Jitter((\%) - 96210\text{Jitter(Abs)} + 272.7\text{Shimmer} - 199.8\text{Shimmer:DDA} - 66.99\text{NHR} - 0.9981\text{HNR} + 6.584\text{RPDE} - 37.22\text{DFA}$$

Model6 (method=6):

$$total\_UPDRS = 37.45 + 0.3143age - 3.918sex + 0.01711test\_time - 152400Jitter(Abs) - 1170Jitter:PPQ5 + 820.8Jitter:DDP + 410.1Shimmer - 19.73Shimmer(dB) - 183.5Shimmer:DDA - 28.42NHR - 0.751HNR + 9.941RPDE - 27.33DFA + 21.4PPE$$

IV Results and Conclusions

In the models corresponding to different ‘methods’, the selected variables may vary and the regression parameters are different. This indicates that in different "methods", the characteristics of sound may be reflected in different aspects, or the Parkinson's condition of patients may be reflected to varying degrees in different feature data. The variables age, sex and test time are selected in every model and their coefficients do not change much. This is because these variables are not affected by 'method'. It also indicates that they have a relatively stable effect on the development of Parkinson's symptoms.

Hence, when using sound features to predict the total UPDRS score, it is suggested to firstly classify the sounds and then use the corresponding regression models to predict. It is also advised to collect multiple sounds at once and use the prediction results of different models for validation and give a prediction of total UPDRS comprehensively.

However, there are still some limitations in our project. For instance, we only employed multiple linear regression instead of comparing different regression or machine learning methods. In future endeavors, we aim to explore alternative machine learning approaches to construct a more precise model.

In conclusion, through the use of exploratory data analysis, blocking, LSD analysis, and multiple linear regression, we have successfully developed a comprehensive model for predicting Parkinson's Disease based on voice recordings. Nevertheless, we acknowledge the significance of the vocal method as a crucial factor, and we emphasize the necessity of building different multiple linear regression models for different vocal methodologies.

V Team Work

Member	郝悦延	吕康乐	刘程华	庄成	纳日泰
Work	Analysis, Research Method and Coding	Analysis, Research Method, and Coding	EDA and Data Preprocessing	Analysis, Research Method and Coding	Research Method, Conclusion and Results

VI References

[1]Tsanas, Athanasios, Little, Max, A., & McSharry, et al. (2010). Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering.

[2]M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, “Prevalence of Parkinson’s disease in Europe: A collaborative study of population-based cohorts,” Neurology, vol. 54, pp. 21–23,2000.

[3] Jain, S. , & Shetty, S. . (2016). Improving accuracy in noninvasive telemonitoring of progression of Parkinson'S Disease using two-step predictive model. Third International Conference on Electrical. IEEE.

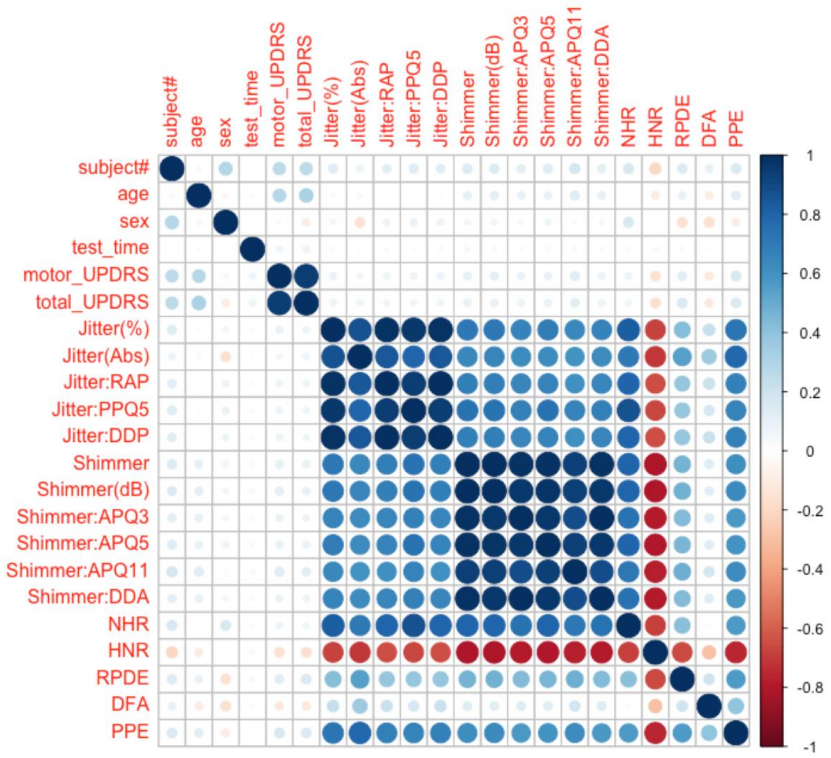
[4] Tsanas, A. . (2012). Accurate telemonitoring of parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. university of oxford.

APPENDIX

Based on the collected data, multiple feature extracting algorithms were applied. The following table was listed out all relevant criteria and their definition:

Table1. Relevant Criteria and Their Definition

MDVP Jitter(%)	**MDVP Jitter(Abs)**	**MDVP: RAP**
MDVP Jitter as a percentage	MDVP Jitter in microseconds	MDVP Relative Amplitude Perturbation
MDVP: PPQ	**Jitter: DDP**	**MDVP: Shimmer**
Period Perturbation Quotient	Difference of Differences between cycles, divided by the average period	Local shimmer
MDVP: Shimmer(dB)	**Shimmer: APQ3**	**Shimmer: APQ5**
Local shimmer in decibels	Three point Amplitude Perturbation Quetient	Five point Amplitude Perturbation Quetient
Shimmer DDA	**NHR**	**HNR**
Difference between consecutive difference between the amplitudes of consecutive periods	Noise-to-Harmonics Ratio	Harmonics to Noise Ratio
RPDE	**DFA**	**PPE**
Recurrence Period Density Entropy	Detrended Fluctuation Analysis	Pitch Period Entropy



Graph 1. Correlation Plot

```
Call:
lm(formula = total_UPDRS ~ . - motor_UPDRS - `subject#`, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.538  -7.386  -1.768   7.419  26.429

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.022e+01  3.241e+00  12.411  < 2e-16 ***
age              3.029e-01  1.506e-02  20.107  < 2e-16 ***
sex             -2.813e+00  3.171e-01  -8.869  < 2e-16 ***
method          -7.809e-02  7.729e-02  -1.010   0.3123
test_time        1.656e-02  2.387e-03   6.937 4.45e-12 ***
`Jitter(%)`      4.715e+01  2.129e+02   0.221   0.8248
`Jitter(Abs)`   -6.476e+04  9.907e+03  -6.537 6.78e-11 ***
`Jitter:RAP`    -4.078e+04  4.676e+04  -0.872   0.3831
`Jitter:PPQ5`   -3.459e+02  1.895e+02  -1.825   0.0680 .
`Jitter:DDP`     1.388e+04  1.559e+04   0.890   0.3734
Shimmer          1.496e+02  6.469e+01   2.313   0.0208 *
`Shimmer(dB)`   -8.797e+00  4.833e+00  -1.820   0.0688 .
`Shimmer:APQ3`  -1.492e+04  4.695e+04  -0.318   0.7506
`Shimmer:APQ5`  -6.159e+01  5.522e+01  -1.115   0.2648
`Shimmer:APQ11`  5.287e+01  2.483e+01   2.129   0.0333 *
`Shimmer:DDA`    4.900e+03  1.565e+04   0.313   0.7542
NHR             -1.512e+01  6.230e+00  -2.427   0.0152 *
HNR             -6.259e-01  6.886e-02  -9.089  < 2e-16 ***
RPDE             4.044e+00  1.821e+00   2.220   0.0265 *
DFA             -3.181e+01  2.330e+00 -13.651  < 2e-16 ***
PPE              1.751e+01  2.922e+00   5.992 2.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.726 on 5854 degrees of freedom
Multiple R-squared:  0.1767,    Adjusted R-squared:  0.1739
F-statistic: 62.81 on 20 and 5854 DF,  p-value: < 2.2e-16
```

Graph 2. Regression Model

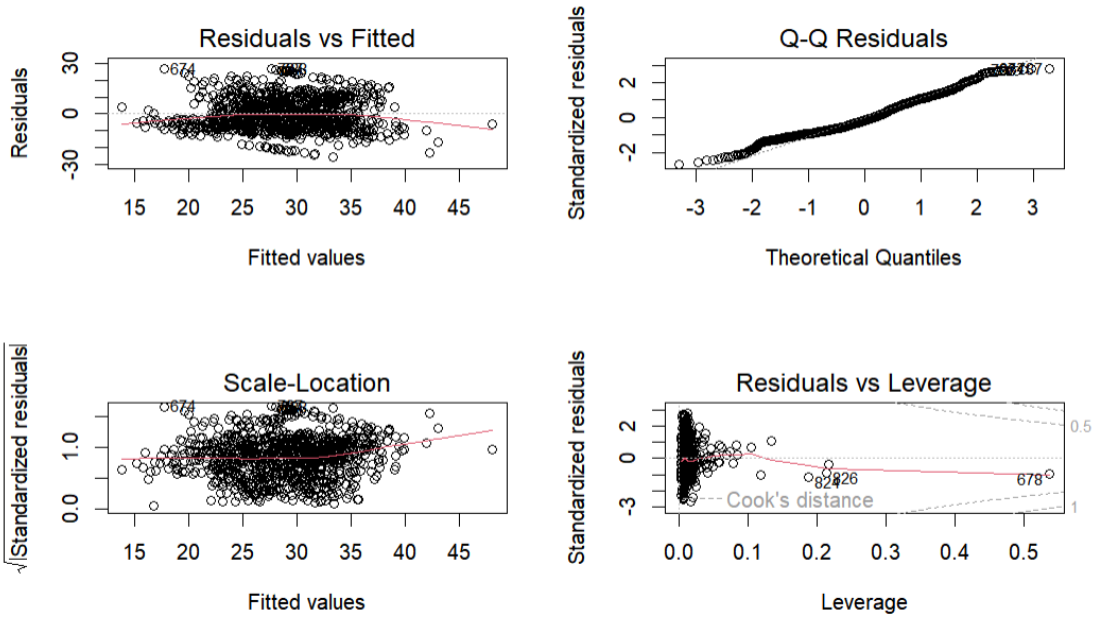
```
Call:
lm(formula = total_UPDRS ~ 0 + age + as.factor(sex) + test_time +
    `Jitter(Abs)` + `Jitter:PPQ5` + `Jitter:DDP` + Shimmer +
    `Shimmer(dB)` + `Shimmer:APQ3` + `Shimmer:APQ11` + NHR +
    HNR + RPDE + DFA + PPE, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.008  -7.365  -1.719   7.424  26.312

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
age               3.040e-01  1.504e-02  20.220  < 2e-16 ***
as.factor(sex)0   3.961e+01  3.187e+00  12.428  < 2e-16 ***
as.factor(sex)1   3.681e+01  3.166e+00  11.626  < 2e-16 ***
test_time         1.667e-02  2.385e-03   6.991 3.04e-12 ***
`Jitter(Abs)`    -6.338e+04  9.489e+03  -6.679 2.63e-11 ***
`Jitter:PPQ5`    -3.881e+02  1.457e+02  -2.664 0.007746 **
`Jitter:DDP`      3.210e+02  5.485e+01   5.853 5.10e-09 ***
Shimmer           1.311e+02  6.280e+01   2.088 0.036882 *
`Shimmer(dB)`    -8.440e+00  4.788e+00  -1.763 0.077984 .
`Shimmer:APQ3`   -2.470e+02  6.608e+01  -3.738 0.000188 ***
`Shimmer:APQ11`   4.196e+01  2.249e+01   1.865 0.062163 .
NHR              -1.649e+01  6.113e+00  -2.698 0.006994 **
HNR              -6.234e-01  6.826e-02  -9.132  < 2e-16 ***
RPDE              4.120e+00  1.808e+00   2.279 0.022690 *
DFA              -3.162e+01  2.302e+00 -13.733  < 2e-16 ***
PPE               1.798e+01  2.803e+00   6.413 1.54e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.724 on 5859 degrees of freedom
Multiple R-squared:  0.9014,    Adjusted R-squared:  0.9011
F-statistic:  3348 on 16 and 5859 DF,  p-value: < 2.2e-16
```

Graph 3. ANOVA Factor(sex)



Graph 4. Residual Plots (Model 1)