STA2101HF

PREDICTING PARKINSON'S DISEASE USING SPEECH SIGNALS

December 17, 2022

Kanika Chopra
University of Toronto

1 Part 1 - Non-Technical

1.1 The Scientific Problem of Interest

Parkinson's disease (PD) is a neurological disorder resulting in uncontrollable movements affecting over 10 million people worldwide [2]. This can include shaking, stiffness and difficulty with balance and coordination [1]. On top of this, PD can also result in dysphonia, disorders of the voice. Research has shown that 89% of PD patients experience speech and voice disorders, including soft, monotone, breathy and hoarse voice and uncertain articulation [2]. These symptoms can diminish their quality of life and reduce their urge to engage in social settings due to their difficulty communicating. Thus, there is interest in understanding how variation in voice measurements can be used to identify Parkinson's early on.

This paper will investigate how we can leverage speech signals processed from voice recordings to distinguish between patient's with and without PD. This will also involve identifying which voice measurements are significantly relevant to PD.

1.2 Data: How and Why it was collected?

The data was collected by the National Centre for Voice and Speech (NCVS) in Denver Colorado. The NCVS recorded the speech signals using feature extraction methods that were published for voice disorders [3]. This work was done in collaboration with Max Little from the University of Oxford who specializes in signal processing and machine learning.

The team obtained 6-8 voice recordings of vowel phonations from participants with and without PD. Biomedical measurement methods were applied to measure how the quality of vowels is affected by the vibrations and tensions formed in vocal cords [4]. This data was collected for the purpose of leveraging signal processing and machine learning to discriminate healthy people from those with PD based on dysphonia [5]. Further details on the signals included is provided in the next section.

1.3 Preliminary Description of the Data

This data is collected from an observational study aiming to make conclusions about an outcome of interest based on a sample of the population. In this study, the outcome of interest is whether or not a person has PD, represented by the variable health status. This data set was collected from 31 subjects, 23 with PD and 8 without [5]. From these 31 subjects, 195 voice recordings were taken; each recording is the unit of observation. Measurements were computed based on these audio recordings in an attempt to distinguish those with PD

from healthy participants.

The participants' ages ranged from 46 to 85 years. For those diagnosed with PD, their time since diagnosis ranged from 0 to 28 years [3]. However, in the UCI data, we do not have access to data regarding their time since diagnosis and age. There is 1 categorical predictor and 23 continuous predictors processed on the voice recordings. The categorical variable is a combination of the ASCII subject name and their recording number. The continuous variables can be broken into various types of acoustic measurements. We first note that variables with MDVP are measurements made on Multi-Dimensional Voice Programs which is a computerized acoustic analysis [9]. These variables include measures of the average (`MDVP:Fo(Hz)`), maximum (`MDVP:Fhi(Hz)`) and minimum `MDVP:Flo(Hz)`) fundamental frequency. Measures of variation in fundamental frequency include `MDVP:Jitter(%)`, `MDVP:Jittter(Abs)`, `MDVP:RAP`, `MDVP:PPQ`, `Jitter:DDP`. Measures of variation in amplitude include `MDVP:Shimmer`, `MDVP:Shimmer(db)`, `Shimmer:APQ3`, `Shimmer:APQ5`, `MDVP:APQ`, and `Shimmer:DDA`. There are two measures of ratio of noise to tonal components in the voice: `NHR` and `HNR` as well as two nonlinear dynamical complexity measures: `RPDE`, `D2`. We also have DFA which is a signal fractal scaling exponent and lastly three nonlinear measures of fundamental frequency variation: `spread1`, `spread2` and pitch period entropy (`PPE`) which is a measure of dysphonia. The full variable definitions are included in Appendix A.1. These variables are all alternative methods to represent the variability in the acoustics from a voice recording. In this data set, the distribution of the health status based on the total number of recordings is displayed in Figure 1.

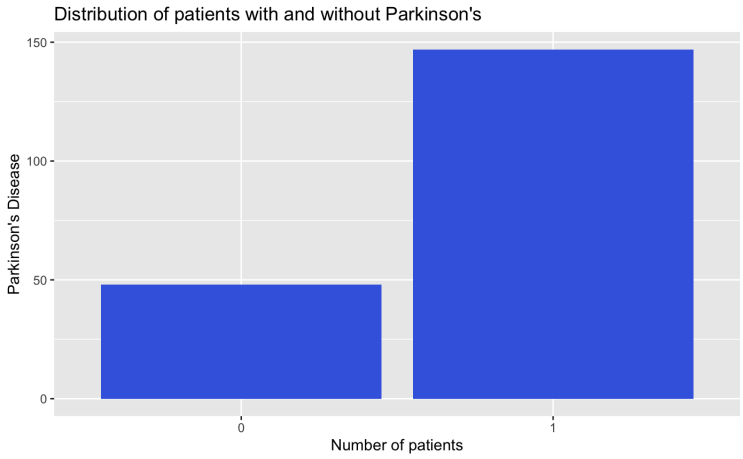

Figure 1: Distribution of Health Status

We can see that we have more audio recordings for subjects with PD than healthy subjects. Hence, we are working with an imbalanced data set. Figure 2 demonstrates how each

of the distribution of the continuous variables differs based on the subject's health status.

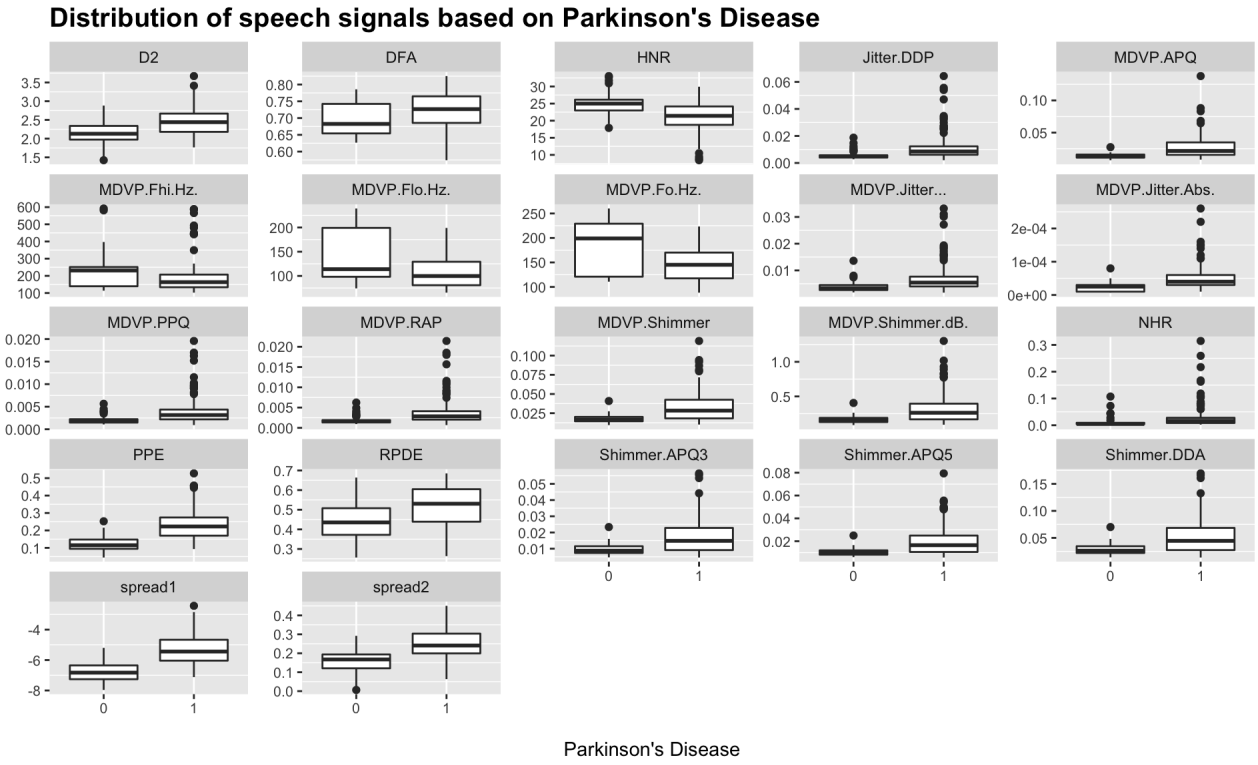

Figure 2: Distribution of predictor values based on health status

From this figure, we can observe that for the majority of measurements, those with PD have a higher distribution of values. The exception being `HNR` and `MDVP.Hz` (maximum, minimum and average). In general, it seems that those with PD have a higher variability in their fundamental frequency and amplitude. We also note that the ranges of the values for each measurement varies greatly, i.e. `D2` ranges from 1 to 4 whereas `HNR` ranges from 5 to 35. This will be considered with the data preprocessing prior to building our model. These box plots also show that there may be some outliers in the data; however, given that our data set is small, we do not remove these values as they may not be influential values. We will investigate influential values in the analysis section.

1.4 Summary for the Non-Statistician

The problem of interest was to use processed speech signal measurements to distinguish between subjects with Parkinson's Disease (PD) from those without. The data used involved a binary response variable, which was the health status of the subject (i.e. PD or not) and con-

tinuous variables which were measurements processed from audio recordings. To use these features to predict whether a participant has PD, we have to build a statistical model. The first step is to preprocess and clean the data to ensure it is structured and compatible with our model. This involved removing variables that were highly dependent on one another. This is done to ensure that we are able to identify the independent relationships between each measurement and the health status. Then, with our transformed data set, we aim to build a model that adequately fits our data. To build this model, we run many statistical tests to check the significance of each of the predictors and their interaction effects. This is done in a backwards selection process, meaning that we remove variables one by one. Each time a variable is removed, we ensure that it does not significantly diminish the fit of our model or increase the variability within our model. We continue with this process until we are left with a simpler model with only the variables of importance. This is following the parsimony principle which states that given two models that perform similarly, the simpler one is preferred. To test that we have an adequate fit, we conduct diagnostic checks with our final model to ensure that we are satisfying the underlying assumptions of the model and that our data is of high quality. This process concluded that average fundamental frequency (`MDVP.Fo.Hz`), detrended fluctuation analysis (`DFA`), nonlinear measure of fundamental frequency variation (`spread2`) and correlation dimensions (`D2`) are the variables significantly related to PD. A one unit increase in the average fundamental frequency would decrease the odds of having PD by 0.4222. Contrarily, a one unit increase in the detrended fluctuation analysis, nonlinear measure of fundamental frequency variation (`spread2`) and correlation dimensions would result in an increase in the odds of having PD by 1.9750, 2.6024 and 3.5402 respectively. Thus, increases in average fundamental frequency result in a significant decrease in the odds for having PD whereas the remaining predictors in the model result in a significant increase in the odds of having PD. Furthermore, we conclude that the interaction effects were not significant between any predictors.

Once we have a sufficient fit with our model, the goal is to distinguish between those with PD and those without using a separate test set. We then evaluate the accuracy of these predictions to conclude that the model is strong at identifying those with PD; however, it needs improvement at predicting those without PD. This is likely attributed to the imbalanced data set as the model has been exposed more PD subjects than non-PD subjects.

2 Part 2 - Technical

2.1 Models and Analysis

Model Building

Since the outcome of interest is a binary variable, a binary logistic regression model with the logit link was fit. Prior to building the model, the correlation between variables was checked to avoid multicollinearity when building the model. Figure 3 displays a heat map representing the Pearson correlation between all of the predictor variables. Based on the scale, the higher the value, the deeper blue the square will be. We take a look at variables that have a correlation greater than 0.5 with multiple other predictors. We note that the `jitter` and `shimmer` variables have high correlations with almost every other predictor. Similarly, the ratios `NHR` and `HNR` as well as the nonlinear measures of fundamental frequency, `spread1` and `PPE` also have high correlations with multiple other variables. Removing these variables leaves us with `MDVP.Fo.Hz, MDVP.Fhi.Hz, MDVP.Flo.Hz, RPDE, DFA, spread2` and `D2` as our variables of interest. This is an important step to ensure there is no multicollinearity as otherwise we would have dependent variables which diminishes the statistical significance of our independent estimates.

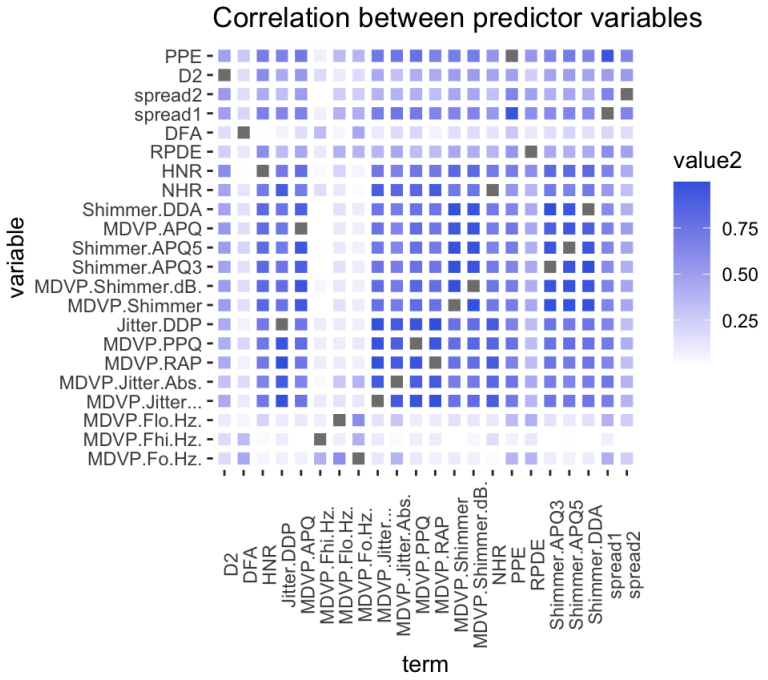

Figure 3: Distribution of predictor values based on health status

Next, we address the issue with having varying ranges amongst our variables by stan-

dardizing all of our variables so that they have a mean of 0 and standard deviation of 1. This is done to ensure that the scales of our variables does not impact the contribution of the variable to our outcome of interest. We then split our data randomly into 80% training and 20% testing with our training data set having only 159 recordings.

We can now use this cleaned and transformed data to build the model. We try two different methods for building our model, both using backward selection. The first uses backward selection using a likelihood ratio test (LRT) with `drop1` starting with a full model with all 7 predictors and their interactions. The LRT method compares the log likelihood ratio statistic from the two nested models to a Chi-squared distribution. This determines whether it is beneficial to add more parameters or if the simpler model is preferred. The second method performs backward selection using `stepAIC` with the same initial model. This method involves backward selection to minimize the model's Akaike Information Criteria (AIC) with a final set of features. This helps to simplify the model without sacrificing the model fit and performance.

With the LRT method, it was shown through multiple iterations of backward selection that none of the interaction effects were significant. Removing these interaction effects did not cause any significant changes in the estimates or standard errors of the model. The process was then continued with the main effects model. Again, we removed variables that were shown to not be significant using the LRT statistic and that did not significantly increase the model deviance. Thus, the final model from the LRT method included `MDVP.Fo.Hz, DFA, spread2` and `D2`. The model summary indicated that all variables were significant based on a threshold of $\alpha = 0.05$. Thus, based on this model selection process, these four speech signals were significantly related to PD.

Next, using step-wise backward selection based on AIC, `MDVP.Fo.Hz, DFA, spread2` and `D2` were included in the final model. This is interestingly the same final model as what was given using the LRT method. Lastly, for curiosity sake, the significance of variables from the main effects model was investigated. The summary showed that the only variables that were significant were `DFA, spread2` and `D2`. Thus, from our LRT and step AIC methods, we have an additional variable, `MDVP.Fo.Hz`. Since this variable appeared in both backward selection methods, we opt to keep it in our final model. With our final model, we use an ANOVA test against the full main effects model to test which model is preferred; we get a p-value of 0.75 meaning we cannot reject our null hypothesis that the simpler model is preferred. Therefore, our final model is the simpler model chosen through both methods of backward selection.

Diagnostic Tests

Now that we have our final logistic regression model, we want to run diagnostic tests to ensure that the assumptions of logistic regression are held and that the model fits the data well. For a logistic regression model, we are testing that we have linearity with the logit of our predictor, a lack of strongly influential outliers and the absence of multicollinearity.

Appendix A.2 shows a relatively linear relationship with the logit and our final predictor variables; thus, demonstrating the linearity assumption. This figure is using loess smoothing to clearly show the relationship without any noisy data values [6]. Next, to check for any outliers or extreme values, Figure 4 shows the Cook's Distance. This plot clearly shows that there are three indices which may be outliers. However, this only poses an issue if they are influential outliers. Hence, the standardized residual error is assessed to determine if these values are concerning. A quick check shows that the standardized residual error associated with each of these indices is within the (-3, 3) range; hence, we have no influential observations in our data. To further ensure there were no outliers or extreme values, Appendix A.3 shows a half-normal plot of the jackknife residuals. Through a similar analysis of investigating those potential outliers, it was concluded that there were no influential values. Thus, our second model check is satisfied.

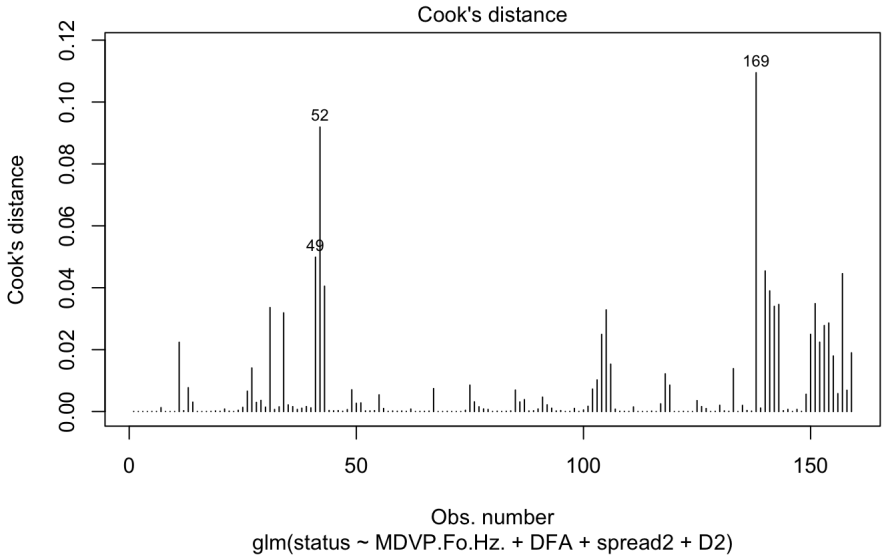

Figure 4: Cook's Distance

Lastly, we dealt with multicollinearity earlier by removing predictors that were highly correlated with one another. To ensure our logistic regression model holds, we conduct one more quality check. We check the variance inflation factor (VIF) of each of our variables, which measures the strength of correlation between independent variables in a regression

analysis [7]. As a rule of thumb, VIFs greater than 4 would require further analysis while VIFs greater than 10 indicate multicollinearity [8]. Appendix A.4 shows that the VIFs for each predictor are relatively small; hence, we confirm that there is no multicollinearity within our final model.

2.1.1 Model Performance

Using the above diagnostics, we have confirmed that our model is a decent fit for our data set. The next step would be to use it for distinguishing PD on a new test set. We predict based on our test set consisting of 36 voice recordings. Our logistic model predicts the probability of belonging to the positive class; hence, using a threshold of 0.5 on our probabilities, we are able to get binary values. To assess the models performance, we look at our confusion matrix in Appendix A.5 as well as accuracy, precision, recall and the F1-score in Table 1.

We have decent number of correct predictions; however, we note that we do predict 1 when the expected value is 0 a total of 7 times which is larger than the number of correct predictions for 0. This is represented in our precision score which is lower due to the false positive predictions. This is likely due to our imbalanced data as our model is more familiar with the data corresponding to an expected value of 1. This is furthermore exemplified by our high recall which represents the ratio of true positives by the sum of true positives and false negatives. The combination of these results is demonstrated with the F1-score, a weighted average of the recall and precision scores. Therefore, the classifier is good at predicting PD but not those without PD.

Accuracy	Precision	Recall	F1-Score
77.78%	77.42%	96.00%	85.71%

Table 1: Classification Metrics

2.2 Summary for a Statistician

To address the problem of interest of distinguishing those with PD from those without using speech signals, a binary logistic regression model was fit. Prior to fitting the model, to handle multicollinearity, the correlation between predictor variables was computed. Variables that had high correlations with multiple variables were dropped, leaving us with 7 variables. These variables were then standardized so that variables with a higher scale would not have a larger effect on the model than those with a smaller scale. The model was then selected

using backward selection with both LRT and `stepAIC` methods. With the LRT method, we used `drop1` to eliminate insignificant variables without causing the model's deviance to increase. This deemed that all of the interaction effects were insignificant and that the only significant main effects were average fundamental frequency (`MDVP.Fo.Hz`), detrended fluctuation analysis (`DFA`), nonlinear measure of fundamental frequency variation (`spread2`) and correlation dimensions (`D2`). The same final model was concluded by using the `stepAIC` method. Thus, this model had satisfied minimizing the AIC and was also the best model by comparison using the LRT method. In our final model, all of our predictor variables were statistically significant using a threshold of $\alpha = 0.05$. ANOVA tests against the full main effects model confirmed that the simpler model provided an adequate fit to the data; hence, using the parsimony principle, this model was the final model.

Our logistic regression model measures the log-odds ratio of PD; however, for interpretation, we are more interested in the odds-ratio. The odds ratio for the predictors and their respective 95% confidence intervals are summarized in Appendix A.6. From the table, we can interpret that a one unit increase in the average fundamental frequency would decrease the odds of having PD by 0.4222. Contrarily, a one unit increase in the detrended fluctuation analysis, nonlinear measure of fundamental frequency variation (`spread2`) and correlation dimensions would result in an increase in the odds of having PD by 1.9750, 2.6024 and 3.5402 respectively. Thus, increases in average fundamental frequency results in a significant increase in the odds for those without PD whereas the remaining predictors in the model result in a significant increase in the odds of having PD. Furthermore, we conclude that the interaction effects were not significant between any predictors. From our analysis, this model has proven to be a sufficient fit to the data; however, there is always room for improvement.

2.3 Limitations and Future Work

Although the model was able to distinguish those who had PD, it performed poorly with those without PD. This is one area of improvement which can be potentially remedied by collecting more non-PD data to solve the class imbalance problem. In addition, more variation in audio recordings would be beneficial. In the data set, multiple audio recordings were taken from a single patient; hence, recordings are not independent from one another. Thus, a larger subject size would also be beneficial in improving the statistical significance and accuracy of this model. Furthermore, alternative classification models such as SVM models, decision trees, random forests, etc. could be used to compare against the performance of the logistic regression model. These improvements can be investigated in future studies; however, this work provides an interesting application of using auditory features to identify PD.

References

[1] NIH National Institute on Aging. *Parkinson's Disease: Causes, Symptoms, and Treatments*. Accessed: 2022-12-01. 2022. URL: https://www.nia.nih.gov/health/parkinsons-disease#:~:text=Parkinson's%5C%20disease%5C%20is%5C%20a%5C%20brain,gradually%5C%20and%5C%20worsen%5C%20over%5C%20time..

[2] Parkinson's Foundation. *Who has Parkinson's?* Accessed: 2022-12-01. 2022. URL: https://www.parkinson.org/understanding-parkinsons/statistics#:~:text=More%5C%20than%5C%2010%5C%20million%5C%20people,have%5C%20Parkinson's%5C%20disease%5C%20than%5C%20women..

[3] M. A. Little et al. "Suitability of dysphonia measurements for telemonitoring of Parkinson's diseases". In: *IEEE transactions on bio-medical engineering* 56(4) (2009), p. 1015. URL: https://doi.org/10.1109/TBME.2008.2005954.

[4] Sam M.S. *PHONATION*. Accessed: 2022-12-01. 2013. URL: https://psychologydictionary.org/phonation/.

[5] Machine Learning Repository. *Parkinsons Data Set*. Accessed: 2022-12-01. 2008. URL: https://archive.ics.uci.edu/ml/datasets/parkinsons.

[6] Glen S. Team. *Lowess Smoothing in Statistics: What is it?* Accessed: 2022-12-01. URL: https://www.statisticshowto.com/lowess-smoothing/.

[7] Investopedia Team. *Variance Inflation Factor (VIF)*. Accessed: 2022-12-01. 2022. URL: https://www.investopedia.com/terms/v/variance-inflation-factor.asp.

[8] The Pennsylvania State University. *Detecting Multicollinearity Using Variance Inflation Factors*. Accessed: 2022-12-01. 2019. URL: https://online.stat.psu.edu/stat462/node/180/#:~:text=The%5C%20general%5C%20rule%5C%20of%5C%20thumb,of%5C%20serious%5C%20multicollinearity%5C%20requiring%5C%20correction..

[9] S. Zelcer et al. "Multidimensional voice program analysis (MDVP) and the diagnosis of pediatric vocal cord dysfunction". In: *Annals of allergy, asthma & immunology official publication of the American College of Allergy, Asthma, & Immunology* 88(6) (2002), pp. 601–608. URL: https://doi.org/10.1016/S1081-1206(10)61892-3.

A Appendix

A.1 Variable Definitions

Variable	**Definition**
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%)	Kay Pentax MDVP jitter as a percentage
MDVP:Jitter(Abs)	Kay Pentax MDVP absolute jitter in microseconds
MDVP:RAP	Kay Pentax MDVP Relative Amplitude Perturbation
MDVP:PPQ	Kay Pentax MDVP five-point Period Perturbation Quotient
Jitter:DDP	Average absolute difference of differences between cycles, divided by the average period
MDVP:Shimmer	Kay Pentax MDVP local shimmer
MDVP:Shimmer(dB)	Kay Pentax MDVP local shimmer in decibels
Shimmer:APQ3	Three point Amplitude Perturbation Quotient
Shimmer:APQ5	Five point Amplitude Perturbation Quotient
MDVP:APQ	Kay Pentax MDVP 11-point Amplitude Perturbation Quotient
Shimmer:DDA	Average absolute difference between consecutive differences between the amplitudes of consecutive periods
NHR	Noise-to-Harmonics Ratio
HNR	Harmonics-to-Noise Ratio
RPDE	Recurrence Period Density Entropy
DFA	Detrended Fluctuation Analysis
D2	Correlation dimensions
Spread1, Spread2	Non-linear measures of fundamental frequency variation
PPE	Pitch period entropy

Table A.1: Variable Definitions

A.2 Linearity Condition

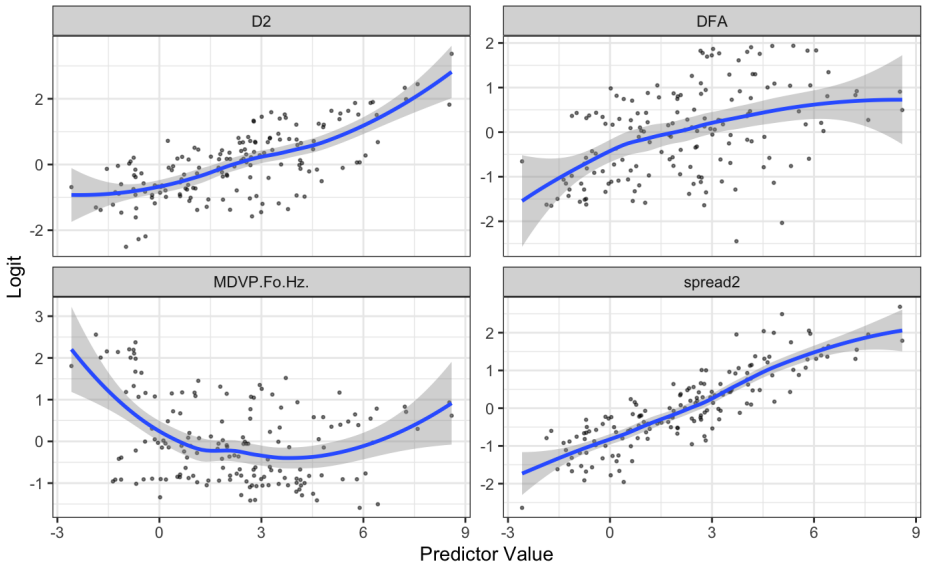

Figure A.2: Linearity Assumption for all predictors

A.3 Outlier Analysis

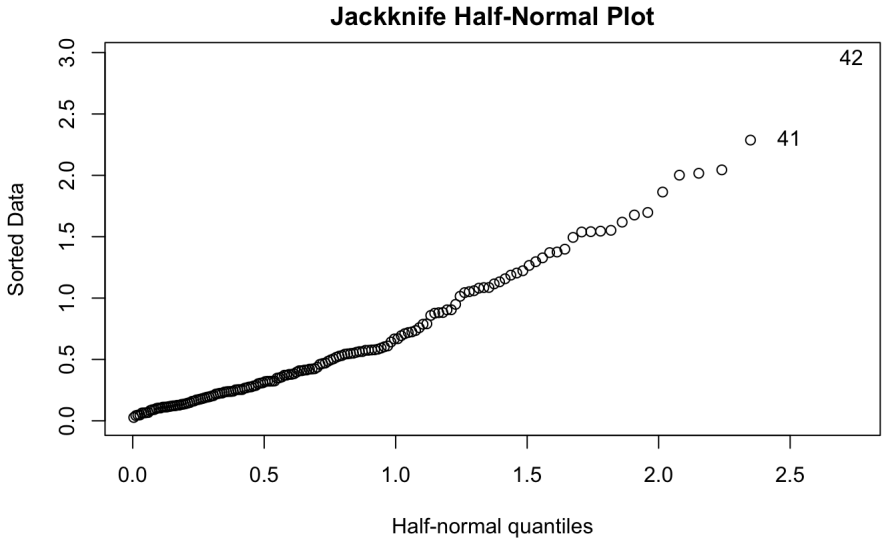

Figure A.3: Half-Normal Jackknife Plot

A.4 Variance Inflation Factors

MDVP.Fo.Hz	DFA	spread2	D2
1.290146	1.165339	1.130790	1.299703

Table A.4: Variance Inflation Factors for predictors

A.5 Confusion Matrix

Confusion Matrix	Actual Health Status		
		No PD	**PD**
Predicted Health Status	**No PD**	4	1
	PD	7	24

Table A.5: Confusion Matrix

A.6 Odds Ratio Estimates

Variable	**Estimate**	**2.5%**	**97.5%**
MDVP.Fo.Hz	0.422	0.221	0.747
DFA	1.975	1.064	3.844
spread2	2.602	1.268	5.795
D2	3.540	1.704	8.568

Table A.6: Odds Ratio of Variables