Early Biomarkers of Parkinson's Disease Based on Natural Connected Speech

Anja Probst[1]

[1] University of Geneva

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.
Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.
One sentence clearly stating the **general problem** being addressed by this particular study.
One sentence summarizing the main result (with the words "**here we show**" or their equivalent).
Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.
One or two sentences to put the results into a more **general context**.
Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords
Word count: X

The authors made the following contributions. Anja Probst: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Anja Probst, 24 rue du Général-Dufour, 1211 Genève 4. E-mail: anja.probst@etu.unige.ch

Introduction

Context of the Project

Patients with the neurodegenerative disease Parkinson's have numerous symptoms ranging from cognitive impairments to motor symptoms. Those symptoms may appear relatively late in the disease when the neurodegeneration has already widely spread in different areas of the brain (mainly Basal Ganglia). Main symptoms of PD are motor dysfunctions including abnormalities in the production and sound of speech of such patients (up to 90%). These abnormalities in speech and voice are called hypokinetic dysarthria which is characterized by a decreased quality of the speech, where the voice, sound formation as well as the articulation is impaired. As I mentioned before, often motor impairments are detected relatively late in the disease. To improve diagnostics and to detect the disease in a much earlier stage, the detection of biomarkers related to neurodegeneration could lead to a better prognosis and therapy of PD.

Therefore, the investigation of prodromal speech changes could be an appropriate and suitable approach. To investigate this approach, an automated speech monitoring system was developed, that uses a segmentation method for the precise estimation of voiced and unvoiced segments of speech, respirations, and pauses. Further proposed was a set of acoustic speech features based on the segmentation algorithm applicable to connected speech, allowing the description of complex vocal disturbances due to neurodegeneration including respiratory deficits, dysphonia, imprecise articulation, and dysrhythmia.

In this data analysis project, the main focus was to explore, if there are any speech patterns that support the usage of an automated speech monitoring system to detect prodromal parkinsonian neurodegeneration based on natural connected speech.

130 subjects were tested. 30 subjects with early, untreated Parkinson's disease (PD) where the disease is already manifested. 50 subjects with REM sleep behaviour disorder (RBD), which is a disease where its relatively likely to develop PD in a later phase. As a control group, 50 healthy subjects (HD) were included.

Manual Variable Selection

Due to the constraints of this project, I reduced the data set from originally 62 variables to the best fitting 7. As I am looking specificially into the aspect of speech, and to evaluate if speech is a good predictor for PD, I chose speech related variables that were assessed empirically, and should represent the hypothesis the best. Note that patient group will be extracted from the variable Participant_code. The resulting data set is summarized in Table 1

Data Description

For each sample in this data set ($n = 130$), we have the following information:

- Demographic information:

  - Age (years)
  - Gender (M for male, F for female)

- Speech examination - Speaking task of reading passage: speakers read a standardized, phonetically-balanced text of 80 words twice
  - Duration_Of_Pause_Intervals_Reading: Duration of pause intervals (DPI) describes the quality of speech timing, as pauses can be heavily influenced by the ability to properly initiate speech, it is measured in miliseconds (ms)
  - Rate_Of_Speech_Timing_Reading: Rate of speech time (RST) includes voiced, unvoiced and pause intervals, it is measured in intervals per minute (-/min)
- Speech examination - Speaking task of monologue: participants were instructed to provide monologue about their interests, job, family or current activities for approximately 90 seconds
  - Duration_Of_Pause_Intervals_Monologue: Duration of pause intervals (DPI) describes the quality of speech timing, as pauses can be heavily influenced by the ability to properly initiate speech, it is measured in milliseconds (ms)
  - Rate_Of_Speech_Timing_Monologue: Rate of speech time (RST) includes voiced, unvoiced and pause intervals, it is measured in intervals per minute (-/min)
- Group: based on Participant Code
  - PD: subjects with Parkinson's disease
  - RBD: subjects with REM sleep behaviour disorder
  - HC: healthy controls

Table 1
Summary of the Data Set used in this Analysis

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	130	64.331	10.134	34	58.25	72	83
Gender	130						
... F	27	20.8%					
... M	103	79.2%					
Speech.Timing.Rate.Reading	130	327.277	47.385	140	297.25	358.75	457
Speech.Timing.Rate.Monologue	130	288.338	52.892	112	258	328.75	412
Pause.Interval.Duration.Reading	130	166.646	46.488	96	138.25	185	388
Pause.Interval.Duration.Monologue	130	229.069	79.697	117	177	263.25	611
Group	130						
... HC	50	38.5%					
... PD	30	23.1%					
... RBD	50	38.5%					

```
## 'data.frame':    130 obs. of  7 variables:
##  $ Age                              : int  58 68 68 75 61 58 79 59 73 66 ...
```

```
##  $ Gender                           : Factor w/ 2 levels "F","M": 1 1 2 2 2 2 2 1 2 2 ..
##  $ Speech.Timing.Rate.Reading       : int  354 340 211 140 269 317 269 338 374 281 ...
##  $ Speech.Timing.Rate.Monologue     : int  333 285 247 112 230 181 289 370 288 258 ...
##  $ Pause.Interval.Duration.Reading  : int  146 173 377 360 211 186 214 145 117 213 ...
##  $ Pause.Interval.Duration.Monologue: int  158 295 280 397 206 611 251 118 194 246 ...
##  $ Group                            : Factor w/ 3 levels "HC","PD","RBD": 2 2 2 2 2 2 2 2
```

Data Pre-Processing

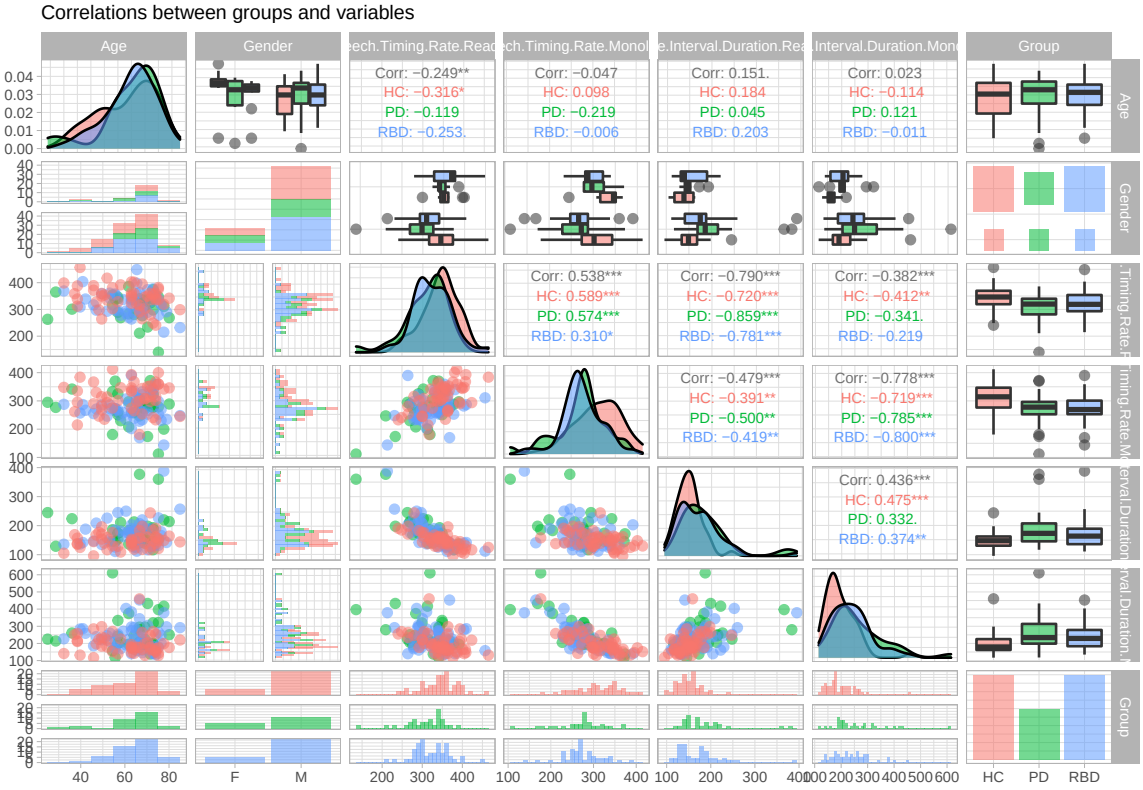


Figure 1

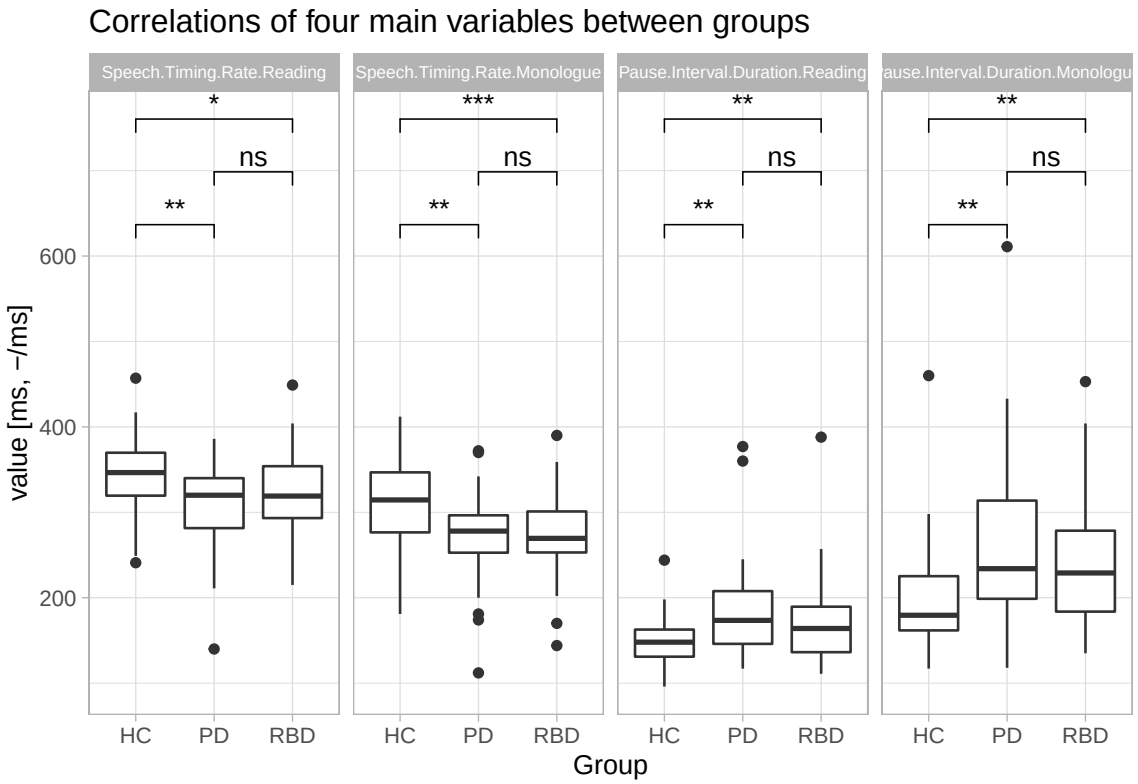

Correlations of four main variables between groups
Speech.Timing.Rate.Reading
Speech.Timing.Rate.Monologue
Pause.Interval.Duration.Reading
ause.Interval.Duration.Monologu
*
ns
**
***
**
ns
**
**
ns
**
**
ns
600
400
200
value [ms, −/ms]
HC
PD
RBD
Group

Figure 2

Data Analysis

Binomial Regression

Above, we have seen that there are almost no differences between the groups PD and RBD, so in a first step, we will limit our investigation to creating a binomial model predicting the group HC or PD. Indeed, the paper from which the data was extracted discusses the hard problem of differentiating PD from RBD, which might very well be impossible with generalised linear models. We will revisit this problem in the section Multinomial Regression.

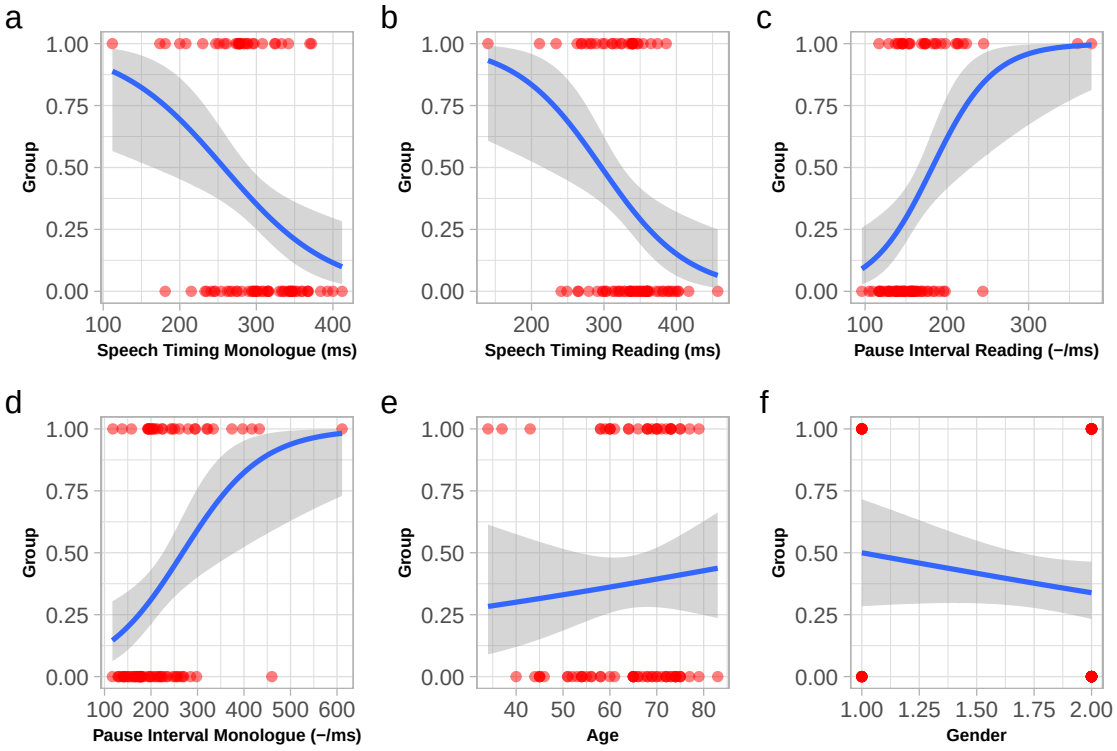


Figure 3

In a first step, a simple linear regression model based on a single predictor is built and visualized for each of the selected variables. As can be seen by visual inspection of the data points (red), none of the predictors is sufficient to predict the response variable (Group) on its own, given the respective overlap between the two groups. Hence, a series of multiple linear regression models have to be built and evaluated. As I would have to test 64 models (all possible combinations plus intercept only) to be certain to have found the best one, I chose to use the automated model selection function `dredge` from the R package `MuMIn`. Starting from the global binomial model `Group ~ .` as an input, `dredge` enumerates all possible models and evaluates them based on their AIC.

Table 2
A full regression table of the best model (selected using dredge) without interactions.

Predictor	*b*	95% CI	*z*	*p*
Intercept	-5.28	[−8.60, −2.53]	-3.42	.001
GenderM	-1.76	[−3.15, −0.49]	-2.63	.009
Pause Interval Duration Monologue	0.01	[0.00, 0.02]	2.03	.042
Pause Interval Duration Reading	0.02	[0.01, 0.05]	2.39	.017

```
m.full <- glm(
  data=df.binom, Group ~ .,
  family=binomial,
  na.action = "na.fail"
)

d <- dredge(m.full, rank = "AIC")

m.best.no.interactions = get.models(d, 1)[[1]]
```

Looking at interactions ...

```
##
## Call:
## glm(formula = Group ~ Pause.Interval.Duration.Reading:Pause.Interval.Duration.Monologue +
##     Gender, family = binomial, data = df.binom)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8600  -0.7777  -0.5119   0.9345   2.0621
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                   -1.954e+00
## GenderM                                                       -1.650e+00
## Pause.Interval.Duration.Reading:Pause.Interval.Duration.Monologue  7.070e-05
##                                                                Std. Error
## (Intercept)                                                     7.296e-01
## GenderM                                                         6.492e-01
## Pause.Interval.Duration.Reading:Pause.Interval.Duration.Monologue  1.976e-05
##                                                                z value
## (Intercept)                                                     -2.678
## GenderM                                                         -2.541
```

```
## Pause.Interval.Duration.Reading:Pause.Interval.Duration.Monologue   3.577
##                                                                  Pr(>|z|)
## (Intercept)                                                      0.007396 **
## GenderM                                                          0.011061 *
## Pause.Interval.Duration.Reading:Pause.Interval.Duration.Monologue 0.000347 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.850  on 79  degrees of freedom
## Residual deviance:  81.561  on 77  degrees of freedom
## AIC: 87.561
##
## Number of Fisher Scoring iterations: 5
```

PCA. As there has been significant correlation between the predictors in the ggpairs plot as well as some extreme changes in coefficients when adding additional variables, there exists the possbility of collinearity negatively affecting the models. Indeed, we observe variance inflation factors of more than 2.5 between all experimental predictors. This warrants and attempt at solving the potential collinearity issue.

```
##                               Age                            Gender
##                          1.204231                          1.407117
##        Speech.Timing.Rate.Reading      Speech.Timing.Rate.Monologue
##                          3.073223                          2.916549
##   Pause.Interval.Duration.Reading Pause.Interval.Duration.Monologue
##                          2.599297                          2.652620
```

```
## Importance of components:
##                            PC1    PC2     PC3     PC4
## Standard deviation     1.6786 0.8762 0.53096 0.36403
## Proportion of Variance 0.7045 0.1919 0.07048 0.03313
## Cumulative Proportion  0.7045 0.8964 0.96687 1.00000
```

Running ggpairs shows, that there is no longer any correlation between the variables.

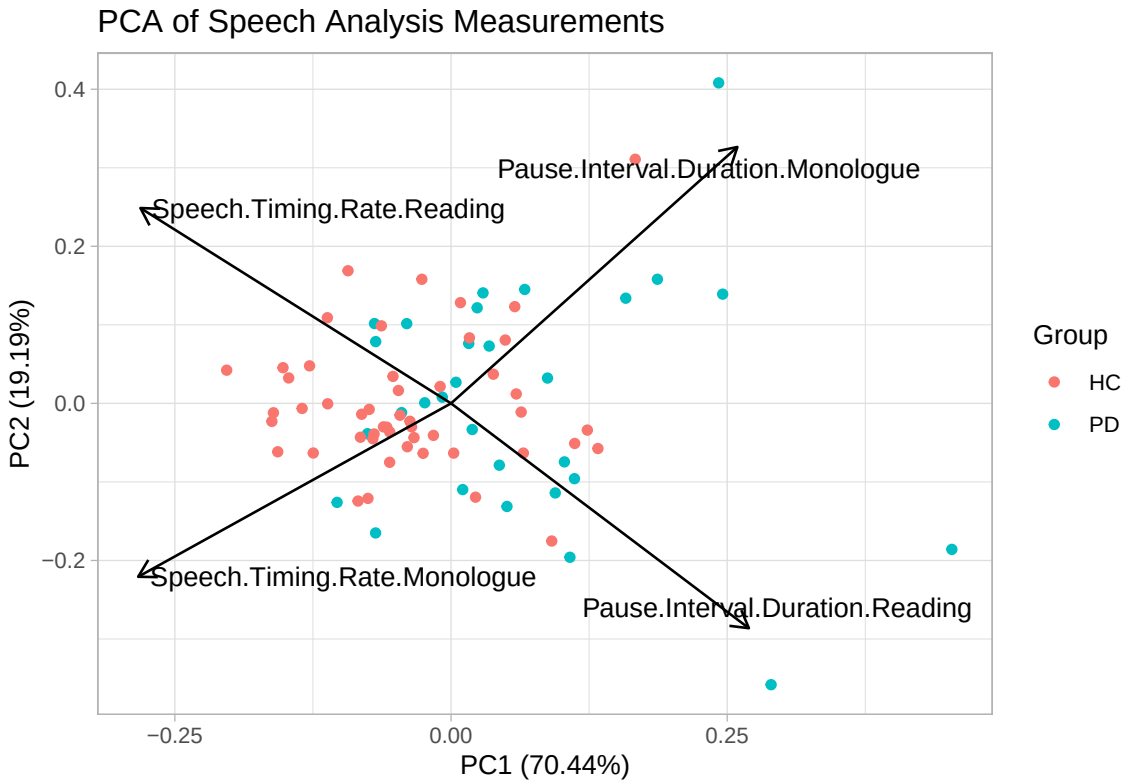
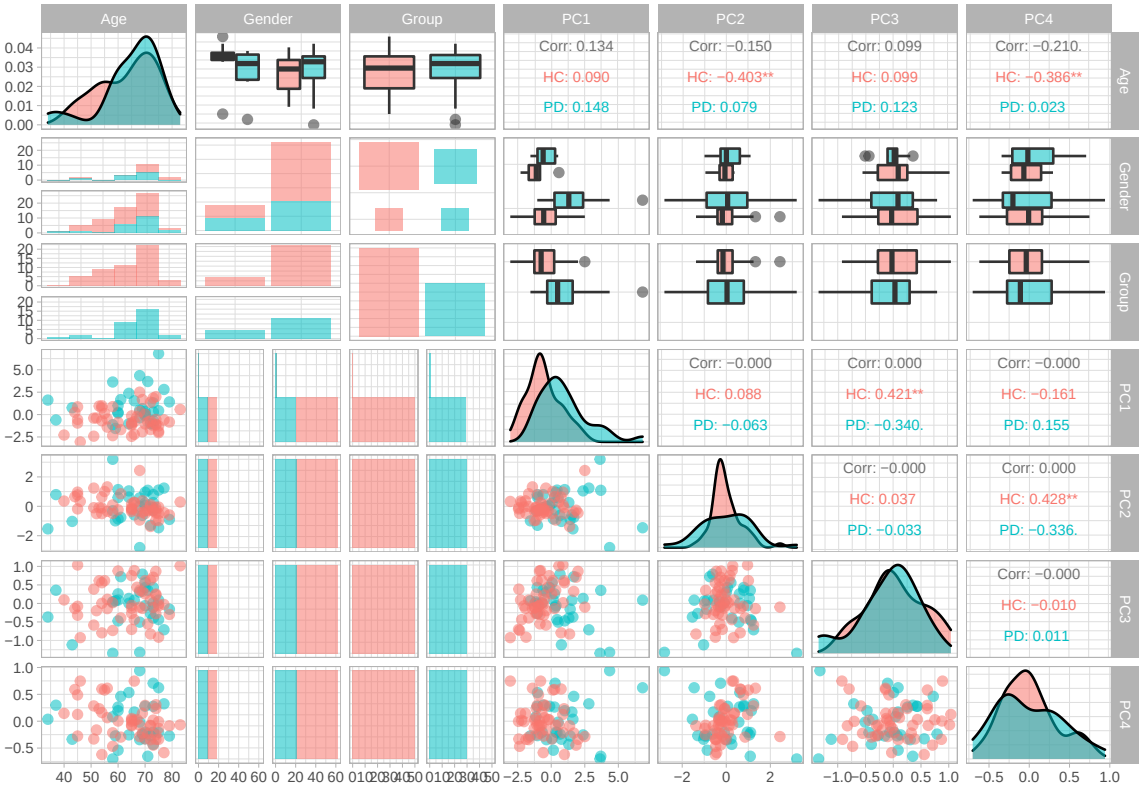
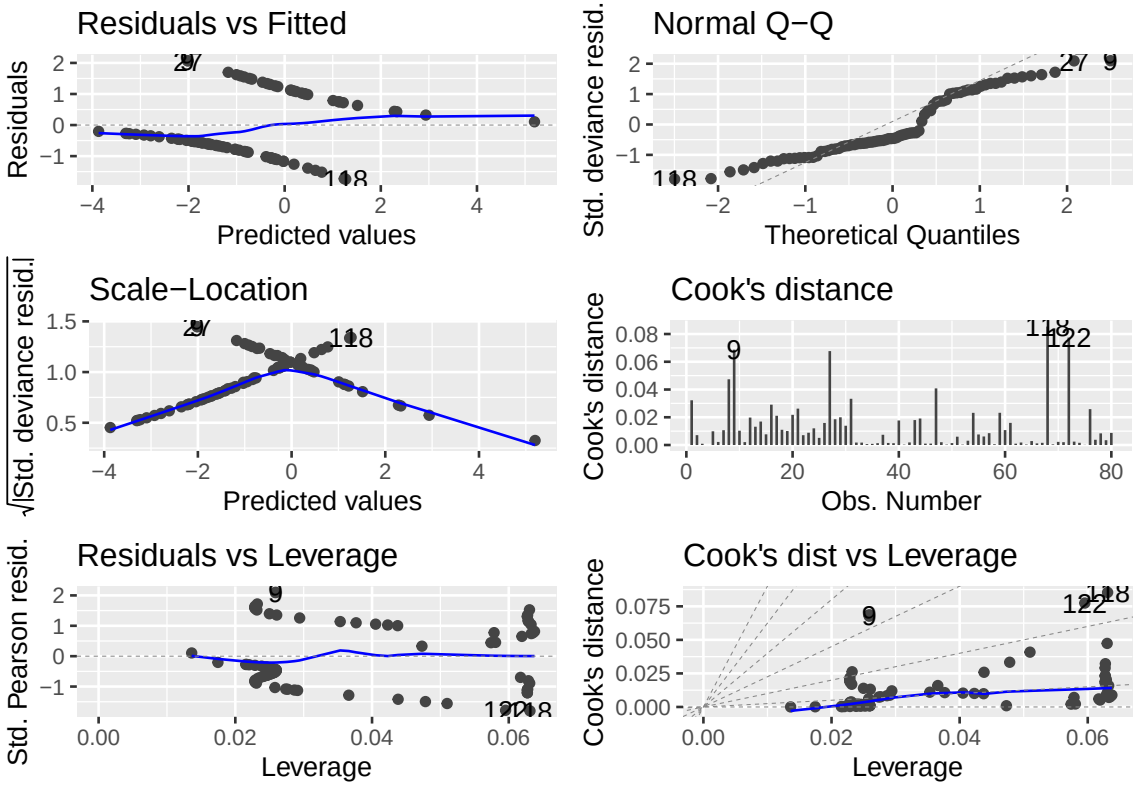

PCA of Speech Analysis Measurements
Pause.Interval.Duration.Monologue
Speech.Timing.Rate.Reading
Speech.Timing.Rate.Monologue
Pause.Interval.Duration.Reading
PC2 (19.19%)
PC1 (70.44%)
Group
HC
PD

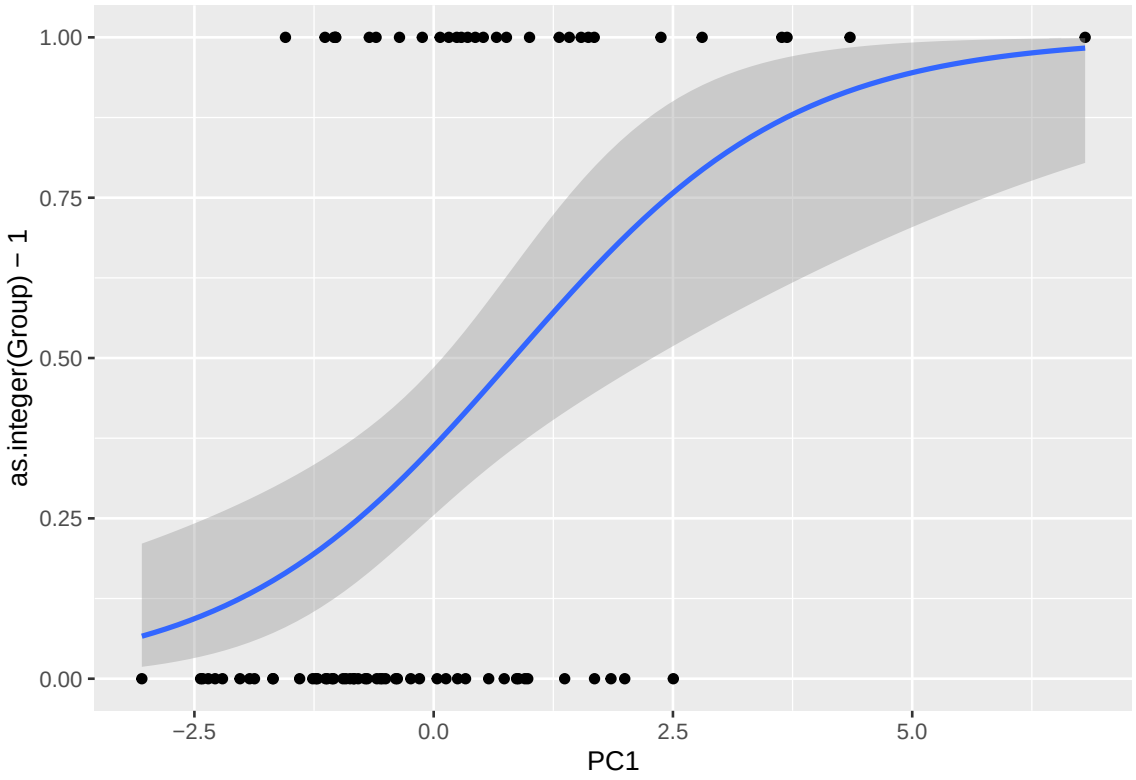
Figure 4


Age
Gender
Group
PC1
PC2
PC3
PC4
Corr: 0.134
HC: 0.090
PD: 0.148
Corr: −0.150
HC: −0.403**
PD: 0.079
Corr: 0.099
HC: 0.099
PD: 0.123
Corr: −0.210.
HC: −0.386**
PD: 0.023
Corr: −0.000
HC: 0.088
PD: −0.063
Corr: 0.000
HC: 0.421**
PD: −0.340.
Corr: −0.000
HC: −0.161
PD: 0.155
Corr: −0.000
HC: 0.037
PD: −0.033
Corr: 0.000
HC: 0.428**
PD: −0.336.
Corr: −0.000
HC: −0.010
PD: 0.011

```
##
## Call:
## glm(formula = Group ~ PC1 + Gender, family = "binomial", data = df.biom.pca.joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7424  -0.7892  -0.4558   0.9875   2.0717
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7417     0.5401   1.373 0.169734
## PC1           0.9199     0.2456   3.746 0.000179 ***
## GenderM      -1.8062     0.6809  -2.653 0.007986 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.850  on 79  degrees of freedom
## Residual deviance:  81.454  on 77  degrees of freedom
## AIC: 87.454
##
## Number of Fisher Scoring iterations: 5
```




```
## Analysis of Deviance Table
##
## Model 1: Group ~ PC1 + Gender
## Model 2: Group ~ Gender + Pause.Interval.Duration.Monologue + Pause.Interval.Duration.Re
##     1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        77     81.454
## 2        76     80.274  1   1.1799   0.2774
```



Multinomial Regression

To predict over all three groups (HC, PD, RBD), we have to use a more complex multinomial model.

```
## # weights:  15 (8 variable)
## initial  value 99.973718
## iter  10 value 86.144554
## iter  20 value 84.687289
## final  value 84.684738
## converged
```

```
##
##       HC PD RBD
##   HC  28  1  10
##   PD   5  5   8
##   RBD  8  1  25
```

```
## [1] 63.74
```

Conclusion

Manual Model Plot

References