Basic Analysis

To predict severity and progression of Parkinson's disease, using the total UPDRS score, many audio quality features were analysed. These include test time, various frequency measures (jitter), various amplitude measures (shimmer), noise and tonal control (NHR and HNR), dynamic complexity measure (RPDE), signal fractal scaling component (DFA), and fundamental frequency variation (PPE).

First, a linear model incorporating all predictors and using the response variable "total_UPDRS" was applied to preliminarily identify significant predictors. This full model shows numerous significant variables. Namely, test_time, NHR, HNR, RPDE, DFA, PPE, Jitter_Abs., Shimmer (general), Shimmer.APQ5, and Shimmer.APQ11 (Figure A1.a).

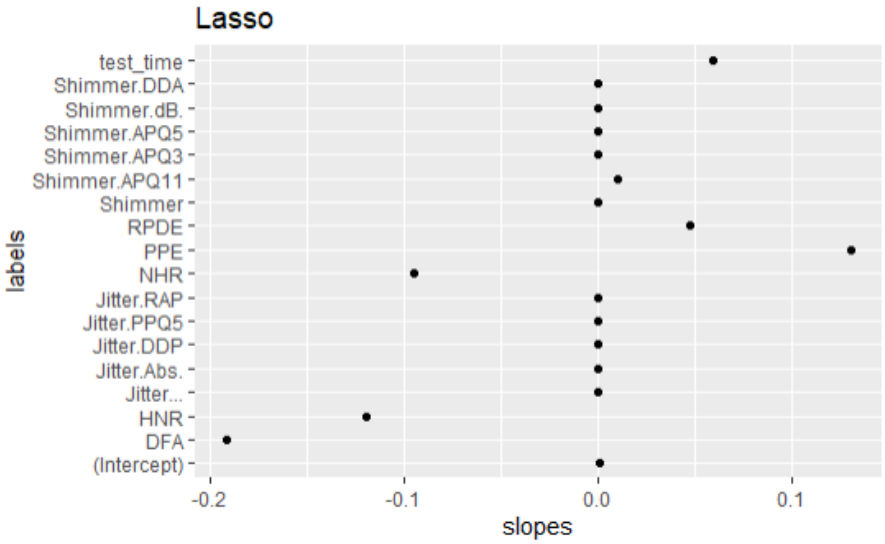

Figure 1 Variable Selection Using Lasso
Shown is a plot of the slope estimates for each predictive variable after Lasso regression was applied.

To improve the model and reduce bias from unnecessary predictors, Lasso regression was applied to select for significant features. From this analysis, 7 features of interest were identified: DFA, HNR, NHR, PPE, RPDE, test_time, and Shimmer.APQ11 (Figure 1). These variables were then used to create a reduced model giving a more accurate look at the data. The results of a linear regression using only these predictors showed that all seven were significant predictors of total UPDRS (Figure A1.b).

To ensure the accuracy of the predictions for the two models, the residuals were examined to check for normality and heteroscedasticity. According to the scatter and Q-Q plots of the residuals, the full and reduced models have heteroscedastic and roughly normally distributed residuals (Figure 2).

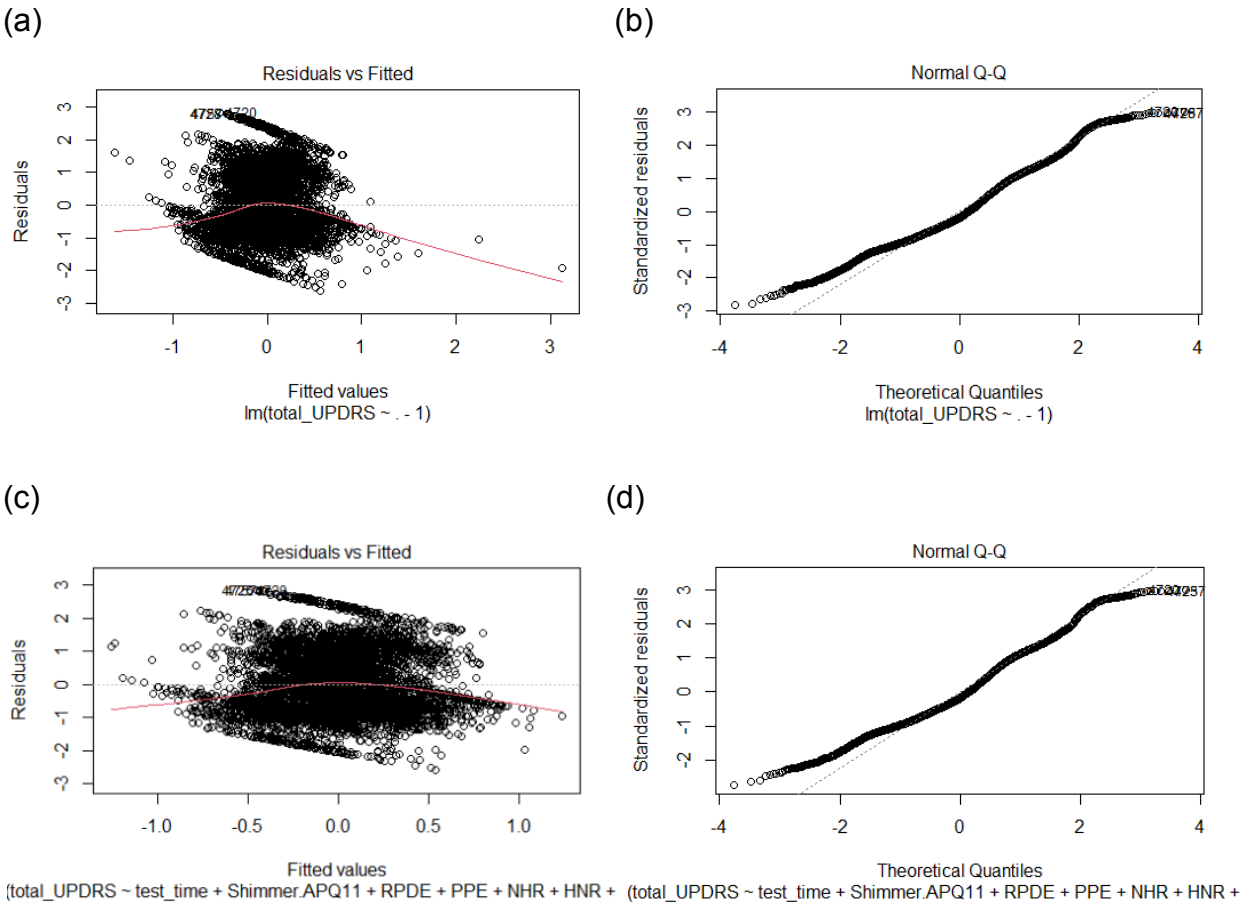

Figure 2 Residual Plots
(a) and (b) show the scatter plot and Q-Q plots respectively for the residuals of the full model. (c) and (d) show the same, but for the reduced model.

Further analysis was performed relating to Shimmer.APQ11 since it had the smallest slope of the features selected by lasso. The null hypothesis,

$$H_0: \beta_{Shimmer.APQ11} = 0$$

states that the coefficient associated with Shimmer.APQ11 is equal to zero, meaning that the predictor has no significant impact on the model. Conversely, the alternative hypothesis,

$$H_a: \beta_{Shimmer.APQ11} \neq 0$$

states that the coefficient is non-zero and should stay in the model.

$$t = \frac{\widehat{\beta_{Shimmer.APQ11}} - 0}{S.E(\widehat{\beta_{Shimmer.APQ11}})}$$

$$t = \frac{0.05281}{0.02152} = 2.454$$

The t-statistic of 2.454 on 5686 degrees of freedom has a p-value of ~0.01966. The null hypothesis can, therefore, be rejected at the $\alpha = 0.05$ level and the slope for Shimmer.APQ11 is expected to be non-zero.

A 95% confidence interval for this predictor, shows, with 95% confidence, that the true value of the coefficient lies in the interval $(0.01063, 0.09499)$. This is according to the following calculation, where 't_k(df)' represents the t value at a critical point 'k' with degrees of freedom 'df',

$$\widehat{\beta_j} \pm t_{1-(0.5\alpha)}(n - p - 1) \cdot S.E(\widehat{\beta_j})$$

$$0.05281 \pm 1.96 \cdot 0.02152 = (0.01063, 0.09499)$$

Based on the interval, the impact of Shimmer.APQ11 on total UPDRS could be as small as 0.01 per 1 unit increase, which is not a particularly large effect. Future analyses may benefit from determining whether this variable is truly useful in predicting UPDRS.

An anova test was performed to compare the full and reduced models and determine whether a difference was established between the predictions of the two models. The null hypothesis,

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

states that for each predictor 'j', the given slope 'β_j' is equal to zero, meaning that the predictor has no significant impact on the model. Conversely, the alternative hypothesis,

$$H_a: \beta_j \neq 0 \text{ for at least one } j,\ j = 1, \ldots, p$$

states that there is at least one regressor that is significant. From the associated F-statistic with 5686 degrees of freedom, a p-value of 3.11×10^{-11} was obtained (Figure A2), which is smaller than an $\alpha = 0.05$ threshold. This indicates a high level of significance in the differences between models and the null hypothesis that the models are the same can be rejected. Based on this analysis and the results of the Lasso regression, it appears that the seven predictors used in the reduced model are ideal predictors of UPDRS.

Appendix A

R output

(a)

```
Call:
lm(formula = total_UPDRS ~ . - 1, data = tot_df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6178 -0.6973 -0.1820  0.7077  2.8463

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
test_time      0.08829    0.01240   7.121 1.20e-12 ***
Jitter...      0.06863    0.11639   0.590 0.555426
Jitter.Abs.   -0.10601    0.03220  -3.293 0.000999 ***
Jitter.RAP   -13.80989   14.19452  -0.973 0.330642
Jitter.PPQ5   -0.02475    0.06819  -0.363 0.716636
Jitter.DDP    13.88979   14.19602   0.978 0.327903
Shimmer        0.28441    0.16238   1.751 0.079917 .
Shimmer.dB.   -0.10831    0.10810  -1.002 0.316407
Shimmer.APQ3 -25.52941   60.42833  -0.422 0.672694
Shimmer.APQ5  -0.20346    0.08931  -2.278 0.022755 *
Shimmer.APQ11  0.20057    0.04799   4.179 2.97e-05 ***
Shimmer.DDA   25.32614   60.42825   0.419 0.675150
NHR           -0.23069    0.03502  -6.587 4.87e-11 ***
HNR           -0.24593    0.02859  -8.602  < 2e-16 ***
RPDE           0.05478    0.01778   3.082 0.002068 **
DFA           -0.27176    0.01561 -17.414  < 2e-16 ***
PPE            0.17956    0.02588   6.937 4.43e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9456 on 5858 degrees of freedom
Multiple R-squared:  0.1083,    Adjusted R-squared:  0.1057
F-statistic: 41.85 on 17 and 5858 DF,  p-value: < 2.2e-16
```

(b)

```
Call:
lm(formula = total_UPDRS ~ test_time + Shimmer.APQ11 + RPDE +
    PPE + NHR + HNR + DFA - 1, data = tot_df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5971 -0.7127 -0.1682  0.7171  2.8278

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
test_time      0.08513    0.01243   6.848 8.24e-12 ***
Shimmer.APQ11  0.05281    0.02152   2.454 0.014156 *
RPDE           0.05668    0.01676   3.382 0.000725 ***
PPE            0.19601    0.02053   9.547  < 2e-16 ***
NHR           -0.23690    0.01981 -11.959  < 2e-16 ***
HNR           -0.17864    0.02678  -6.670 2.79e-11 ***
DFA           -0.26997    0.01451 -18.601  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9505 on 5868 degrees of freedom
Multiple R-squared:  0.09745,    Adjusted R-squared:  0.09638
F-statistic: 90.51 on 7 and 5868 DF,  p-value: < 2.2e-16
```

Figure A1 Summary of Linear Models

(a) gives the output for the full model including all predictive variables. (b) gives the output for the reduced model, including only those variables deemed important by Lasso regression.

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   5868 5301.6
2   5858 5237.9 10    63.642 7.1176 3.113e-11 ***
```

Figure A2 Summary of ANOVA

Shown is the output for an F-test comparing the full and reduced models. This includes degrees of freedom, sums of squares, the F-statistic and the p-value for the analysis.