Predicting Parkinson's Disease Using Various Vocal Measures

Tri Cao

December 8, 2021

Introduction

Parkinson's Disease, or more commonly referred to as Parkinson's is a neurological disease that impairs the motor skills in an individual. Nerve cells in the brain are damaged or destroyed as a result of this disease which affects the communication between nerve cells. Parkinson's disease is made up of a series of five stages, with each subsequent stage being defined by more substantial symptoms. In stage one, symptoms are mild and can be difficult to catch. They typically occur on one side of the body, and include hand tremors, muscle stiffness, and changes in posture, walking, and facial expressions. It can take between months and years to advance to stage two consisting of a severe stage one symptom. Stage three is characterized by issues with balance and reflexes. In stage four movement becomes more limited and assistance is necessary for everyday tasks. In stage five patients commonly experience delusions and dementia. At this stage, quality of life tends to be very poor as individuals require assistance to be wheeled around or may be entirely bed ridden. Throughout these stages, non-motor symptoms are frequent and can include cognitive decline, such as poor memory and planning, mood disorders, issues with sleep, loss of smell, vision problems, and the main focus of this paper, speech problems.

There are currently no treatments to reverse the disease or prevent the disease from worsening. Most treatments that are available today focus on reducing or lessening the symptoms that an individual with Parkinson's experiences. For example, some of the treatments to assist with motor function can include; deep brain stimulation, physical therapy, and exercise.

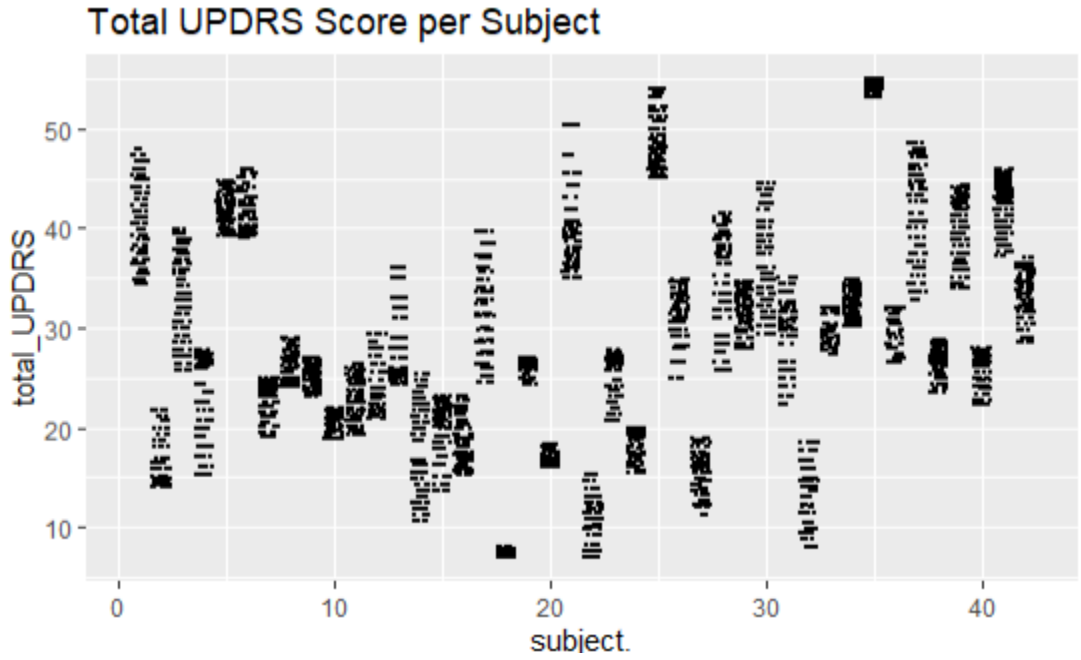
Methods

For the purposes of our analysis we will be primarily focused on the speech problems that are typically found within the early-stage Parkinson's patients. The Parkinson's Telemonitoring Dataset that we will be analyzing is from the UCI Machine Learning Repository. This dataset was created by University of Oxford's Athanasios Tsanas and Max Little and contains important measurements that are used when determining the Unified Parkinson's Disease Rating Scale score (UPDRS score) of an individual. These features include the subject's id, the age of a subject (subject), the sex of a subject (sex, male: 0; female: 1) , the duration since the trial started (test_time), the motor UPDRS score (motor_UPDRS), the total UPDRS score (total_UPDRS), various measurements of variation in fundamental frequency (Jitter), several measures of amplitude variation (Shimmer), two measurements of ration of noice and tonal components in the voice (NHR, HNR), a nonlinear dynamical complexity measure (RPDE), signal fractal scaling exponent (DFA), and a nonlinear measure of fundamental

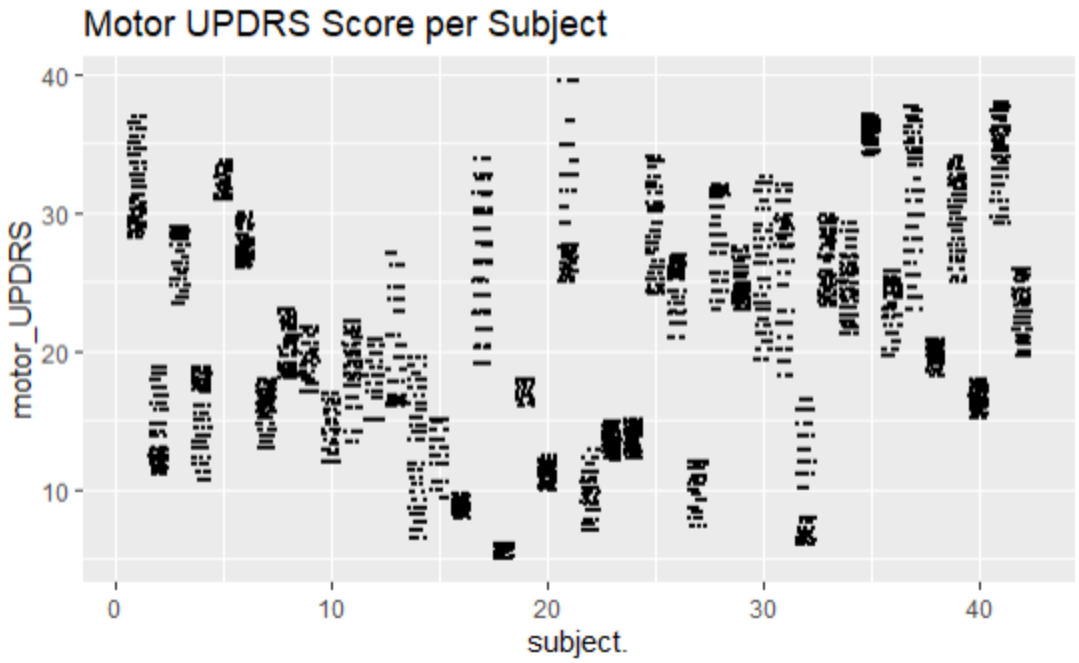
frequency variation (PPE). This data will be used to run linear regression analysis on the variables and optimize an equation that can be used to predict a patient's score. Subsequently, a Linear Mixed Effects Regression model will be fit to the data to determine if some of the features are significant in the prediction. These analyses will shed some light onto what the common symptoms of Parkinson's Disease are in order to come up with a better way to classify the stage of Parkinson's an individual is in.

Initial Results

Upon an initial look at the data there are two main features in the dataset, the total_UPDRS score with values ranging from (0-199) and the motor_UPDRS score which is a subset of the total UPDRS score with values (0-52). These scores are what determines the Parkinson's stage the subject is in but we want to pick one in order to eliminate any correlation or unknown relationships between the data.The initial graph of the Subject's UPDRS measurements is shown in the plot below.



From this data we can see that each subject's scores are relatively uniform and there is not a lot of variance in the measurements for each subject. The figure below will show the relationship between the patient and the motor UPDRS score.



When these two are put together we can see that they are fairly uniform and don't contain many outliers. For our final analysis we will focus on the total UPDRS score because it encompasses all of the features present in the data including the motor UPDRS score.

Basic Analysis

To predict severity and progression of Parkinson's disease, using the total UPDRS score, many audio quality features were analyzed. 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).

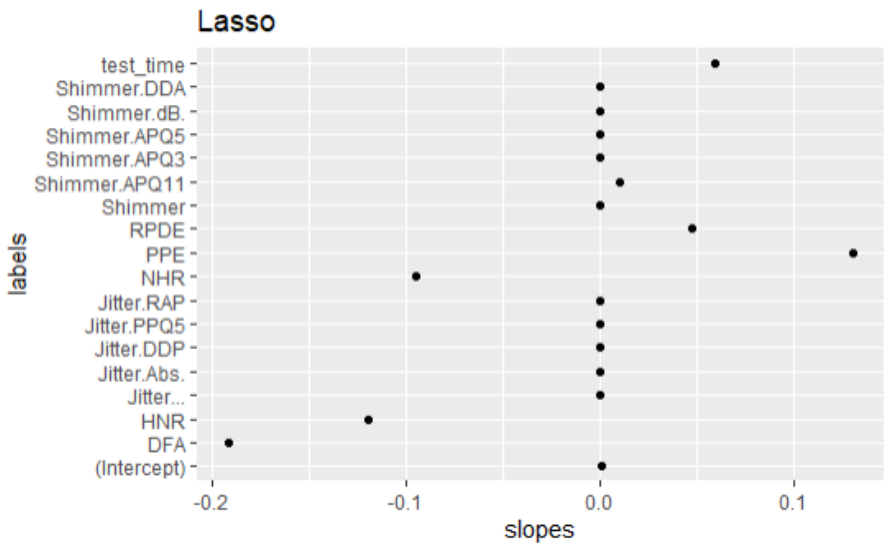
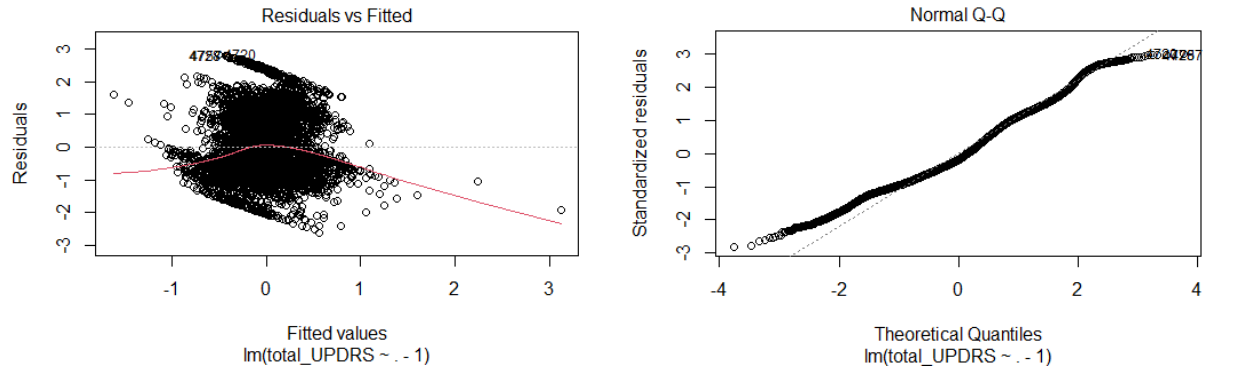


Figure 1 Variable Selection Using Lasso
Shown is a plot of the slope estimates for each predictive variable after Lasso regression was applied.

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).

(a) (b)



(c) (d)

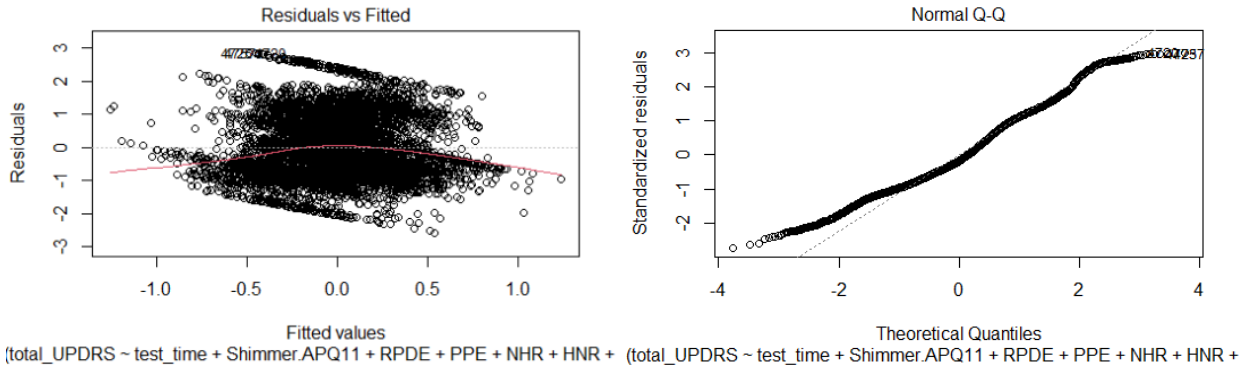


Figure 2 Residual Plots

(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.

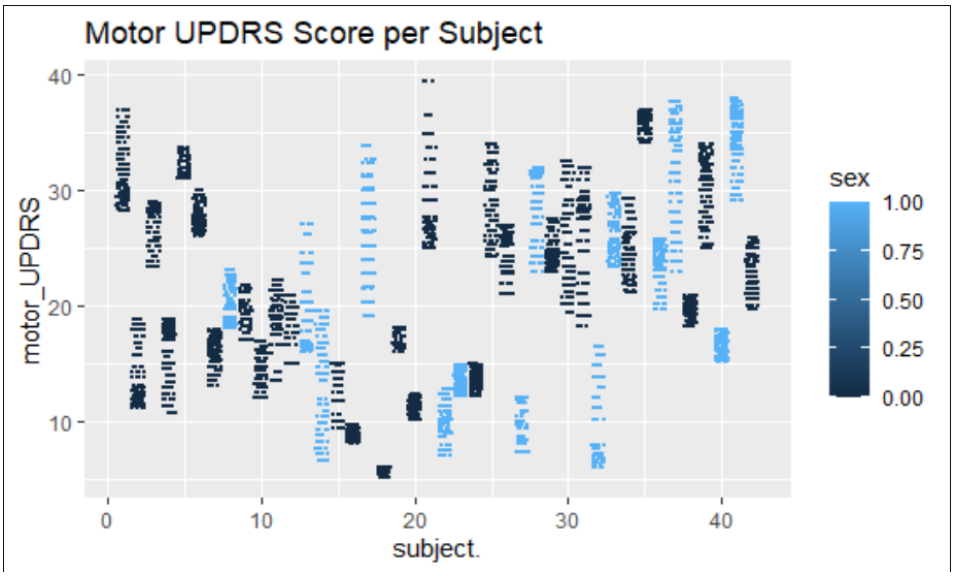
LMER Analysis

After we have reviewed the results of Parkinson's DIsease in a patient, we would like to see if there is a difference in the total_UPDRS score between the two sexes, Male and Female. We would like to see if there is a correlation between the sex of a patient and the severtiy of their UPDRS score. To do this we have created a LMER model with the random effects being the subject and the sex of the subject as the fixed effect. The general equation for our model is as follows:

$$y_{ij} = \mu + v_i + \epsilon_{ij} \; for \; 1 \leq i \leq I \; and \; 1 \leq j \leq n_0$$

Where y_{ij} is the total Unified Parkinson's Disease Rating Scale score of a patient, μ is the mean score of the patients, v_i is the random effect adjustment for each subject and ϵ_{ij} is the residual score for each subject.

The initial look at the data with the sexes color coded shows that females (sex label = 1 and light blue) appears to have smaller values compared to males (sex label = 0 and darkblue). Our data is also not balanced so we will be fitting a maximum likelihood to the model.



For our analysis, we will be performing two hypothesis tests. The first one being a significance test for the random effect and the second being a significance test for the fixed effects.

This first hypothesis will be determining if the random effect is significant in the full model by comparing it to the model without this effect. With the reduced model having only the sex of a subject. A difference of deviance is performed on the two models and a p-value is determined using a Chi-Squared distribution. The null hypothesis is

$$H_0: \ Full\ Model \ = \ Reduced\ Model \ = \ 0,$$

and the alternative is

$$H_a: \ Full\ Model \ \neq \ Reduced\ Model.$$

A Chi-squared distribution is fit to the model and we get a value of 15511 with a p-value of about 0. This prompts us to reject the null hypothesis and we are certain that the random effect is a significant predictor in the model.

When a test for the fixed effect is performed. The reduced model has only the subject as a random effect. A difference of deviance is then performed on the two models and a p-value is determined using a Chi-Squared distribution. The null hypothesis is once again

$$H_0: \ Full\ Model \ = \ Reduced\ Model \ = \ 0,$$

and the alternative is

$$H_a: \ Full\ Model \ \neq \ Reduced\ Model.$$

A Chi-Squared fit to the model yields a value of 0.634 and has a p-value of 0.4259. We therefore fail to reject the null hypothesis at the critical value of 0.05. This implies that the sex of a patient does not significantly predict the total_UPDRS score.

Conclusion

Overall, it can be seen that there are some good predictors to better predict the score of a person who may be experiencing early stage Parkinson's Disease. From the basic analysis, we were able to determine that the most influential factors in determining an individual's UPDRS score were the time, Shimmer.APQ11, RPDE, PPE, NHR, HNR and DFA. If we were given the measurements of these individual scores, we could make a good estimate for their score on the Unified Parkinson's Disease Rating Scale and could therefore treat the symptoms before it progresses more. Since the world is becoming increasingly virtual coupled with the fact that people with Parkinson's Disease may not want to leave the comfort of their own home where their symptoms can be controlled, these findings may be crucial in the medical world because it would allow for in-home measurements and studies. For the LMER analysis, we were able to come to the conclusion that there is not significant correlation between the UPDRS score and a patient's sex. This sheds some light into the intricacies of Parkinson's Disease, and we can claim that Parkinson's Disease does not favor or discriminate against either sex.

In the future, if we are given a large unlabeled dataset with many different features presented, we could accurately predict the score of an individual and recommend them different actions to take based on the stage that they are classified into. These models along with future automation could allow for quick and accurate Parkinson's screening. This could eliminate the resources required to have a specialist be present to screen an individual and would allow for a more prompt diagnosis and treatment regiment.

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

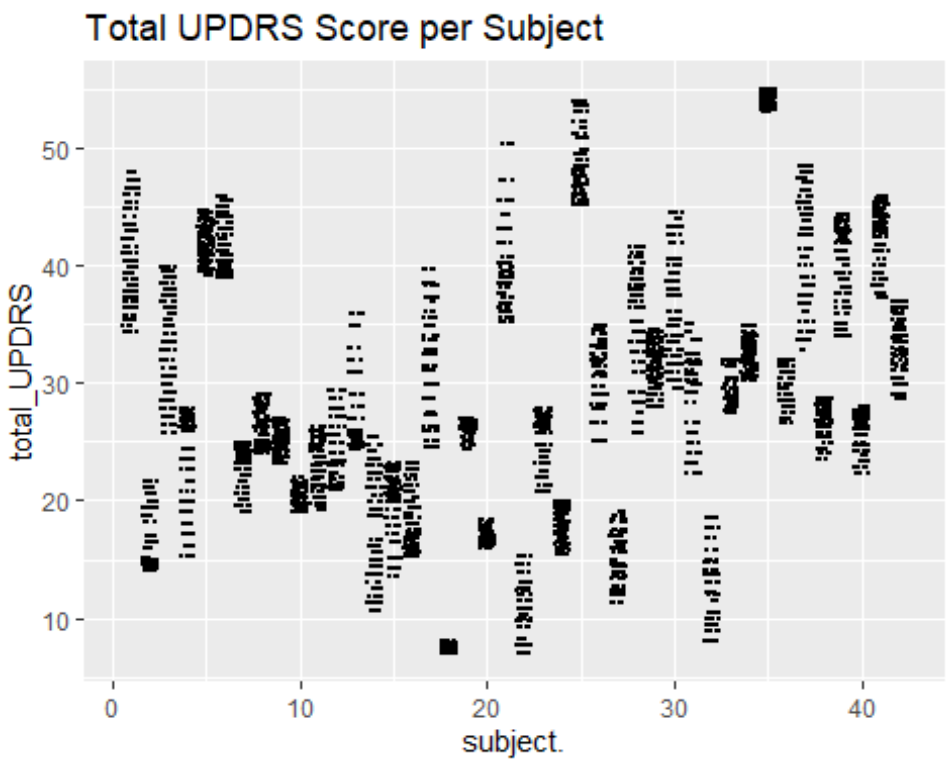
Shown is the output for an F-test comparing the full and reduced models. This includes degrees of freedom, sums of squares, the F-statistic and the p-value for the analysis.

Appendix B

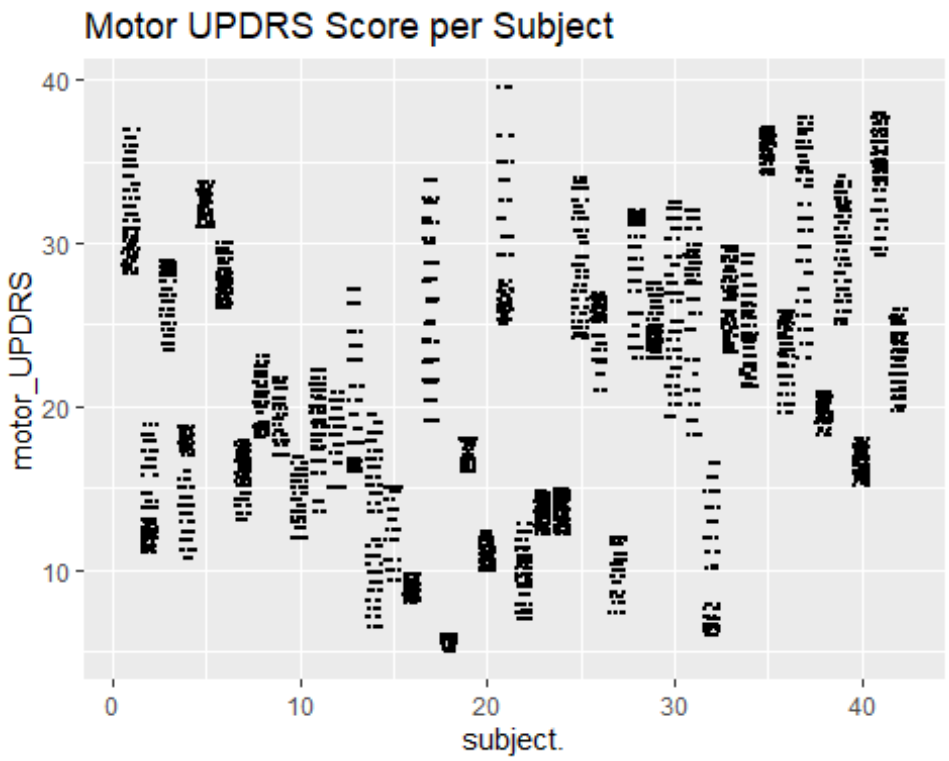
Code

```
categorical_col = c(1,2,3)
full_data = read.csv("D:/R/data_467/data467_project/parkinsons_updrs.data")
data = full_data[-categorical_col]
for (i in 1:length(data)){
  data[,i] = scale(data[,i], center = TRUE, scale = TRUE)
}
subject_data = full_data
head(data)
```

```
##    test_time motor_UPDRS total_UPDRS   Jitter... Jitter.Abs. Jitter.RAP
## 1 -1.6319513   0.8491244   0.5027024  0.08289818  -0.2842180  0.3274246
## 2 -1.5005486   0.8796314   0.5490563 -0.56074568  -0.7566586 -0.5337008
## 3 -1.3692936   0.9101384   0.5953167 -0.23892375  -0.5393359 -0.3000125
## 4 -1.2576661   0.9359710   0.6346615 -0.15535673  -0.4851442 -0.3448294
## 5 -1.1080747   0.9706604   0.6874638 -0.49851492  -0.6638379 -0.6585480
## 6 -0.9769133   1.0011674   0.7337243 -0.46651053  -0.5871358 -0.5753165
##   Jitter.PPQ5 Jitter.DDP    Shimmer Shimmer.dB. Shimmer.APQ3 Shimmer.APQ5
## 1  -0.0286346  0.3284775 -0.3245661  -0.3516122   -0.2096910   -0.4233205
## 2  -0.4761714 -0.5347790 -0.5339707  -0.5731071   -0.5451114   -0.5655438
## 3  -0.3207395 -0.2989574 -0.6690579  -0.5644210   -0.7415288   -0.7023662
## 4  -0.1706673 -0.3448413 -0.4236559   0.0696623   -0.4605008   -0.4497248
## 5  -0.5297687 -0.6596258 -0.6582200  -0.5861362   -0.7830786   -0.6513579
## 6  -0.4520527 -0.5753275 -0.4553956  -0.4211008   -0.5360460   -0.4065177
##   Shimmer.APQ11 Shimmer.DDA         NHR          HNR        RPDE       DFA
## 1    -0.5434195  -0.2096865 -0.29869541 -0.009203976 -1.21396224 -1.478374
## 2    -0.5299101  -0.5451066 -0.35193510  1.282540523 -1.05502898 -1.247774
## 3    -0.6454902  -0.7415238 -0.19935262  0.318684263 -0.78479294 -1.540008
## 4    -0.3928152  -0.4607479 -0.07174823  0.644475136 -0.53644116 -1.062024
## 5    -0.4648651  -0.7833254 -0.34334103  1.036216523 -0.68913591 -1.297843
## 6    -0.2427112  -0.5357894 -0.37997892  0.295147154 -0.01963576 -1.139737
##          PPE
## 1 -0.6506028
## 2 -1.2184810
## 3 -0.1032714
## 4  1.2369692
## 5 -0.2839301
## 6 -0.2687386

#TOTAL UPDRS
ggplot(full_data, aes(subject., total_UPDRS)) +
  geom_jitter(size = 0.01)+
  ggtitle('Total UPDRS Score per Subject')
```

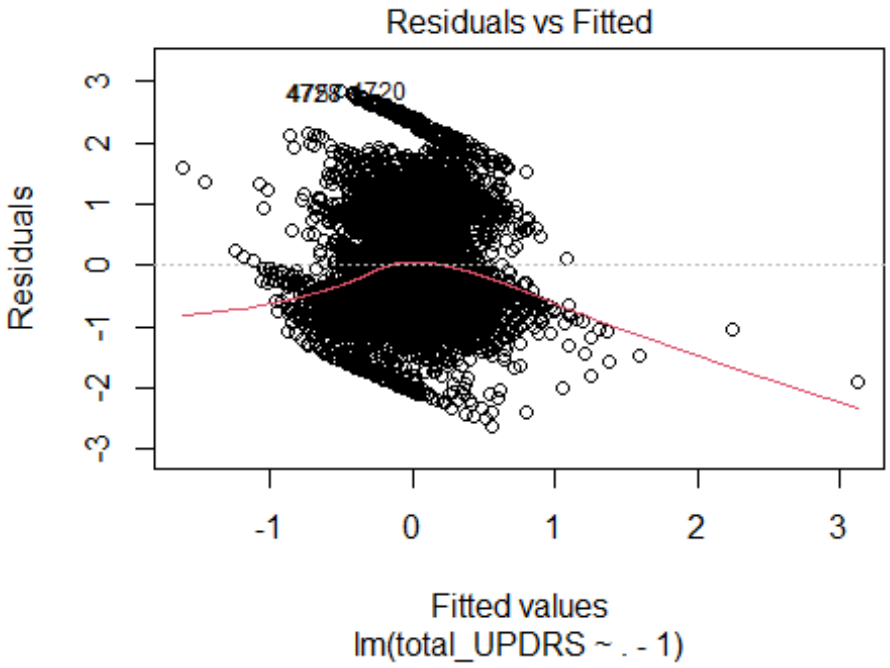


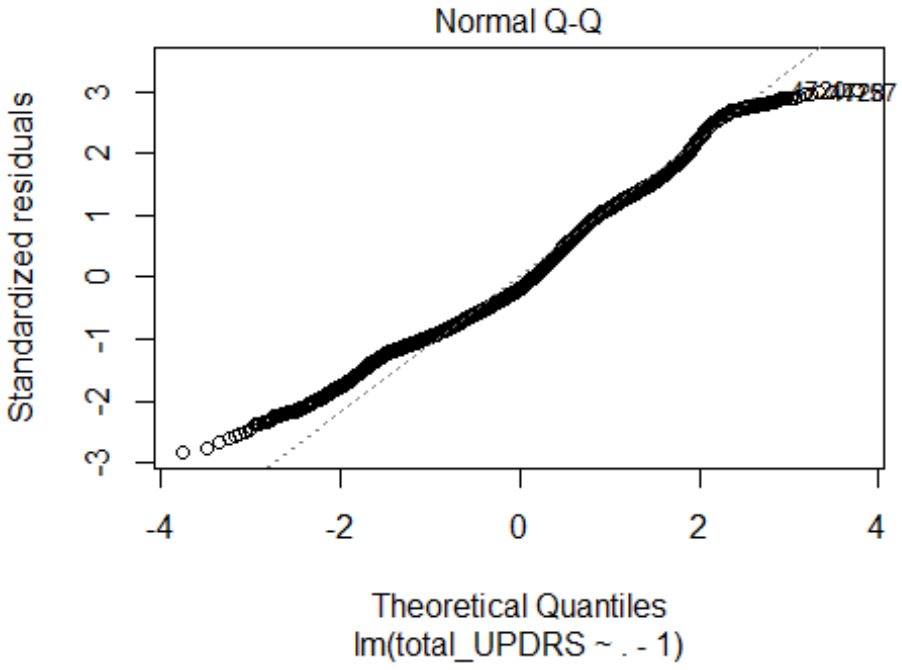

```
#MOTOR UPDRS
ggplot(full_data, aes(subject., motor_UPDRS)) +
  geom_jitter(size = 0.01)+
  ggtitle('Motor UPDRS Score per Subject')
```

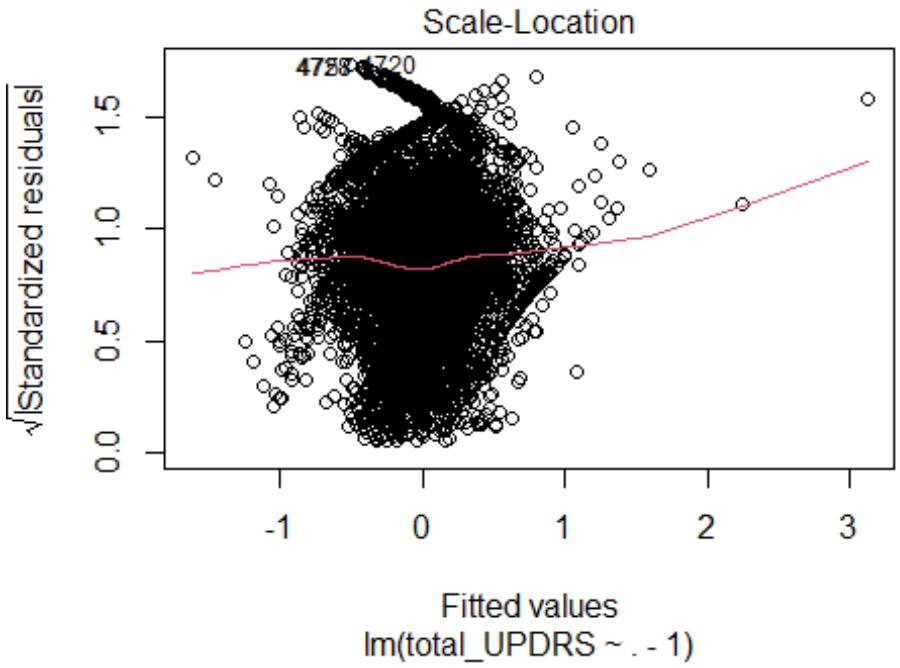


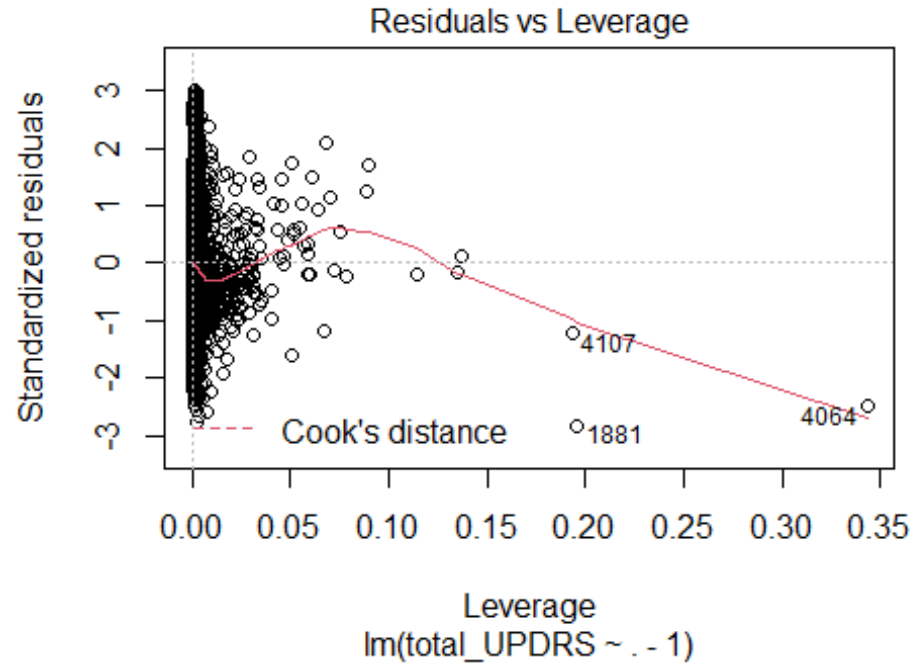

```
set.seed(7)

tot_df = data[-2] #Removing motor_UPDRS

#FULL MODEL
full_model = lm(total_UPDRS ~.-1, data = tot_df)
summary(full_model)

##
## 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

plot(full_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(total_UPDRS ~ . - 1)

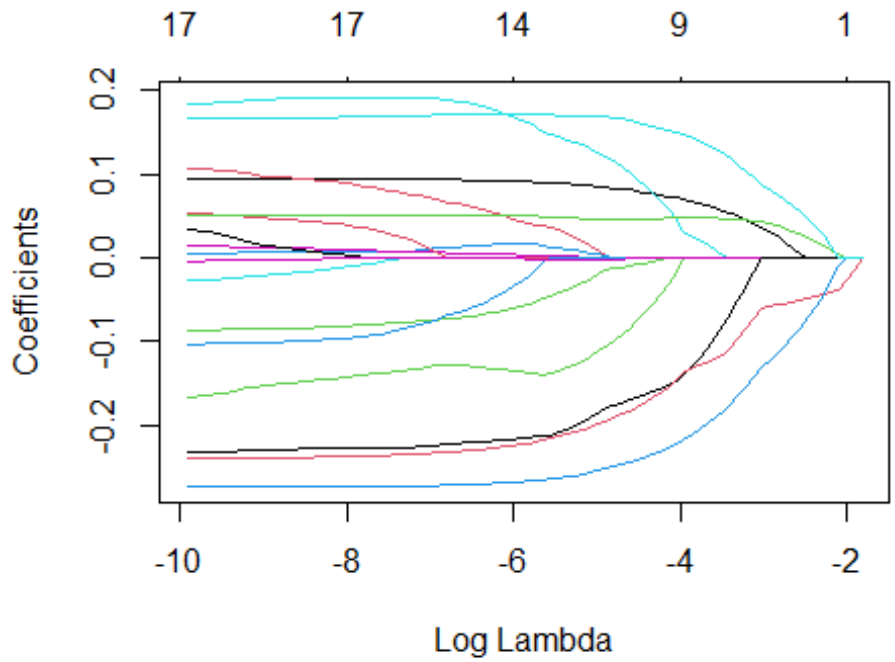

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(total_UPDRS ~ . - 1)



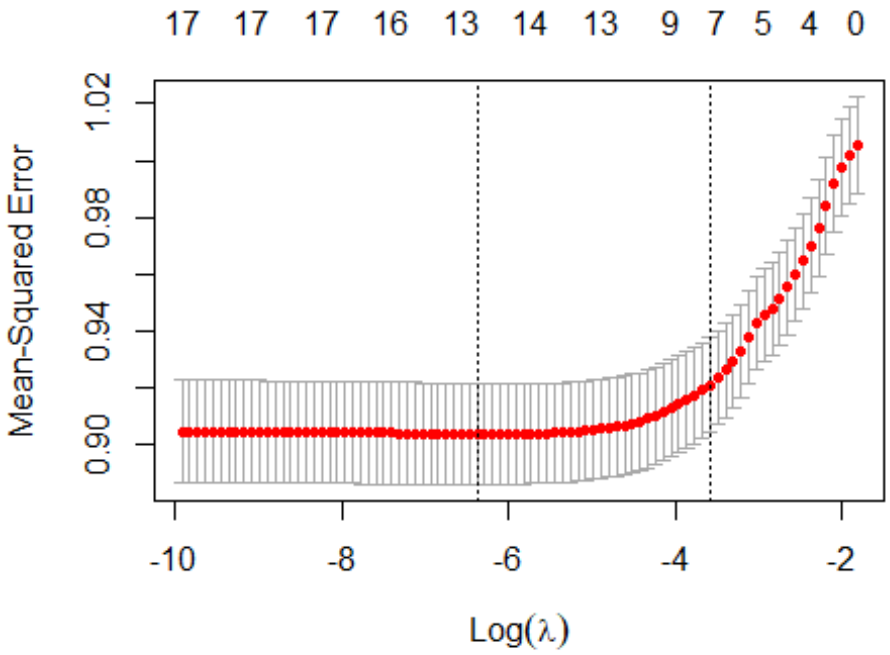



```r
#CROSS VALIDATION
tot_split = initial_split(tot_df , prop = 0.7, strata = 'total_UPDRS')
tot_train = training(tot_split); tot_test = testing(tot_split)
```

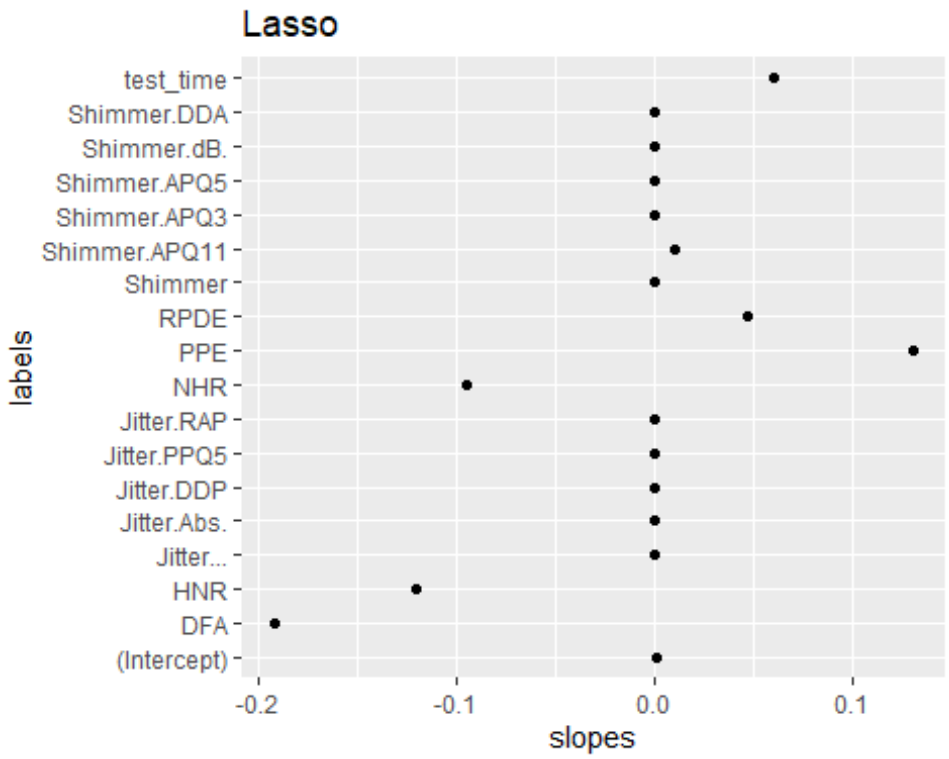
```r
tot_train_y = tot_train$total_UPDRS
tot_train_x = model.matrix(lm(total_UPDRS ~ .-1, data = tot_train))


# #RIDGE
# tot_ridge = glmnet(x = tot_train_x, y = tot_train_y, alpha = 0)
# plot(tot_ridge, xvar = 'lambda')
#
# tot_ridge_cv = cv.glmnet(x = tot_train_x, y = tot_train_y, alpha = 0)
# plot(tot_ridge_cv)
#
# ridge_coef = coef(tot_ridge, tot_ridge_cv$lambda.1se)
# ridge_coef_df = as.data.frame(as.matrix(ridge_coef))
# colnames(ridge_coef_df) = c('slopes')
# ridge_coef_df$labels = rownames(ridge_coef_df)
# ggplot(data = ridge_coef_df, mapping = aes(x = slopes, y = labels)) +
geom_point()+ggtitle('Ridge')
# ridge_coef_df

#LASSO
tot_lasso = glmnet(x = tot_train_x, y = tot_train_y, alpha = 1)
plot(tot_lasso, xvar = 'lambda')
```




```r
tot_lasso_cv = cv.glmnet(x = tot_train_x, y = tot_train_y, alpha = 1)
plot(tot_lasso_cv)
```




```
lasso_coef = coef(tot_lasso, tot_lasso_cv$lambda.1se)
lasso_coef_df = as.data.frame(as.matrix(lasso_coef))
colnames(lasso_coef_df) = c('slopes')
lasso_coef_df$labels = rownames(lasso_coef_df)
ggplot(data = lasso_coef_df, mapping = aes(x = slopes, y = labels)) +
geom_point() +ggtitle('Lasso')
```

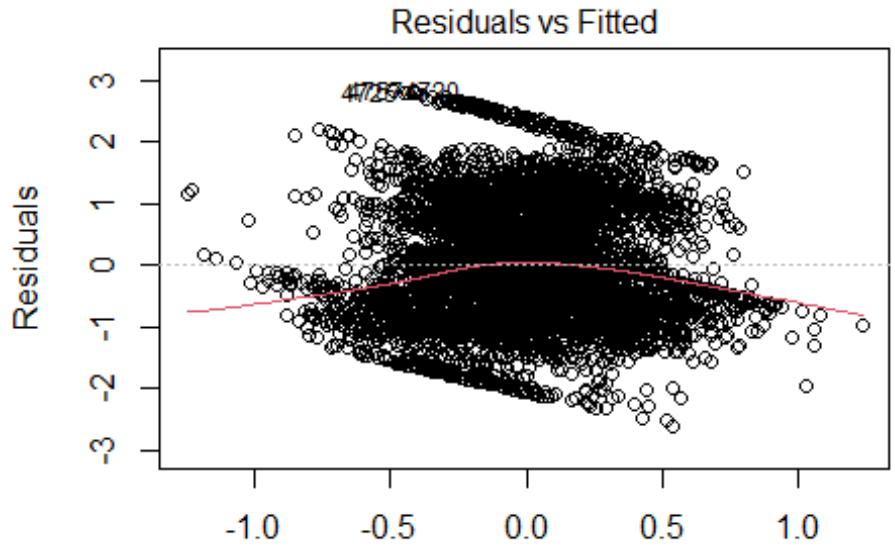



```
lasso_coef_df
```

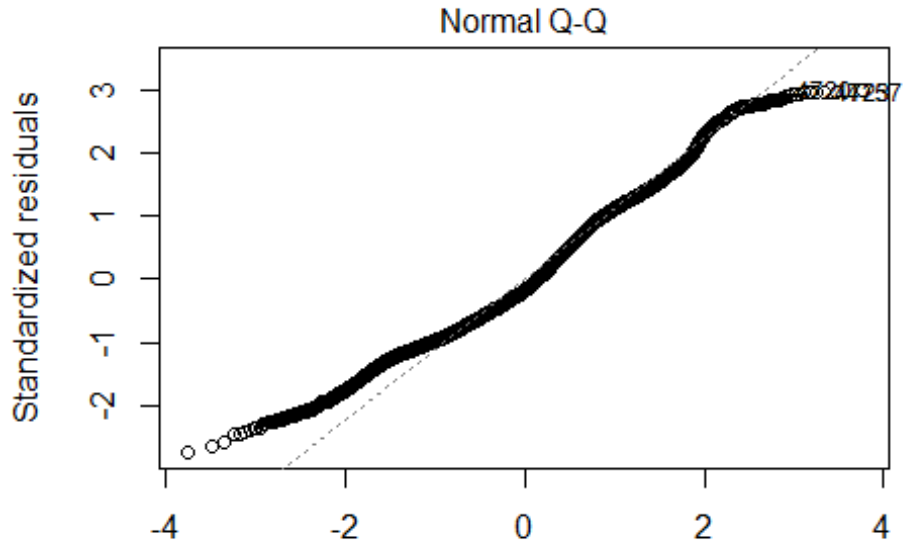
```
##                      slopes        labels
## (Intercept)    0.0009700013   (Intercept)
## test_time      0.0596532999     test_time
## Jitter...      0.0000000000     Jitter...
## Jitter.Abs.    0.0000000000   Jitter.Abs.
## Jitter.RAP     0.0000000000    Jitter.RAP
## Jitter.PPQ5    0.0000000000   Jitter.PPQ5
## Jitter.DDP     0.0000000000    Jitter.DDP
## Shimmer        0.0000000000       Shimmer
## Shimmer.dB.    0.0000000000   Shimmer.dB.
## Shimmer.APQ3   0.0000000000  Shimmer.APQ3
## Shimmer.APQ5   0.0000000000  Shimmer.APQ5
## Shimmer.APQ11  0.0105209293 Shimmer.APQ11
## Shimmer.DDA    0.0000000000   Shimmer.DDA
## NHR           -0.0952388334           NHR
## HNR           -0.1200046925           HNR
## RPDE           0.0473549429          RPDE
## DFA           -0.1917464936           DFA
## PPE            0.1308031817           PPE
```

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

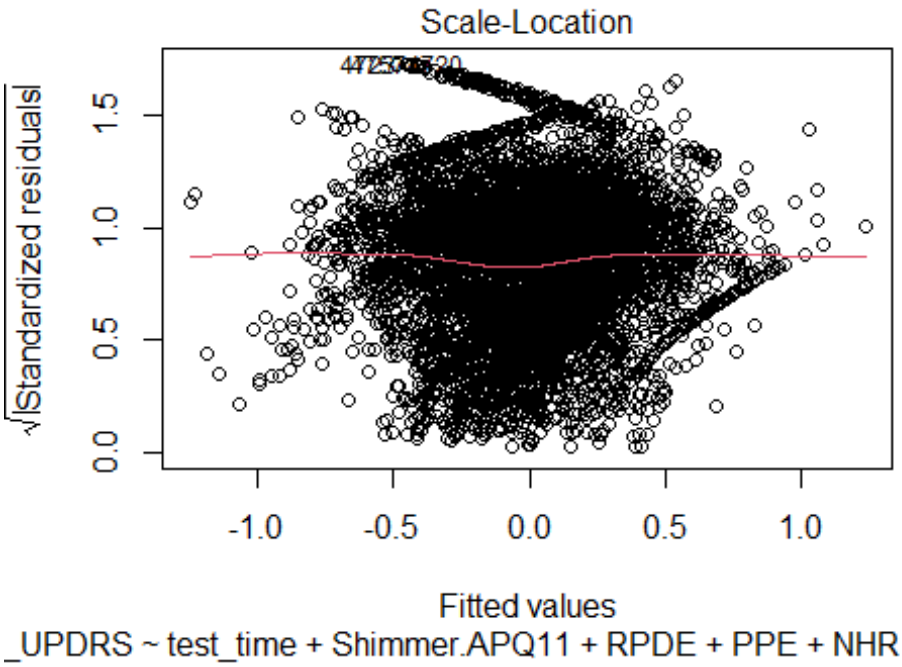
```
##
## Call:
## lm(formula = total_UPDRS ~ test_time + Shimmer.APQ11 + RPDE +
##     PPE + NHR + HNR + DFA - 1, data = tot_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5971 -0.7127 -0.1682  0.7171  2.8278
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## 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

plot(red_model)
```

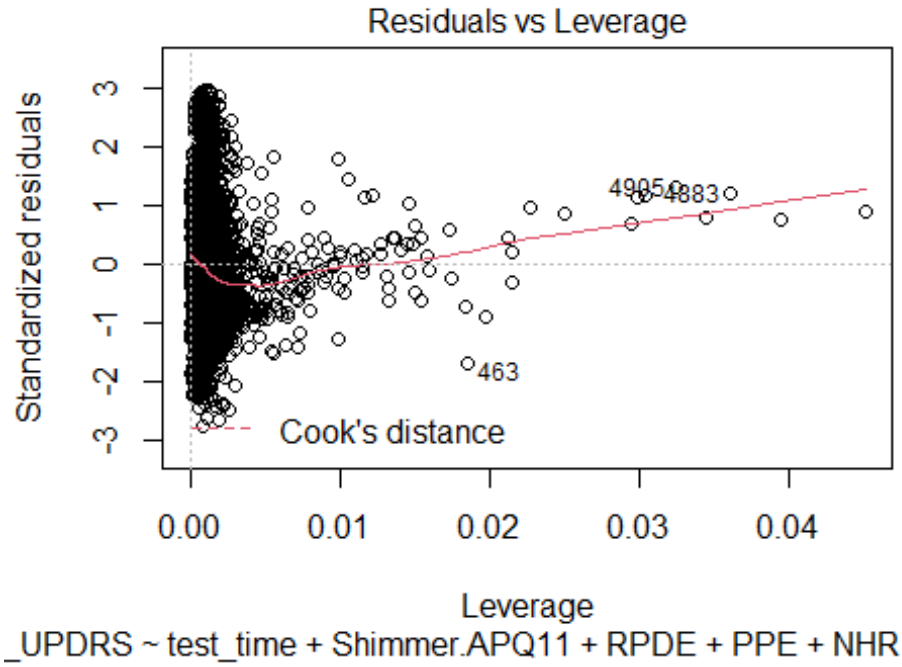


_UPDRS ~ test_time + Shimmer.APQ11 + RPDE + PPE + NHR + HN



_UPDRS ~ test_time + Shimmer.APQ11 + RPDE + PPE + NHR + HN






```
anova1 = anova(red_model, full_model)
anova1
```

```
## Analysis of Variance Table
##
## Model 1: total_UPDRS ~ test_time + Shimmer.APQ11 + RPDE + PPE + NHR +
##     HNR + DFA - 1
## Model 2: total_UPDRS ~ (test_time + Jitter... + Jitter.Abs. + Jitter.RAP +
##     Jitter.PPQ5 + Jitter.DDP + Shimmer + Shimmer.dB. + Shimmer.APQ3 +
##     Shimmer.APQ5 + Shimmer.APQ11 + Shimmer.DDA + NHR + HNR +
##     RPDE + DFA + PPE) - 1
##   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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.95, anova1$Df[2], anova1$Res.Df[1])
```

```
## [1] 1.832312
```

```
dt(2.454,5686)
```

```
## [1] 0.01966379
```

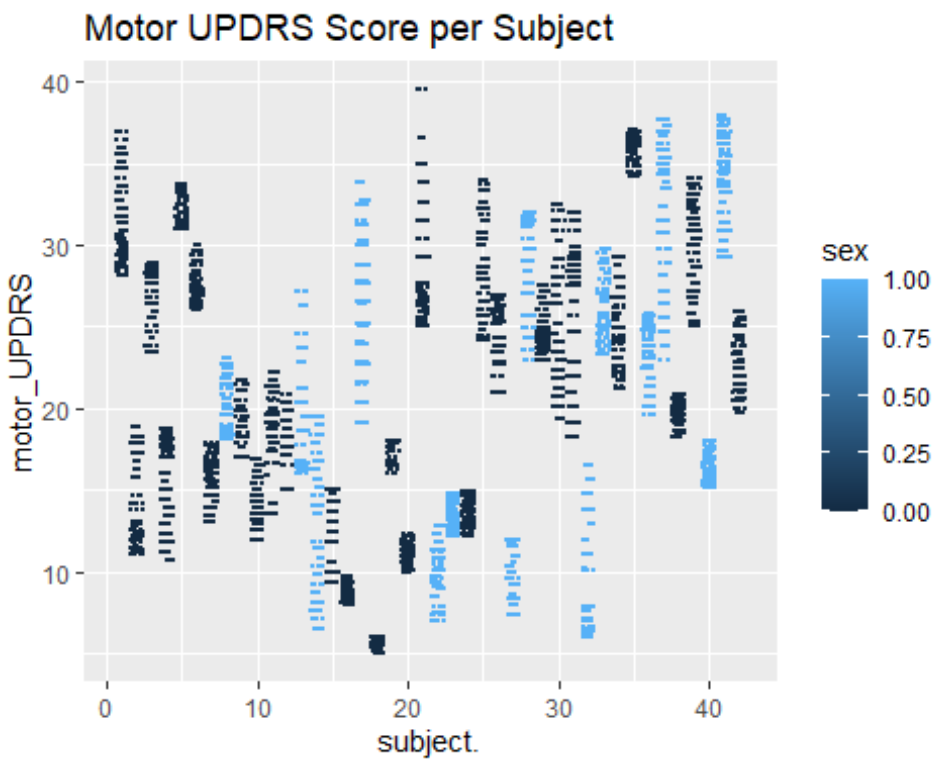
```
grouped_subject = subject_data %>%
  group_by(subject.)

head(grouped_subject)
```

```
## # A tibble: 6 x 22
## # Groups:   subject. [1]
##   subject.   age   sex test_time motor_UPDRS total_UPDRS Jitter...
Jitter.Abs.
##      <int> <int> <int>     <dbl>       <dbl>       <dbl>     <dbl>
<dbl>
## 1        1    72     0      5.64        28.2        34.4   0.00662
0.0000338
## 2        1    72     0     12.7         28.4        34.9   0.003
0.0000168
## 3        1    72     0     19.7         28.7        35.4   0.00481
0.0000246
## 4        1    72     0     25.6         28.9        35.8   0.00528
0.0000266
## 5        1    72     0     33.6         29.2        36.4   0.00335
0.0000201
## 6        1    72     0     40.7         29.4        36.9   0.00353
0.0000229
## # ... with 14 more variables: Jitter.RAP <dbl>, Jitter.PPQ5 <dbl>,
## #   Jitter.DDP <dbl>, Shimmer <dbl>, Shimmer.dB. <dbl>, Shimmer.APQ3
<dbl>,
## #   Shimmer.APQ5 <dbl>, Shimmer.APQ11 <dbl>, Shimmer.DDA <dbl>, NHR <dbl>,
## #   HNR <dbl>, RPDE <dbl>, DFA <dbl>, PPE <dbl>
```

```r
gs_lmer = lmer(formula = total_UPDRS ~ sex + (1 | subject.), data =
grouped_subject)
summary(gs_lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: total_UPDRS ~ sex + (1 | subject.)
##    Data: grouped_subject
##
## REML criterion at convergence: 28948.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2925 -0.5928  0.0396  0.5703  3.7142
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject. (Intercept) 110.064  10.491
##  Residual               7.666   2.769
## Number of obs: 5875, groups:  subject., 42
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   29.431      1.983   14.84
## sex           -2.679      3.435   -0.78
##
## Correlation of Fixed Effects:
##     (Intr)
## sex -0.577
```

```r
ggplot(grouped_subject, aes(subject., motor_UPDRS, color = sex)) +
  geom_jitter(size = 0.01)+
  ggtitle('Motor UPDRS Score per Subject')
```




```
gs_lm = lm(total_UPDRS ~ sex, data = grouped_subject)
gs_me = lmer(total_UPDRS ~ (1 | subject.), data = grouped_subject)
anova(gs_lmer, gs_me)

## refitting model(s) with ML (instead of REML)

## Data: grouped_subject
## Models:
## gs_me: total_UPDRS ~ (1 | subject.)
## gs_lmer: total_UPDRS ~ sex + (1 | subject.)
##         npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
## gs_me      3 28963 28983 -14478    28957
## gs_lmer    4 28964 28991 -14478    28956 0.634  1     0.4259

anova(gs_lmer, gs_lm)

## refitting model(s) with ML (instead of REML)

## Data: grouped_subject
## Models:
## gs_lm: total_UPDRS ~ sex
## gs_lmer: total_UPDRS ~ sex + (1 | subject.)
##         npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
## gs_lm      3 44473 44493 -22234    44467
## gs_lmer    4 28964 28991 -14478    28956 15511  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

Mayo Foundation for Medical Education and Research. (2020, December 8). *Parkinson's disease*. Mayo Clinic. Retrieved December 6, 2021, from https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055.

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE transactions on bio-medical engineering, 57(4), 884–893. https://doi.org/10.1109/TBME.2009.2036000

UCI Machine Learning Repository: Parkinsons telemonitoring data set. (n.d.). Retrieved December 6, 2021, from https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring.