Outliers and Residuals Analysis

Report Prepared by:

Kyra Nicole Melenciano Feliz

March 11th, 2022

Table of Contents

Abstract **2**

Outliers in Multiple Regression **3**

Investigating outliers 3

Dealing with outliers 3

Residuals Analysis **4**

Model 1 4

Model 2 7

Models Performance **10**

Appendix **11**

R Code Script 11

Abstract

This report centers around outliers and residuals analysis in multiple regression models, using the parkinsons data set.

The report is organized as follows. Section I deals with outlier detection in multiple regression models, steps to investigate them, approaches to dealing with them and reasons to remove them or not. Section II shows the results of two different multiple regression models and an analysis of the model's residuals. Section III concludes the report with a comparison of both models' performance using different evaluation metrics.

I. Outliers in Multiple Regression

Outliers are observations that have extreme outcomes that deviate from other observations on the data. In the case of linear regression, outliers are observations that have large residuals and their presence may affect the interpretation of the model, which is why it's important to investigate and handle outliers in order to get the right insights from the data.

Investigating outliers

In the case that you have produced a multiple regression model with reasonable performance and later discover a small number of outliers present, the next step would be to investigate the nature of the outlier to determine how to deal with it.

In regression, to investigate the outliers, one has to do a residual analysis. The standardized residuals are an indicator for outliers, measuring how far the observation is in terms of the standard error from the regression line. Observations with standardized residuals greater than 3 in absolute value are considered outliers.

Another thing to consider is the fact that the "outlier" could actually be a leverage point -an observation with extreme predictor values-. These kinds of outliers can be detected by examining the leverage statistic. Values of this statistic above 2 times the number of predictors plus 1 divided by the number of observations $[\frac{2(p+1)}{n}]$ are considered observations with high leverage.

Outliers could also be influential points, which are observations that including or excluding them can change the results of the regression analysis. These can be identified using Cook's distance, with observations that exceed 4 divided by the number of observations minus the number of predictors minus 1 $[\frac{4}{(n-p-1)}]$ considered as influential points.

All of these statistics can be identified in the residual analysis, which will be performed in the next section for two regression models with the parkinsons data set.

Dealing with outliers

The simplest approach to dealing with outliers would be to remove them. However, when fitting a model, outliers should not be removed without good reason. For instance, when an extreme value is a legitimate observation (it is not because of data entry errors), it should be left in the data set because it represents valuable information that is part of the study.

Excluding extreme values when they are legitimate observations distorts the regression results because you're forcing the subject area to appear less variable than it is in reality. A final model that is fit to the data would not be very helpful if it ignores the most exceptional cases.

II. Residuals Analysis

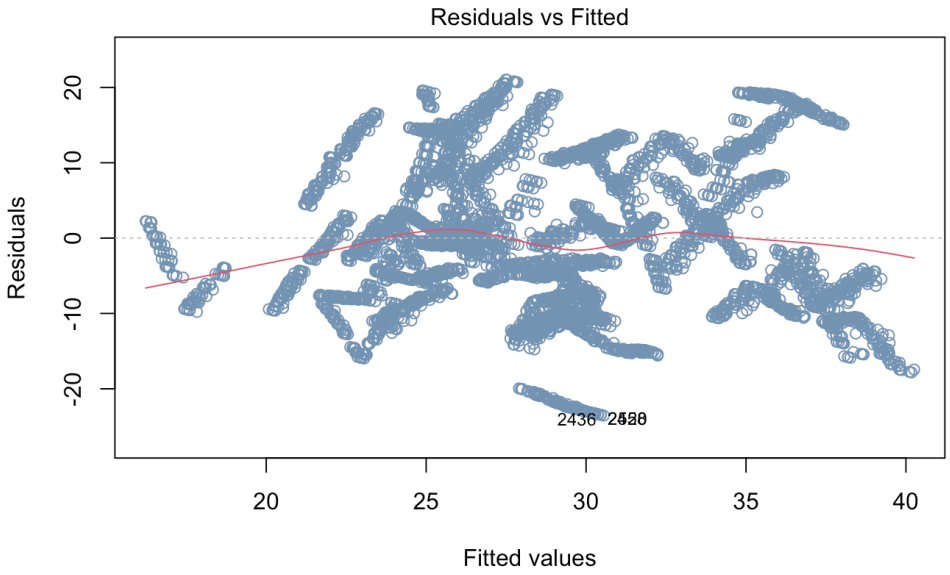
Model 1

To perform the residuals analysis, we'll be fitting a multiple regression model to predict 'total_UPDRS' with 'age', 'sex', 'test_time', 'Shimer', 'Jitter' and 'subject' as predictors. The model will have all the variables formatted as numeric.

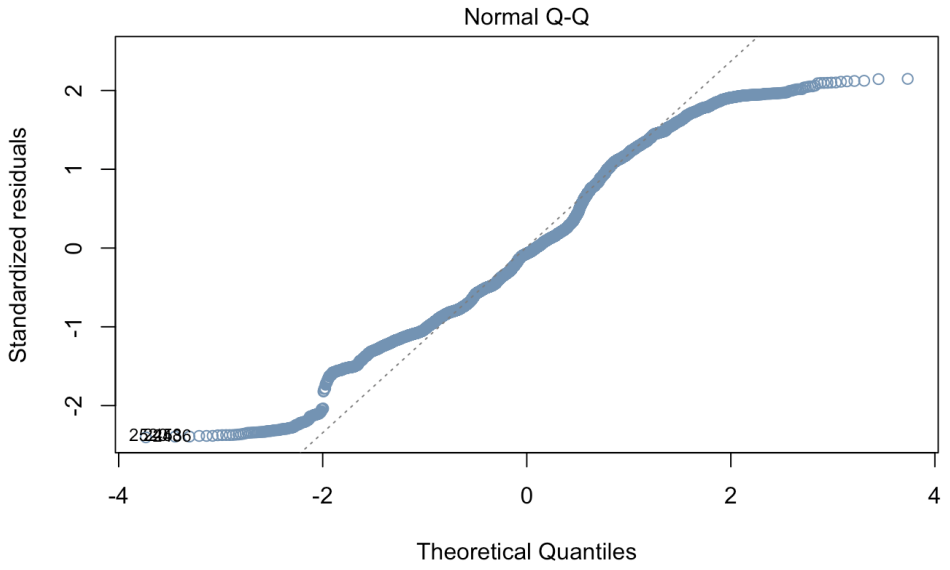
```
Call:
lm(formula = total_UPDRS ~ age + sex + test_time + Shimmer +
    Jitter + subject, data = parkinsons_numeric)

Residuals:
     Min       1Q   Median       3Q      Max
-23.5506  -7.6630  -0.6865   7.9207  21.0356

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  2.317095   1.127978   2.054                0.040 *
age          0.384318   0.015634  24.581 < 0.0000000000000002 ***
sex         -4.143292   0.307914 -13.456 < 0.0000000000000002 ***
test_time    0.013864   0.002524   5.493          0.0000000414 ***
Shimmer     16.139683  12.394215   1.302                0.193
Jitter      12.906873  76.043324   0.170                0.865
subject      0.250287   0.011463  21.834 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.804 on 5289 degrees of freedom
Multiple R-squared:  0.1969,    Adjusted R-squared:  0.196
F-statistic: 216.1 on 6 and 5289 DF,  p-value: < 0.00000000000000022
```

From these results, we can see that Shimmer and Jitter are not statistically significant for the model. Only 19.6% (adjusted R-squared) of the variability within the data is explained by the model. The model itself is statistically significant.

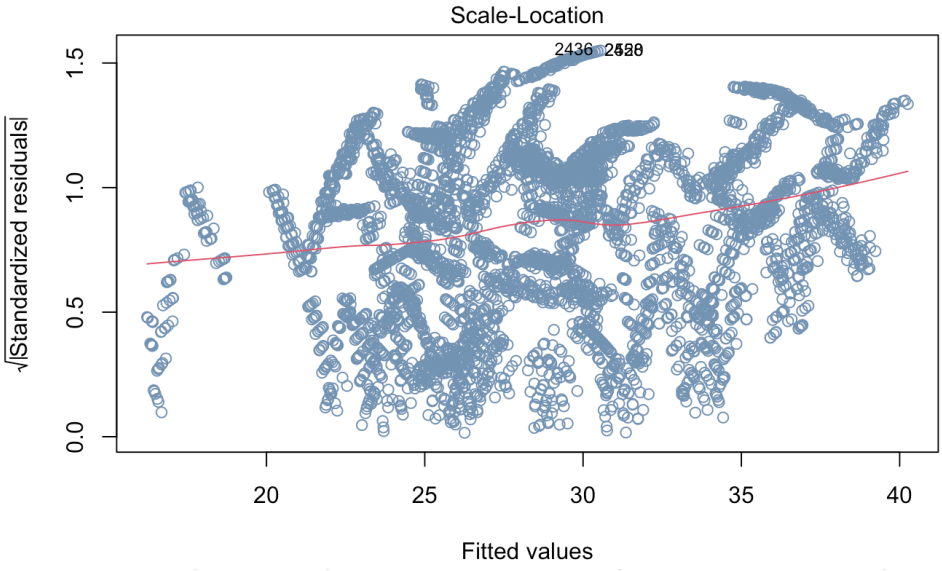
For the residual analysis, let's check the diagnostic plots from R.



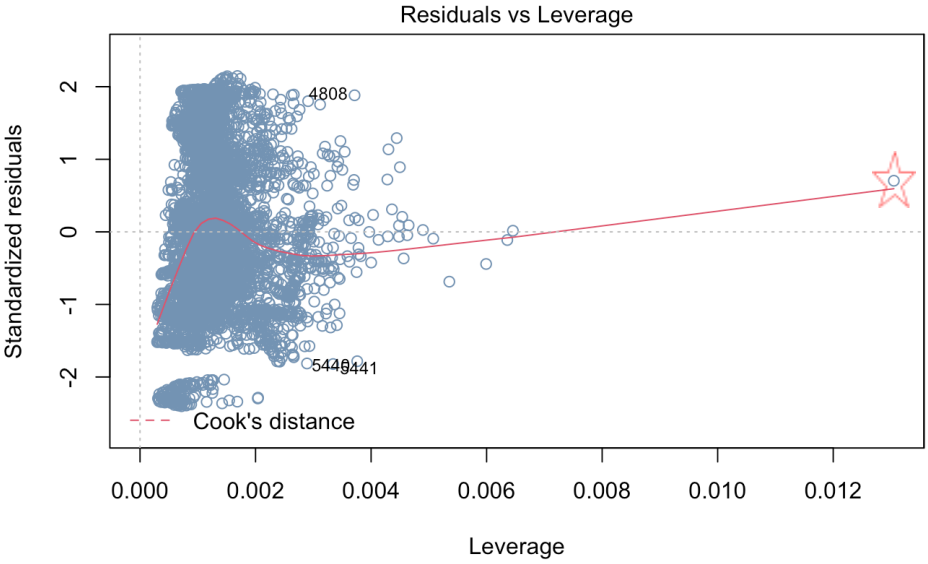
The ‘Residuals vs Fitted’ graph can be checked to confirm the linearity assumption. The plot should show no pattern and the red line should be approximately horizontal at zero. In this case, there is no pattern in the residual plot and the line is almost horizontal at zero, which hints to a linear relationship between the predictors and dependent variable.



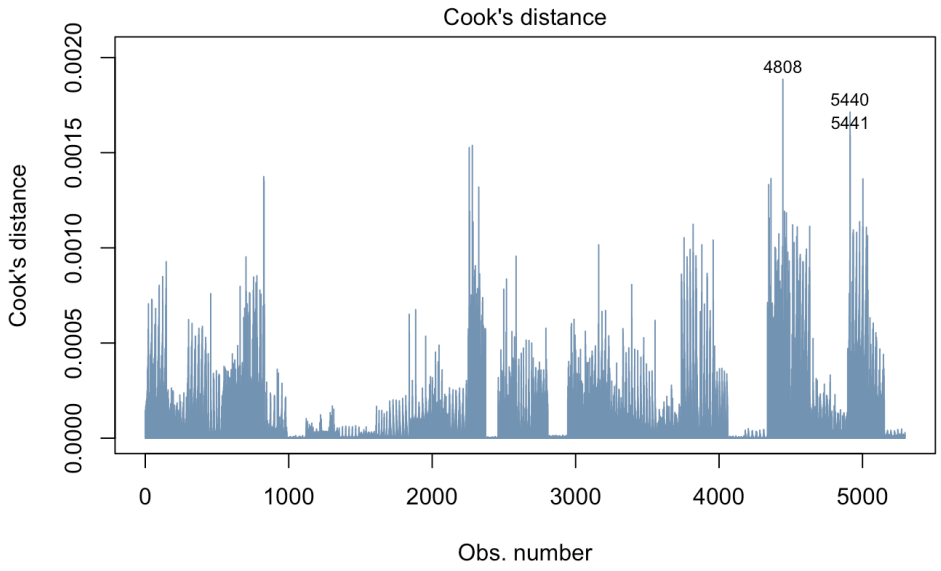
The ‘Normal Q-Q’ plot is used to check the normality assumption.To assume normality the residuals should approximately follow the straight line. In this case, the points curve towards the end hinting that our data is not normally distributed.



The 'Scale-Location' plot is used to check homogeneity of variance by showing residuals equally spread along the ranges of predictors. There should be a horizontal line with equally spread points. That's not the case for our model. The variance of the residuals increases with the predicted value (the line is not horizontal) suggesting heteroscedasticity (when the residuals errors have non-constant variance).



The 'Residuals vs Leverage' plot can be used to identify outliers and high leverage points. The plot highlights the top 3 most extreme points. However, none of the outliers exceed 3 standard deviations. The observation with highest leverage is highlighted inside the star.



The 'Cook's distance' plot shows the 3 most extreme values labeled. However, from the plot shown before we can see that all of the residuals lie within Cook's distance, as no observation is shown below the dashed red line.

Model 2

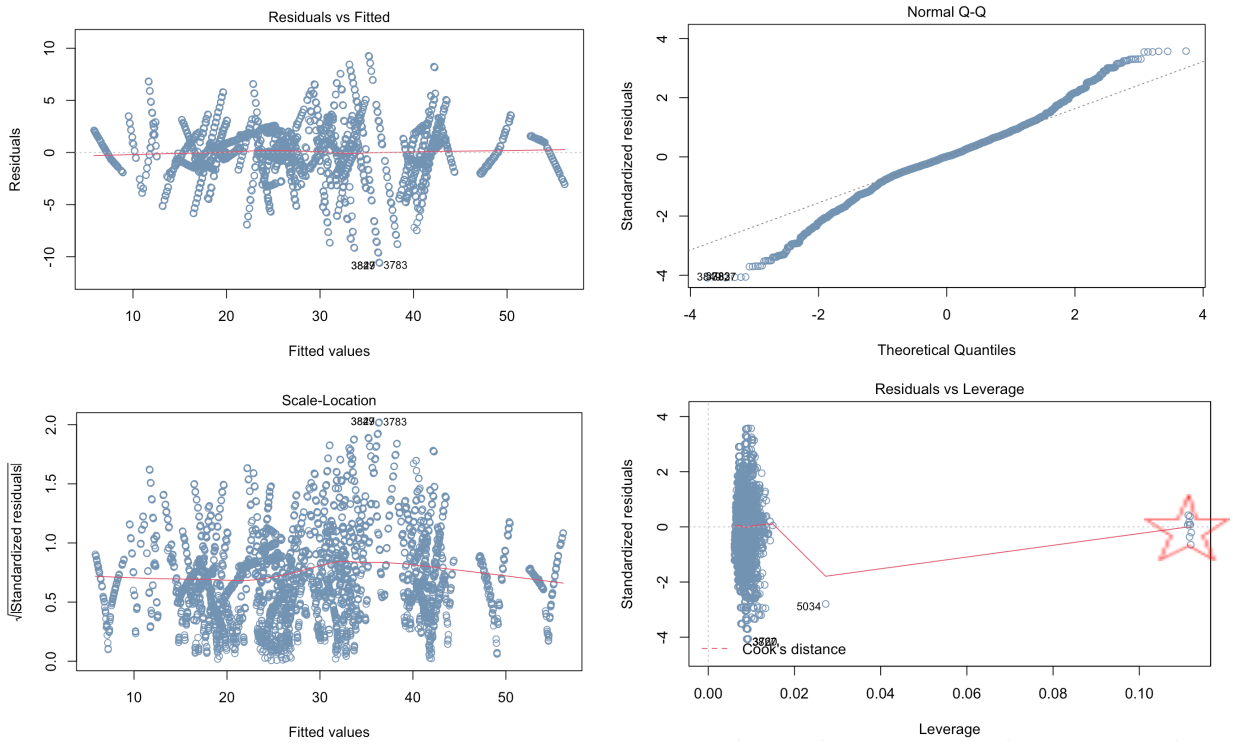
For this model, the same predictors as before will be used to predict 'total_UPDRS'. However, in this case sex and subject will be formatted as factor.

```
Call:
lm(formula = total_UPDRS ~ age + sex + test_time + Shimmer +
    Jitter + subject, data = parkinsons_factor)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6080  -1.2972   0.0261   1.4930   9.2809

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -8.823e+00  1.865e+00   -4.731 2.29e-06 ***
age          6.661e-01  2.782e-02   23.942  < 2e-16 ***
sex1         4.447e+00  2.878e-01   15.450  < 2e-16 ***
test_time    1.886e-02  6.811e-04   27.690  < 2e-16 ***
Shimmer     -1.412e+00  4.205e+00   -0.336   0.7370
Jitter      -1.597e+01  2.490e+01   -0.642   0.5212
subject2    -1.505e+01  3.674e-01  -40.958  < 2e-16 ***
subject3     2.635e+00  3.776e-01    6.979 3.34e-12 ***
subject4    -1.835e+01  3.675e-01  -49.941  < 2e-16 ***
subject5    -7.617e-01  3.556e-01   -2.142   0.0323 *
subject6     6.562e+00  2.889e-01   22.710  < 2e-16 ***
subject7    -1.759e+01  3.026e-01  -58.142  < 2e-16 ***
subject8    -2.005e+01  3.603e-01  -55.652  < 2e-16 ***
```

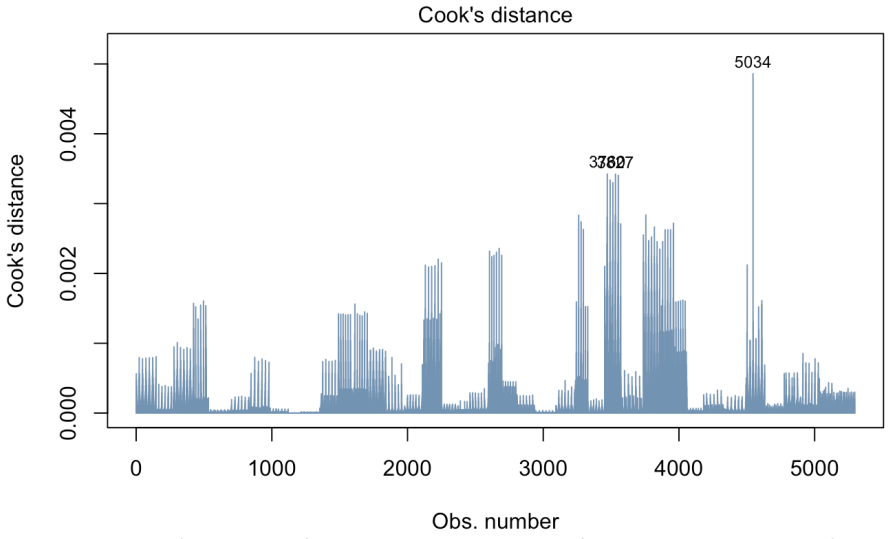
```
subject9    -1.312e+01  3.230e-01  -40.610  < 2e-16 ***
subject10   -1.171e+01  3.600e-01  -32.520  < 2e-16 ***
subject11   -6.589e+00  4.201e-01  -15.685  < 2e-16 ***
subject12   -9.789e+00  3.220e-01  -30.404  < 2e-16 ***
subject13   -1.799e+01  3.745e-01  -48.038  < 2e-16 ***
subject14   -1.786e+01  4.383e-01  -40.745  < 2e-16 ***
subject15   -1.676e+01  2.748e-01  -61.004  < 2e-16 ***
subject16   -1.741e+01  2.827e-01  -61.589  < 2e-16 ***
subject17   -9.277e+00  3.338e-01  -27.791  < 2e-16 ***
subject18   -2.859e+01  2.854e-01 -100.170  < 2e-16 ***
subject19   -3.328e+00  4.593e-01   -7.245 4.95e-13 ***
subject20   -2.049e+01  2.791e-01  -73.421  < 2e-16 ***
subject21   -9.573e-01  3.573e-01   -2.679   0.0074 **
subject22   -2.425e+01  4.696e-01  -51.650  < 2e-16 ***
subject23   -1.114e+01  4.127e-01  -26.996  < 2e-16 ***
subject24   -1.460e+01  3.287e-01  -44.410  < 2e-16 ***
subject25    5.463e+00  3.773e-01   14.479  < 2e-16 ***
subject26    5.914e+00  5.808e-01   10.182  < 2e-16 ***
subject27   -1.987e+01  4.542e-01  -43.753  < 2e-16 ***
subject28   -1.171e+01  3.736e-01  -31.349  < 2e-16 ***
subject29   -1.360e+01  4.113e-01  -33.070  < 2e-16 ***
subject30    1.141e+01  5.649e-01   20.201  < 2e-16 ***
subject31   -1.373e+01  3.962e-01  -34.649  < 2e-16 ***
subject32   -8.068e+00  9.606e-01   -8.399  < 2e-16 ***
subject33   -1.072e+01  3.461e-01  -30.973  < 2e-16 ***
subject34    3.449e-01  3.351e-01    1.029   0.3033
subject35    1.404e+01  2.953e-01   47.526  < 2e-16 ***
subject36   -8.121e+00  9.363e-01   -8.673  < 2e-16 ***
subject37    7.386e+00  4.725e-01   15.631  < 2e-16 ***
subject38   -1.065e+01  2.688e-01  -39.618  < 2e-16 ***
subject39    3.439e+00  2.828e-01   12.160  < 2e-16 ***
subject40   -2.795e+01  5.762e-01  -48.506  < 2e-16 ***
subject41           NA         NA       NA       NA
subject42           NA         NA       NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.609 on 5251 degrees of freedom
Multiple R-squared:  0.9435,    Adjusted R-squared:  0.9431
F-statistic:  1995 on 44 and 5251 DF,  p-value: < 2.2e-16
```

In this case, with subject and sex formatted as factors, the programming behind creates different dummy variables for each of the categories, which is why we see 41 coefficients for each subject (except subject 1 that is takes as the base for comparison). The same happens with sex, only showing the coefficient for sex 1 (male). Here, as with the other model's results, Shimmer and Jitter are not statistically significant. The adjusted R-squared is much larger, with the model explaining 94.31% of the data's variability. The model itself is also statistically significant.

Again, let's check the diagnostic plots.



The 'Residuals vs Fitted' plot shows a horizontal red line at zero and no pattern for the residuals, which confirms the linearity assumption. The 'Normal Q-Q' plot is curved at the tails which confirms that the data is not normally distributed. The 'Scale-Location' plot shows a curved line, which means that the assumption of homoscedasticity is not met and the residuals errors have non-constant variance. The 'Residuals vs Leverage' shows the highest leverage points within the star.



Cook's distance shows the 3 most extreme values labeled. However, from the plot Scale-Location plot we can see that all of the residuals lie within Cook's distance, as no observation is shown below the dashed red line.

III. Models Performance

Now let's compare both models' performance with different evaluation metrics.

```
                  Model     RMSE      MAE        R2  R2adj
1 With numeric format 9.797840 8.066617 0.1968818 0.1960
2  With factor format 2.597678 1.917363 0.9435467 0.9431
```

From these we can see that Model 2, with sex and subject formatted as factor has better performance as the R-squared and adjusted R-squared are greater than in model 1, the variability in the data is better explained by model 2. The Mean Absolute Error (MAE) is lower in model 2, which means that the average distance between observed and predicted values is smaller for this model. The RMSE is also lower for model 2, which confirms that this is the better model from the two shown.

To compare the fits of the two models, an F-test can be done with the Analysis of Variance table.

```
Analysis of Variance Table

Model 1: total_UPDRS ~ age + sex + test_time + Shimmer + Jitter + subject
Model 2: total_UPDRS ~ age + sex + test_time + Shimmer + Jitter + subject
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   5289 508404
2   5251  35737 38    472667 1827.7 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference of the two models lies in the variables sex and subject being formatted as factor in model 2. Hence, this ANOVA will test whether or not this formatting leads to a significant improvement over using the numeric format. The results show 38 degrees of freedom, indicating that model 2 has 38 more parameters and a p-value much smaller than 0.01, meaning that formatting sex and subject as factor did lead to a significant improvement in fit over model 1.

Regardless of the fact that model 2 appears to be a good model with good predictive power, I don't think either of these models should be used because they include subject as a predictor and this variable is a unique identifier, which would make predictions for new individuals difficult since they would not have the same ID as any subject included in the model.

IV. Appendix

R Code Script

```
setwd("/Users/kyramelenciano/dana-4810/A4")
library(modelr)
library("tidyverse")
library("ggpubr")
library("summarytools")
library("cowplot")
library("quantable")
library("robustHD")
library("dlookr")
library("caTools")
library("dplyr")
library("caret")
library(forecast)

parkinsons_numeric <- parkinsons3

# formatting subject and sex as numeric

parkinsons_numeric$subject <- as.numeric(parkinsons_numeric$subject)
parkinsons_numeric$sex <- as.numeric(parkinsons_numeric$sex)


model <- lm(total_UPDRS ~ age + sex + test_time +
        Shimmer + Jitter + subject, data = parkinsons_numeric)
summary(model)

# Diagnostic plots

plot(model, col = "#7393B3")
plot(model, 4, col = "#7393B3")


# with subject and sex as factor

parkinsons_factor <- parkinsons3

model2 <- lm(total_UPDRS ~ age + sex + test_time +
        Shimmer + Jitter + subject, data = parkinsons_factor)
summary(model2)
```

```
# Diagnostic plots

plot(model2, col = "#7393B3")
plot(model2, 4, col = "#7393B3")

# Predictions

pred_num <- model %>% predict(parkinsons_numeric)
pred_fact <- model2 %>% predict(parkinsons_factor)

# Evaluation metrics

data.frame(Model = c("With numeric format", "With factor format"),
        RMSE = c(RMSE(pred_num, parkinsons_numeric$total_UPDRS),
              RMSE(pred_fact, parkinsons_factor$total_UPDRS)),
        MAE = c(MAE(pred_num, parkinsons_numeric$total_UPDRS),
              MAE(pred_fact, parkinsons_factor$total_UPDRS)),
        R2 = c(R2(pred_num, parkinsons_numeric$total_UPDRS),
              R2(pred_fact, parkinsons_factor$total_UPDRS)),
        R2adj = c(0.196, 0.9431)
)

# F-test
anova(model,model2)
```