# Table of Contents

[Table of Contents 2](#_Toc86677312)

[Descriptive Analysis – Parkinson 3](#_Toc86677313)

[Descriptive Analysis 3](#_Toc86677314)

[Identify data quality and apply Data cleansing 4](#_Toc86677315)

[Multivariate Regression 6](#_Toc86677316)

[Exclusion 6](#_Toc86677317)

[Variable Selection Method I – Forward Selection 7](#_Toc86677318)

[Forward Selection – Stopping rule with AIC (Akaike Information Criterion) 7](#_Toc86677319)

[Variable Selection Method II – Backward Selection 8](#_Toc86677320)

[Backward Selection – Stopping rule (1) with AIC 8](#_Toc86677321)

[Advantage and disadvantages 10](#_Toc86677322)

[Why less variables is better when implement model with large number of predictors 10](#_Toc86677323)

[Multicollinearity and regression model 10](#_Toc86677324)

[Appendix 12](#_Toc86677325)

[Stepwise – Forward full result table 12](#_Toc86677326)

[Stepwise – Forward full result table 13](#_Toc86677327)

[Stepwise full result table 14](#_Toc86677328)

[Appendix – R Code 15](#_Toc86677329)

# Descriptive Analysis – Parkinson

Parkinson's disease is a kind of nervous system disorder. It starts from a barely noticeable tremor but gradually becomes stiffness or slowing of movement. Tremors are common, but the disorder also commonly causes stiffness or slowing of movement.

This report analyses the given data set from Parkinson's Disease and provides different views after applying standardization.

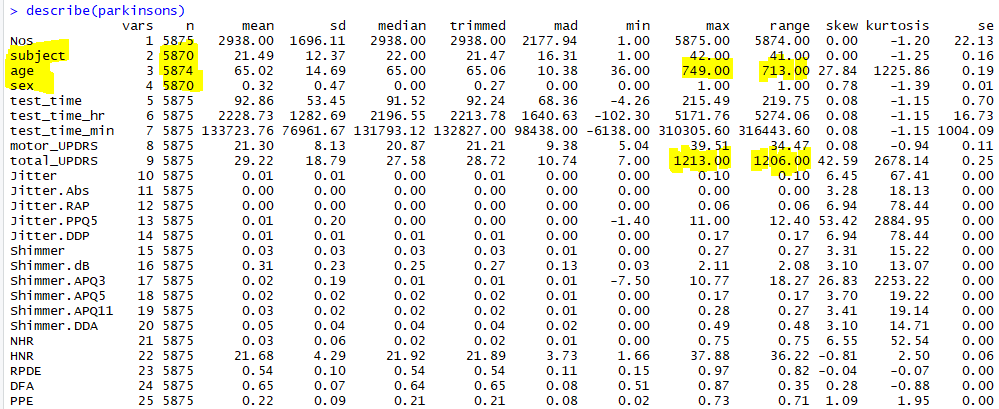
## Descriptive Analysis

The given dataset has 5,875 observations, a collection of biomedical voice data taken over a period from 42 individuals with Parkinson's disease. There are 27 columns in the data frame as below:

* NOS – Numeric variable. the row number of an observation
* Subject ID – Categorical variable. ID is given on test subject
* Age – Categorical variable. Age of the test subject
* Sex – Categorical variable. Gender of the subject, zero is male, and one is female
* Test time, Test time hour and Test time min – Numeric variable. Test time measurement
* Motor\_UPDRS – Numeric variable. Motor UPDRS score of the subject
* Total\_UPDRS – Numeric variable. Total UPDRS score of the subject
* Jitter (%, Abs, RAP, PPQ5, DDP) – Numeric variable. a measure of frequency in a variety of test
* Shimmer (dB, APQ3, APQ5, APQ11, DDA) – Numeric variable. a measure of amplitude in a variety of test
* NHR, HNR – Numeric variable. Measurement of noise to tonal components invoice
* RPDE – Numeric variable. Dynamical complexity measurement
* DFA – Numeric variable. the signal fractal scaling exponent
* PPE – Numeric variable. Fundamental frequency variation measurement.

## Identify data quality and apply Data cleansing

Let's look at the data quality; the overall quality is good, only 13 observations have missing values. In other words, 99.98% observations with value in all reading.

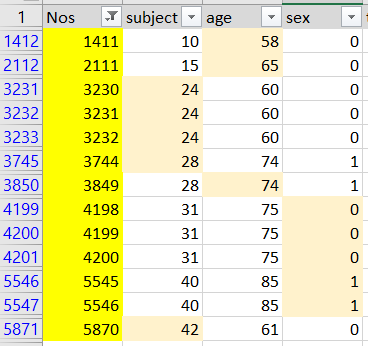
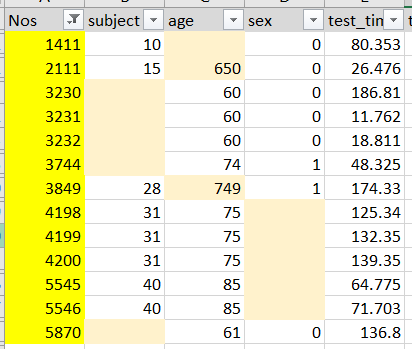


For those missing information, we realize there are:

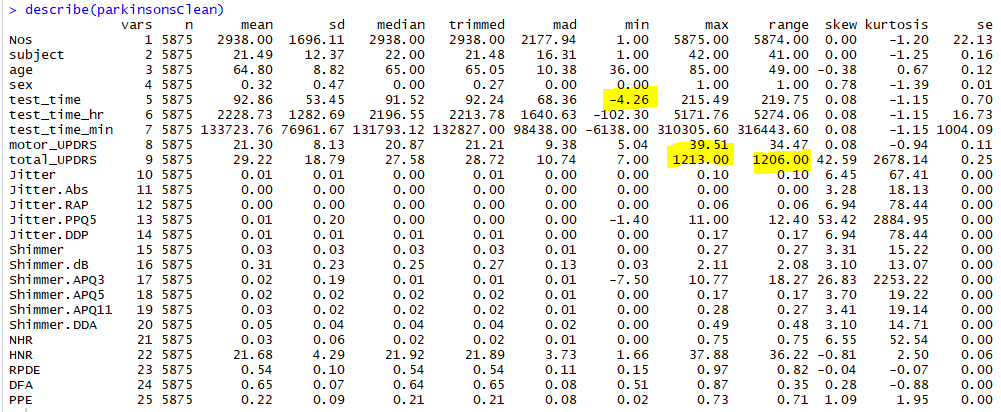
* Five observations have no value in Subject
* One observation has no value in Age, and two invalid values (650, 749) are found and believe it belongs to a typo
* Five observations have no value in Sex

**Resolution** Since this data was collected from the subject over many tests. We can clean up the data with the pattern; for example, in observation #1411, we can look at the range from #1400 to #1420 and identify whether those observations belong to the same subject. If true, we can repair the missing value. By applying this method to all fields, we've got:

Original Data Clean Data

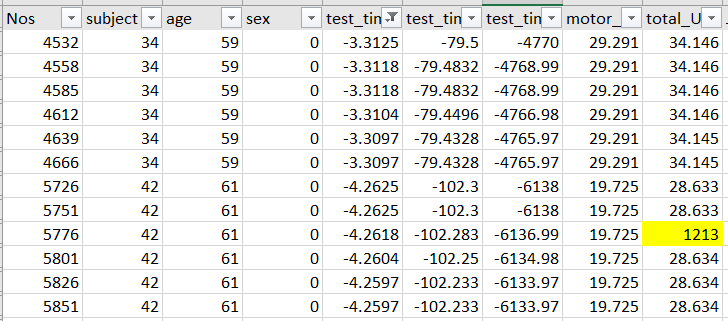


After data cleansing and looking into the dataset's summary again - the total number of variables (Subject, Age and Sex) were aligned to 5,875. However, taking another look at the test\_time does not make any sense with a negative timestamp. Also, there is an outliner (1213.00) on total\_UPDRS, with the mean value at 29.22 and median at 28.72.

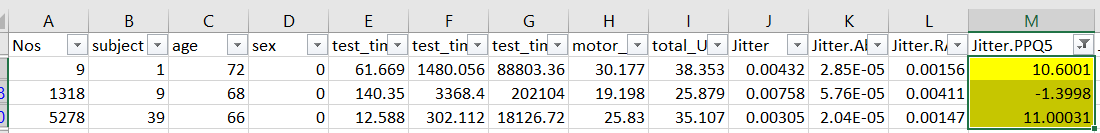
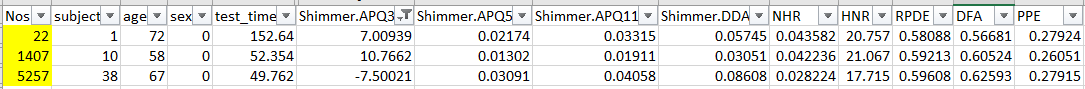


There are 12 observations found with negative test\_time, which includes the outliner of total\_UPDRS. Since 12 observations are only a tiny percent of the population, removing these records will not significantly impact the analysis.

**Resolution**: remove these 12 records in the dataset.



There are three outliners in Shimmer APQ3 and three outliners in Jitter PPQ5, six records removed for the next section experiment. After data cleansing, the total observations are 5,857.



# Multivariate Regression

Multivariate Regression is a method that measures the degree at which more than one predictors and respond variables are linearly related. This method used wisely to predict the behaviour of the response variables by changing the predictor variables.

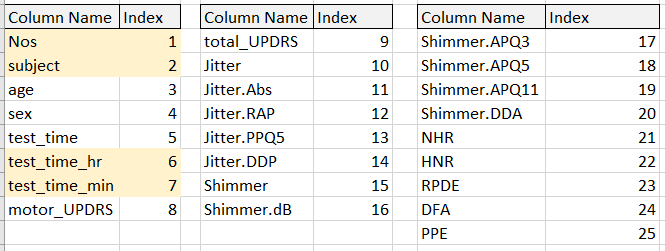
In this report, we would like to explorer the model that can fit for response variable (Total\_UPDRS) from Parkinson Disease dataset. There method introduced below.

## Exclusion

A subset of data created by excluding “Nos”, “subject”,”Test\_time\_hr” and “time\_time\_min” with these reasons:

* Nos – Do not have any relationship with Total\_UPDRS.
* Subject ID – Do not have any relationship with Total\_UPDRS.
* Test time hour and Test time min – these two columns are sub category of Test\_time.

There are totally 25 candidates in the Parkinson Disease dataset, as pevious mentions, column index 1,2,6, and 7 will be excluded.



pks – cleaned Parkinson’s Disease data

pks\_filter – filtered cleaned Parkinson’s Disease data

## Variable Selection Method I – Forward Selection

The step of forward selection are

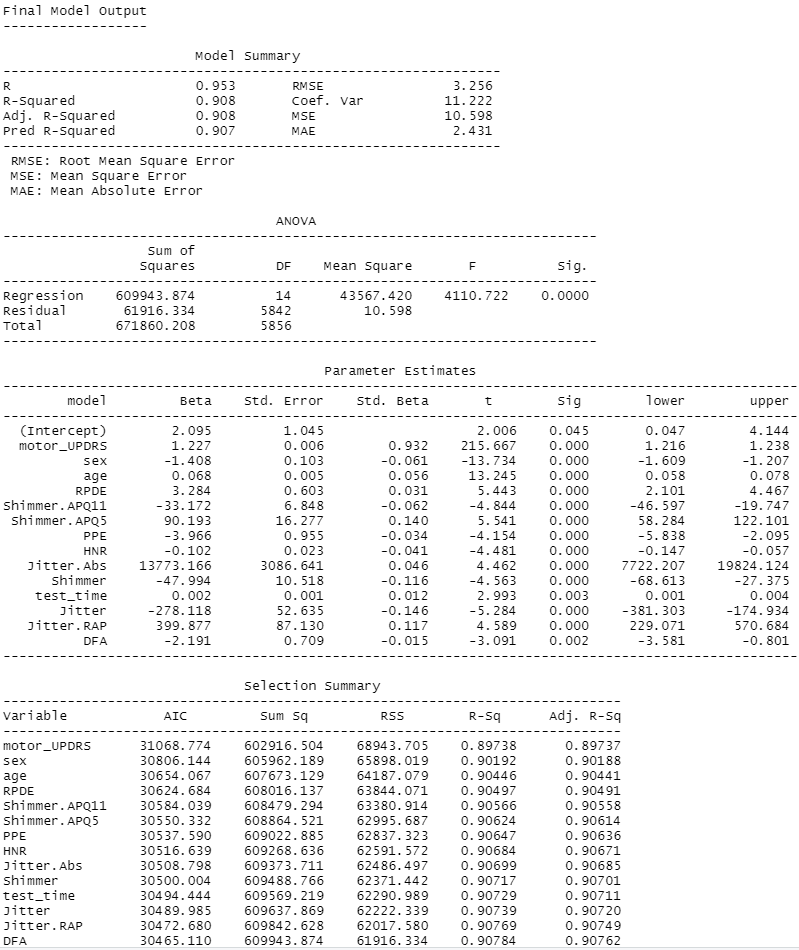
1. Start with an empty model / null model, do not have any predictors.
2. Adding variable one by one bases on the most significant value either on
   1. Smallest p-value, or
   2. The highest increase in R square, or
   3. The highest drop in Residuals Sum of Square (RSS).
3. Loop until all variables are examinated in the model or it stopped on pre-defined rule.
   1. P-Value, or
   2. AIC (Akaike information criterion), or
   3. BIC (Bayesian information criterion), or
   4. A fixed value, for example, 0.30

## Forward Selection – Stopping rule with AIC (Akaike Information Criterion)

AIC defines the threshold by the number of degree of freedom, an example for a binary variable includes in a model it need to have p-value < 0.157. In other word, degree of freedom increase will have lower threshold.

**The Model equation is**

Total\_UPDRS = 2.095 + 1.227(motor\_UPDRS) - 1.408(sex) + 0.068(age) + 3.284(RPDE) -33.172(Shimmer.APQ11) + 90.193(Shimmer.APQ5) – 3.966(PPE) – 0.102(HNR) + 13773.166(Jitter.Abs) – 47.994(Shimmer) + 0.002(test\_time) - 278.118(Jitter) + 399.877 (Jitter.RAP) – 2.191(DFA)



## Variable Selection Method II – Backward Selection

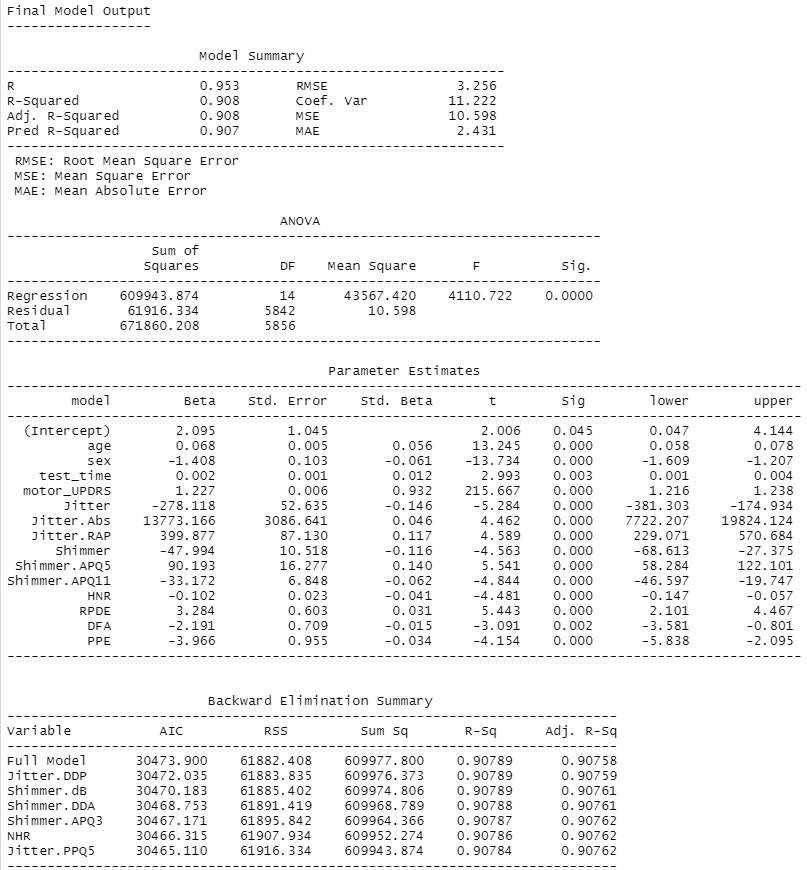
The step of forward selection are

1. Start from a full model iall variables are under consideration.
2. Start to eliminate variable one by one bases on the least significant value either on
   1. Highest p-value, or
   2. The lowest decreasae in R square, or
   3. The lowest increase in Residuals Sum of Square (RSS).
3. Loop until all variables are elminiated in the model or it stopped on pre-defined rule.

## Backward Selection – Stopping rule (1) with AIC

**The Model equation is**

Total\_UPDRS = 2.095 + 0.068(age) - 1.408(sex) + 0.002(test\_time) + 1.227(motor\_UPDRS) - 278.118(Jitter) + 13773.166(Jitter.Abs) + 399.877(Jitter.RAP) – 47.994(Shimmer) + 90.193 (Shimmer.APQ5) - 33.172(Shimmer.APQ11) – 0.102(HNR) + 3.284(RPDE) – 2.191(DFA) – 3.966(PPE)

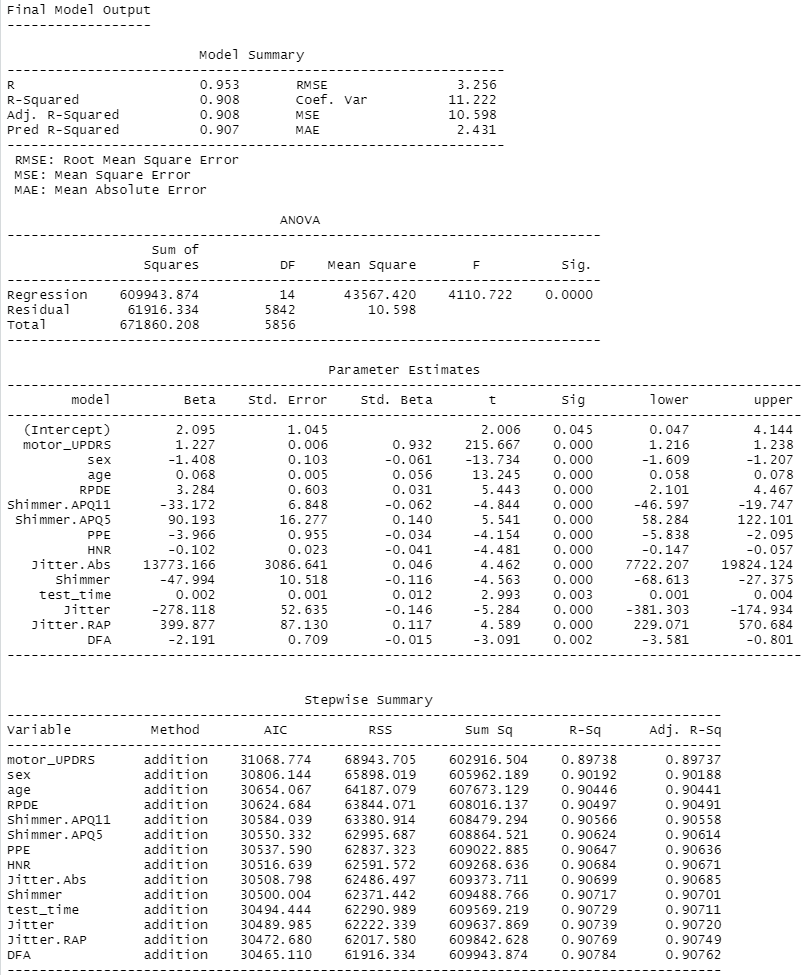


Backward selection will also provide a summary with which variable is eliminated at the end of the result.**Variable Selection Method III – Stepwise Selection**

This is a combination on both test at each step on add or remove variable.

**The Model equation is**

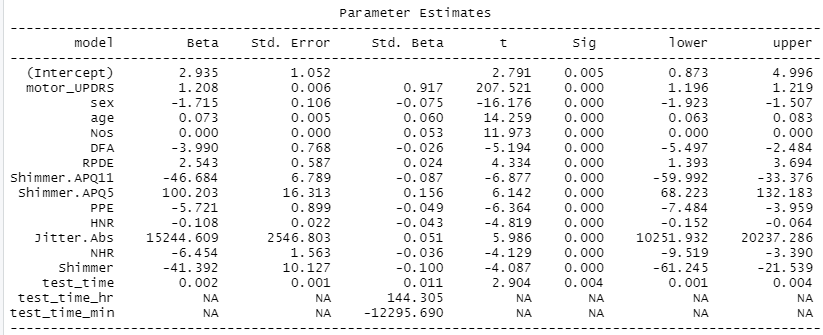
Total\_UPDRS = 2.095 + 0.068 (age) - 1.408 (sex) + 0.002 (test\_time) + 1.227 (motor\_UPDRS) - 278.118 (Jitter) + 13773.166 (Jitter.Abs) + 399.877 (Jitter.RAP) – 47.994 (Shimmer) + 90.193 (Shimmer.APQ5) - 33.172(Shimmer.APQ11) – 0.102 (HNR) + 3.284 (RPDE) – 2.191 (DFA) – 3.966 (PPE)



Findings (filter out test\_time\_hr and test\_time\_min)

Three model equation returning the same when using AIC selection rule as Total\_UPDRS = 2.095 + 0.068(age) - 1.408(sex) + 0.002(test\_time) + 1.227(motor\_UPDRS) - 278.118(Jitter) + 13773.166(Jitter.Abs) + 399.877(Jitter.RAP) – 47.994(Shimmer) + 90.193(Shimmer.APQ5) - 33.172(Shimmer.APQ11) – 0.102(HNR) + 3.284(RPDE) – 2.191(DFA) – 3.966(PPE)

Findings will be differently with model if we include other non meaningful variables, such as “Nos”, “subject”, “test\_time\_hr” , and “test\_time\_min”. The intercepts value has sightly difference and NA on test\_time\_hr and test\_time\_min.



## Advantage and disadvantages

The advantage of stepwise selection is a 1) easy to execute and low cost, and 2) give a direction on which predictors can fit on the response variable.

The disadvantage are 1) computing method which not consider predictors in staticial way (an example of “Nos” variable, 2) instability reducted when having more than 50 variables which required X of events of selection algorithm, and 3) may not consider all possible combination of potential predictors as a variable is added or removed, will not re-consider them when next selection step.

## Why less variables is better when implement model with large number of predictors

In regression analysis, model can provide an explanation on variable relationship however when there is lots of variables which create misleading on statistical values such as R-square values, co-efficient and p-value.

On the other hand, the mission of regression model is to predict the random sample, too many predictors reduce the model’s applicability. In last, the more predictors in the model produce a higher successful rate, however it caused productivity issue (complex the calculation and take time on execution). By 80:20 rule, focus on 20% of most significant predicters will help to create data model with reasonable time plus successful rate.

## Multicollinearity and regression model

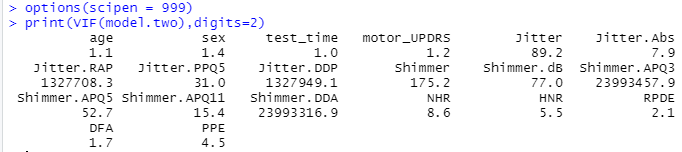
Multiconllinearity is the occurrence of high intercorrelations between predictors in a multiple regression model. It can exist when two predictors are highly correlated, for example, Total\_UPDRS and motor\_UPDRS. There are ways to remove multiconllinearlity however which may also remove an important predictors from the regression model. In statisitic, analyst study the variables relationship before removing.

There are number of ways to detect multicollinearity in the regression model

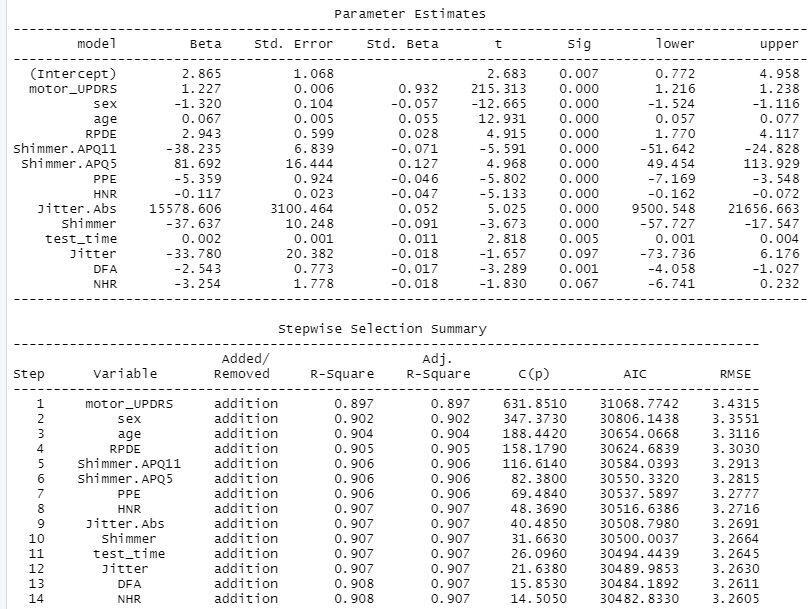
1. Coeffients on different sampling size – split the data into two sets and run the model separately, if you spot the coefficients are very different in two model that may indice a multicollinearity.
2. High Variance Inflaction Factor (VIF) and Low tolerance – both statistical test are indicate on multicollinearity. While VIF is a measurement on variance of the coefficients.
3. Very high standard error for regression coefficients – standard errors higher than the coefficients.

## Identify multicollinearity on Parkinson dataset

Using the VIF function to calculate the variance inflaction factor on each variables, base on point two mentioned above, varibles Jitter.DDP, Jitter.RAP, Shimmer.APQ3 and Shimmer.DDA with very high VIF values which should be ignored from the stepwise selection.

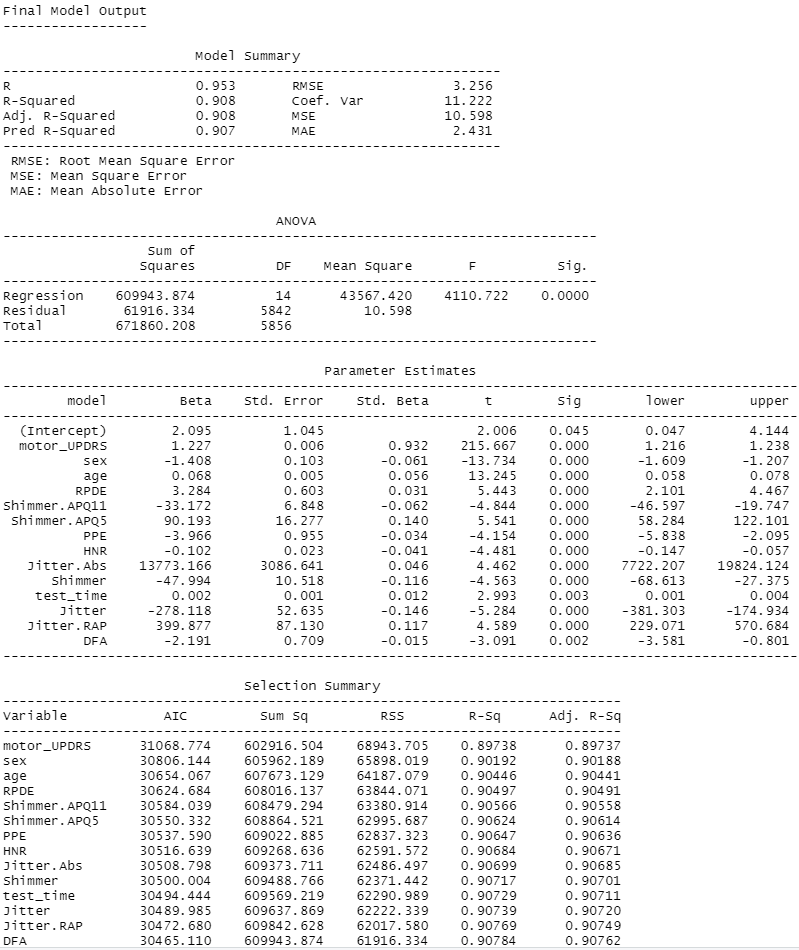


This is the stepwise result after remove those four variables from the candidate list.

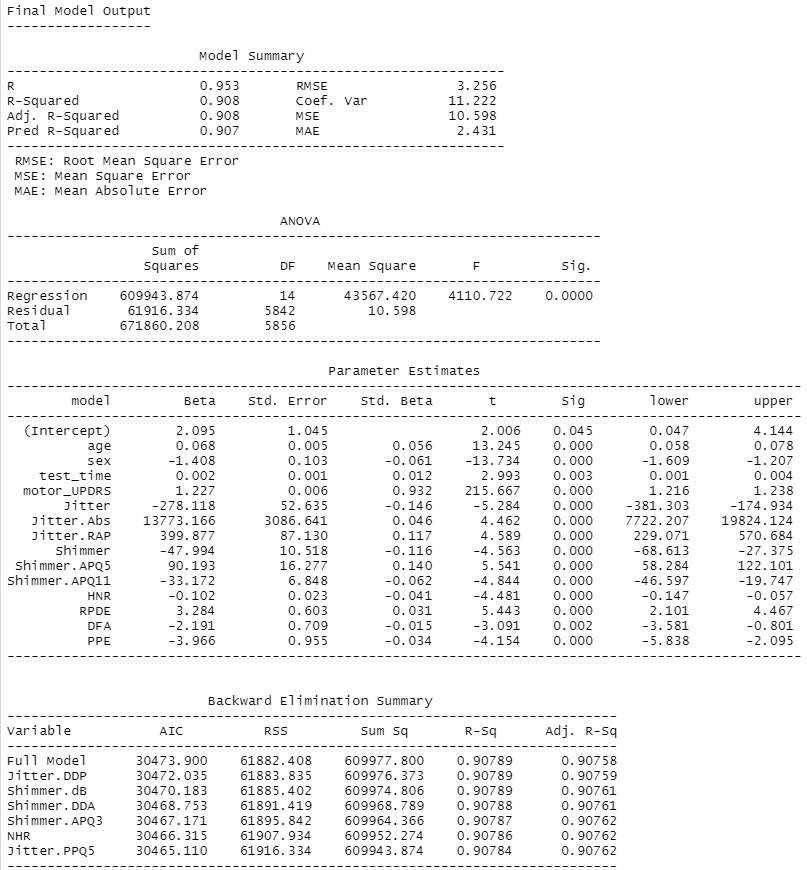


# Appendix

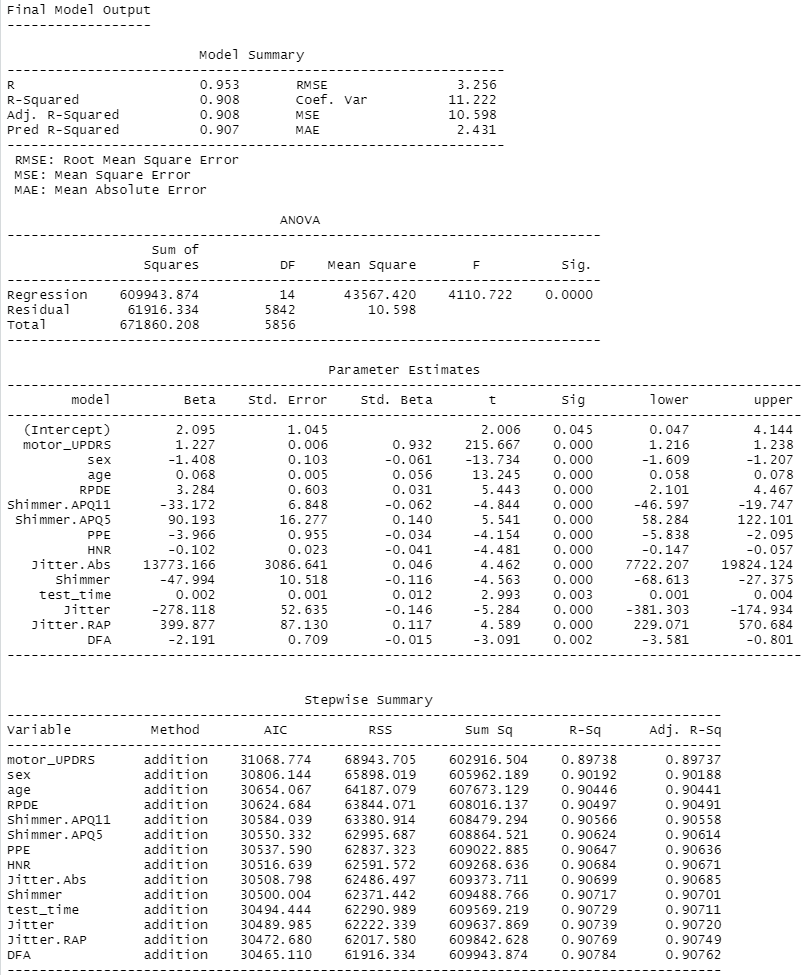
## Stepwise – Forward full result table



## Stepwise – Forward full result table



## Stepwise full result table



# Appendix – R Code

rm(list=ls())

library(tidyverse)

# Read in the Lung Cap Data

#read.table(file.choose(), header=T)

pks <- read.table(file.choose(" "), header=T, sep=",")

#Filter out some data field is not fit for the model, Nos, Test\_time\_hr, test)time\_min

pks\_filter <- pks[ -c(1:2,6:7) ]

#install.packages("PerformanceAnalytics")

#library(PerformanceAnalytics)

#chart.Correlation(pks\_filter, histogram=TRUE, pch=19)

# Attach pks\_filter

attach(pks\_filter)

head(pks\_filter)

# stepwise forward regression

library(olsrr)

model.fwd <- lm(total\_UPDRS ~ ., data = pks\_filter)

summary(model.fwd)

ols\_step\_forward\_p(model.fwd, details = FALSE)

#Stepwise AIC Forward Regression

#Build regression model from a set of candidate predictor variables by entering predictors based on Akaike Information Criteria

ols\_step\_forward\_aic(model.fwd, details = TRUE)

# stepwise forward regression

library(olsrr)

model.bkw <- lm(total\_UPDRS ~ ., data = pks\_filter)

summary(model.bkw)

ols\_step\_backward\_p(model.bkw, details = TRUE)

#Build regression model from a set of candidate predictor variables by entering predictors based on Akaike Information Criteria,

ols\_step\_backward\_aic(model.bkw, details = TRUE)

#stepwise regression

#Build regression model from a set of candidate predictor variables by entering and removing predictors based on p values, in a stepwise manner until there is no variable left to enter or remove any more. The model should include all the candidate predictor variables. If details is set to TRUE, each step is displayed.

model.stpw <- lm(total\_UPDRS ~ ., data = pks\_filter)

ols\_step\_both\_p(model.stpw, details = TRUE)

#Stepwise AIC Regression

#Build regression model from a set of candidate predictor variables by entering and removing predictors based on Akaike Information Criteria, in a stepwise manner until there is no variable left to enter or remove any more. The model should include all the candidate predictor variables. If details is set to TRUE, each step is displayed.

ols\_step\_both\_aic(model.stpw, details = TRUE)

model.stpwf <- lm(total\_UPDRS ~ ., data = pks)

ols\_step\_both\_aic(model.stpwf, details = TRUE)

ols\_step\_backward\_aic(model.stpwf, details = TRUE)

#Select the subset of predictors that do the best at meeting some well-defined objective criterion, such as having the largest R2 value or the smallest MSE, Mallowâs Cp or AIC.

model.all <- lm(total\_UPDRS ~ ., data = pks\_filter)

ols\_step\_best\_subset(model.all, details = TRUE)

#install.packages("regclass")

library(regclass)

model.one <- lm(total\_UPDRS ~ ., data = pks)

VIF(model.one)

model.two <- lm(total\_UPDRS ~ ., data = pks\_filter)

VIF(model.two)

options(scipen = 999)

print(VIF(model.two),digits=2)

head(pks\_filter)

pks\_VIF <- pks\_filter[ -c(8,10,13,16) ]

head(pks\_VIF)

model.three <- lm(total\_UPDRS ~ ., data = pks\_VIF)

VIF(model.three)

options(scipen = 999)

print(VIF(model.three),digits=2)

ols\_step\_both\_p(model.three, details = TRUE)