# Table of Contents

[Table of Contents 2](#_Toc88008225)

[Outliners 3](#_Toc88008226)

[Type of outliners 3](#_Toc88008227)

[How to identify them 3](#_Toc88008228)

[How to deal with them 4](#_Toc88008229)

[Descriptive Analysis – Parkinson 4](#_Toc88008230)

[Clean Data 4](#_Toc88008231)

[Multivariate Regression 5](#_Toc88008232)

[Numeric Data Type Check 5](#_Toc88008233)

[Multivariate Regression Model 8](#_Toc88008234)

[Multivariate Regression Model – factor variables 9](#_Toc88008235)

[Standard Error and MAE 11](#_Toc88008236)

[F-Test 11](#_Toc88008237)

[Conclusion 11](#_Toc88008238)

[Appendix – R Code 12](#_Toc88008239)

# Outliners

In regression model, outliners can distort the prediction and impact the accuracy if they are not handling probably. In general, remove or change the extreme value is not the best option, we should understand the impact and adjust those extreme value to ensure the model is fit for the estimates and prediction.

## Type of outliners

Outliners has two type which can be

1. Univariate – single extreme point of observation
2. Multivariate – a set (two or more data) of unusual data points

## How to identify them

There are number of ways to identify the outliners.

1. Visualization – populate data into graph
   1. Scatter plots – can easily detect outliner when plotting data points onto the graph.
   2. Box plots – like scatter plots can clearly display the IQR, min and max.
   3. Histogram – another graphical presentation, outliner will be display bin with abnormal value.
2. Interquartile Range technique – medium, first quartile and third quartile are good indicator to give an idea on range of data estimate. For data below first quartile or above third quartile and the value is not in a reasonable range, then may need more effort for outliner checking.
3. Statistical test – there are different type of statistical test
   1. Chi-square test – use chi-square distribution with p-value to support conclusion on whether accept or reject the null hypothesis.
   2. Grubb’s test – assume data is normal distributed.
   3. Dixon’s Q test – rare used in data science as normally apply for small dataset size.
   4. The null hypothesis for item 3a – 3c is:
      1. H0; There is no outliner in the dataset
      2. Ha: There is/are outliner(s) in the dataset

## How to deal with them

We should not just remove the outliners because (1) They can be a legitimate observation that may provide a different direction when drafting the model and (2) removing outliners will affect assumption and results (compare analysis on with and without outliner).

# 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.

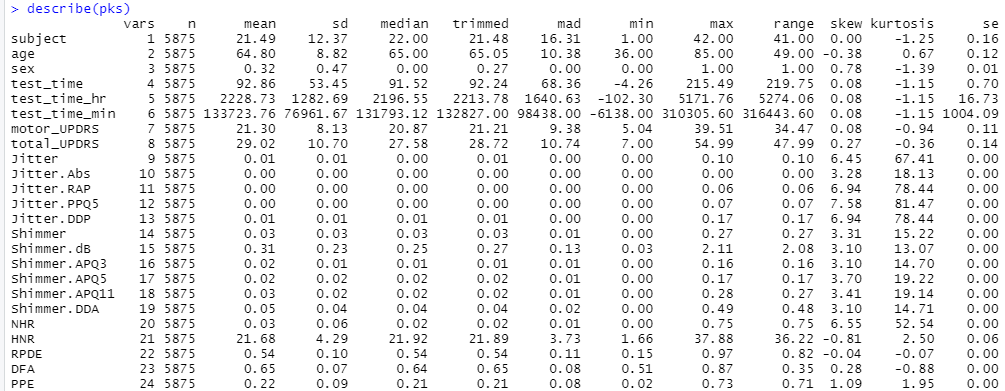
This report analyses the given data set from Parkinson's Disease and discuss the topic of outliner, such as, whether we should remove or keep, why and what are the consequences.

## Clean Data

The given dataset has 5,875 observations, a collection of biomedical voice data taken over a period from 42 individuals with Parkinson's disease.



Looking into the dataset, there is no missing records. Thus, we are move on to outliner detection session.



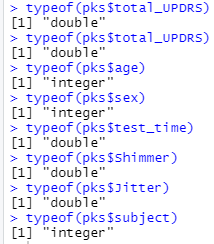
# Multivariate Regression

Multivariate Regression is a method that measures the degree at which more than one predictor and respond variables are linearly related. This method used wisely to predict the behavior 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) with predictors (age, sex, test\_time, shimmer, jitter and subject) from Parkinson Disease dataset.

## Numeric Data Type Check

Check data type of the predictors, all predictors are in numeric type.



Let’s look of these predictors with Quantile-Quantile plots (qq-plot) which aims to find out where the two sets of data (50:50) come from the same distribution. For data set with common distribution, the data point will lie on the reference line.

For predictor - sex, subject and test\_time, the data sets are under common distribution.

For predictor – age, it is a bimodal distribution.

|  |  |
| --- | --- |
| Predictor - Age | Predictor - Sex |
|  |  |
| Predictor – test\_time | Predictor - subject |
|  |  |

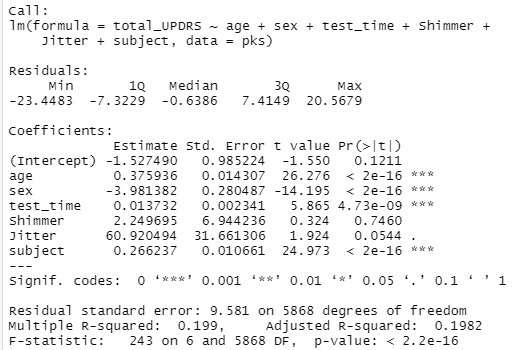
For predictor – Shimmer and Jitter, they are right-skewed

|  |  |
| --- | --- |
| Predictor - Shimmer | Predictor - Sex |
|  |  |

## Multivariate Regression Model

The linear equation on total UPDRS below means one unit of total UPDRS increase required the factor of these predictors -1.5275 intercepts. For Shimmer / Jitter, the p-value is 0.746 and 0.0544, which are greater than the 0.05 confident level. Thus, we accept the null hypothesis, and there is no relationship between Shimmer / Jitter and Total\_UPDRS. Therefore, we remove those predictors from the linear regress model.

Total UPDRS = 0.3759(Age) - 3.9813 (Sex) + 0.0137 (Test\_Time) - 0.2662 (subject) -1.5275

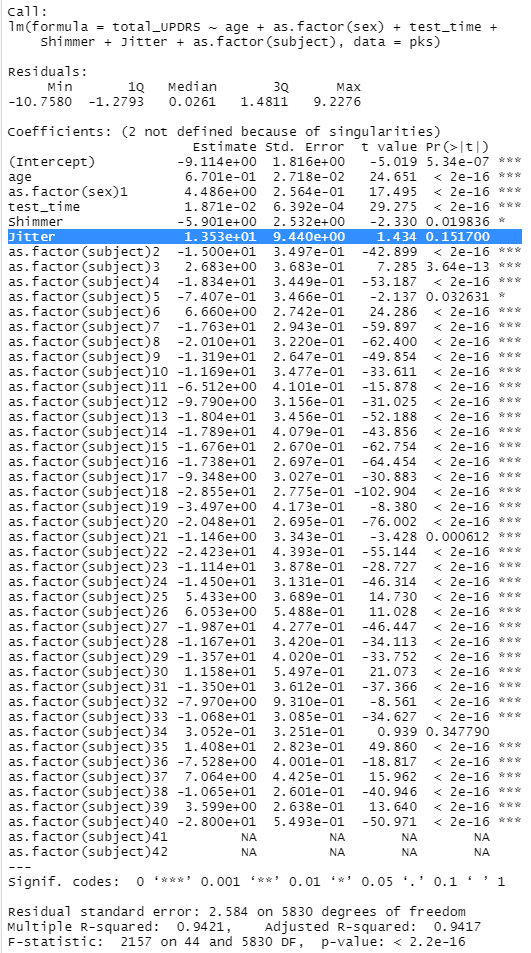


The standard error tells the average distance between the estimated coefficient from the response variable. It also uses it to calculate confidence levels to support the relationship between predictor variables and response variables. P-value helped with these. In this model, most of the p-value are less than 0.05 (with \*\*\* significant codes) indicates we can reject the null hypothesis that there is a relationship between response variables and predictors.

The residual standard error explains the quality of linear regression. The average deviation from the regression line is 9.581, which is quite significant as it was calculated with 5,868 degrees of freedom (data point in consideration). R-square interprets how fit the data model is, with 0.199 (R-square) and 0.1982 (adjusted R-square), it tells around 19.9% of response variable can be explained by predictor variable. In general, the R-square value close to 1 is better.

## Multivariate Regression Model – factor variables

Repeat the same model with ‘sex’ and ‘subject’ defined as factor variables in the multivariate regression model. Look at the model below ‘sex’ and ‘subject’ predictors are broken into categories. Predictor Jitter, subject#34, subject#41, and subject#45 are removed from the linear regress on either p-value is greater than significant level 0.05 or NA.



This time the residual standard error explains the quality of linear regression. The average deviation from the regression line is 2.584, which is quite significant as it was calculated with 5,830 degrees of freedom (data point in consideration). R-square interprets how fit the data model is, with 0.9421 (R-square) and 0.9417 (adjusted R-square), it tells around 94.1% of response variable can be explained by predictor variable. In general, the R-square value close to 1 is better.

The graph on the left is multivariate regression model without factor variable, the other one is with factor variable. You can see the ‘with factor variable’ graph has many observations lie on the line and more interaction on estimates and predictions. Hence this model is better comparing with the previous one.

|  |  |
| --- | --- |
| Without Factor variable | With Factor Variable |
|  |  |
|  |  |
|  |  |
|  |  |

## 

## Standard Error and MAE

Mean Average Error (MAE) is a measurement of error between pairs of the data point. In other words, it calculates the absolute value between forecast and correspondent. For the training dataset, the residual standard error is 9.581, and the MAE is 7.784. The residual standard error is calculated based on prediction, and MAE is calculated as the absolute, which gives a more precise (1.8 difference) on error measurement of the regression model. Like the original dataset, the MAE for model with factor variable is 0.696 (2.584 – 1.888). These two regression models have identified differences in absolute error measurement.

On the other hand, the MAE value from the original dataset is 8.144, and the MAE value for the model with factor variable is 1.888; the model with factor variable has a smaller MAE value, which means the model is better in theory.

## F-Test

The null hypothesis on F-test is there is no independent variable fits the data as well as the predictive model. Refers to the test result, p-value less than significant level 0.05 means both models are fit better than the intercept-only model. Next looking into the F-value, original dataset is 13.753 and the model with factor variable is 0.072; F-value indicate the joint effect with all predictors, smaller variable means all the results are significant while bigger number mean something is significant.

## Conclusion

With all the comparison and statistical support, the multivariate regression model with factor vairbale are better fit for this experiment.

# 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=",")

# Attach pks\_filter

attach(pks)

head(pks)

install.packages('psych')

library('psych')

describe(pks)

typeof(pks$total\_UPDRS)

typeof(pks$age)

typeof(pks$sex)

typeof(pks$test\_time)

typeof(pks$Shimmer)

typeof(pks$Jitter)

typeof(pks$subject)

#-------------- Check QQ Plot and line --------------------

#install.packages('qqplotr')

library('qqplotr')

#qqnorm(): produces a normal QQ plot of the variable

#qqline(): adds a reference line

qqnorm(pks$age, pch = 1, frame = FALSE)

qqline(pks$age, col = "steelblue", lwd = 2)

qqnorm(pks$sex, pch = 1, frame = FALSE)

qqline(pks$sex, col = "steelblue", lwd = 2)

qqnorm(pks$test\_time, pch = 1, frame = FALSE)

qqline(pks$test\_time, col = "steelblue", lwd = 2)

qqnorm(pks$Shimmer, pch = 1, frame = FALSE)

qqline(pks$Shimmer, col = "steelblue", lwd = 2)

qqnorm(pks$Jitter, pch = 1, frame = FALSE)

qqline(pks$Jitter, col = "steelblue", lwd = 2)

qqnorm(pks$subject, pch = 1, frame = FALSE)

qqline(pks$subject, col = "steelblue", lwd = 2)

#------- Define multivariate Regression

library(car)

reg <- lm(total\_UPDRS ~ age+sex+test\_time+Shimmer+Jitter+subject, data=pks)

summary(reg)

#Outliner test

outlierTest(reg)

qqPlot(reg,labels=row.names(pks), id.method="identify",

simulate=TRUE, main="Q-Q Plot")

reg\_factor <- lm(total\_UPDRS ~ age+as.factor(sex)+test\_time+Shimmer+Jitter+as.factor(subject), data=pks)

summary(reg\_factor)

#Outliner test

outlierTest(reg\_factor)

qqPlot(reg\_factor,labels=row.names(pks), id.method="identify",

simulate=TRUE, main="Q-Q Plot")

######### MAE #################

install.packages("Metrics")

library("Metrics")

mae(pks$total\_UPDRS, predict(reg))

mae(pks$total\_UPDRS, predict(reg\_factor))

install.packages("Metrics")

library("Metrics")

######### RMSE#################

rmse(pks$total\_UPDRS, predict(reg))

rmse(pks$total\_UPDRS, predict(reg\_factor))

# Compute the analysis of variance

aov(reg\_factor, data = pks)

# Summary of the analysis

summary(res.aov)

library("STAT")

var.test(reg,reg\_factor, alternative = "two.sided")

var.test(reg\_factor,reg, alternative = "two.sided")