# Table of Contents

[Table of Contents 2](#_Toc84249134)

[Descriptive Analysis – Parkinson 3](#_Toc84249135)

[PDTrain summary 3](#_Toc84249136)

[PDTest summary 4](#_Toc84249137)

[Correlation Map for PDTrain Dataset 5](#_Toc84249138)

[Correlation Map for PDTest Dataset 6](#_Toc84249139)

[Linear Regression on Training Set 7](#_Toc84249140)

[Linear Regression Summary – PD Test dataset 9](#_Toc84249141)

[Standard Error and MAE 10](#_Toc84249142)

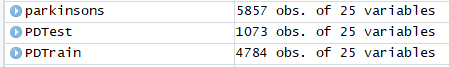
[Conclusion 11](#_Toc84249143)

[Reference 11](#_Toc84249144)

[Appendix – R Code 12](#_Toc84249145)

# Descriptive Analysis – Parkinson

This report continues the experient from assignment 1, cleaned data with 5857 observations. We split the data into 80:20 ratios and got two subsets as PDTrain and PDTest below:



## PDTrain summary

PDTrain dataset is a subset of the Parkinsons dataset, which is applied with the 80/20 rule. 80% subset of data is a clone to the PDTrain dataset. There are 4,784 observations with the summary below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| Nos | 2 | 1460 | 2928 | 2931 | 4397 | 5875 |
| Subject | 1 | 10 | 21.5 | 21.44 | 33 | 42 |
| Age | 36 | 58 | 65 | 64.82 | 72 | 85 |
| Sex | 0 | 0 | 0 | 0.3165 | 1 | 1 |
| test\_time | 0.3958 | 47.3380 | 92.4430 | 93.1543 | 138.4600 | 215.4900 |
| Test\_time\_hr | 9.5 | 1136.1 | 2218.6 | 2235.7 | 3323 | 5171.8 |
| Test\_time\_min | 570 | 68167 | 133118 | 134142 | 199382 | 310306 |
| Motor\_UPDRS | 5.038 | 14.993 | 20.839 | 31.283 | 27.594 | 39.511 |
| Total\_UPDRS | 7 | 21.36 | 27.49 | 29.02 | 36.4 | 54.99 |
| Jitter | 0.00083 | 0.00358 | 0.00491 | 0.006194 | 0.006830 | 0.09999 |
| Jitter.Abs | 2.250e-06 | 2.230e-05 | 3.480e-05 | 2.250e-06 | 5.390e-05 | 4.456e-04 |
| Jitter.RAP | 0.000330 | 0.001580 | 0.002250 | 0.003008 | 0.003310 | 0.057540 |
| Jitter.PPQ5 | 0.000430 | 0.001830 | 0.002495 | 0.003305 | 0.003500 | 0.069560 |
| Jitter.DDP | 0.000980 | 0.004730 | 0.006760 | 0.009024 | 0.009940 | 0.172630 |
| Shimmer | 0.00344 | 0.01910 | 0.2774 | 0.03414 | 0.03981 | 0.26863 |
| Shimmer.dB | 0.0300 | 0.1750 | 0.2540 | 0.3118 | 0.3640 | 2.1070 |
| Shimmer.APQ3 | 0.00172 | 0.00931 | 0.01378 | 0.01722 | 0.02065 | 0.14428 |
| Shimmer.APQ5 | 0.00205 | 0.01082 | 0.01611 | 0.02024 | 0.02385 | 0.16702 |
| Shimmer.APQ11 | 0.00299 | 0.01567 | 0.02282 | 0.02758 | 0.03283 | 0.27546 |
| Shimmer.DDA | 0.00515 | 0.02793 | 0.04134 | 0.05165 | 0.06194 | 0.43283 |
| NHR | 0.000286 | 0.010895 | 0.32324 | 0.032324 | 0.031441 | 0.748260 |
| HNR | 1.659 | 19.366 | 21.905 | 21.666 | 24.468 | 37.875 |
| RPDE | 0.1510 | 0.2696 | 0.5427 | 0.5416 | 0.6145 | 0.9661 |
| DFA | 0.5140 | 0.5967 | 0.6447 | 0.6537 | 0.7122 | 0.8656 |
| PPE | 0.02198 | 0.15684 | 0.20648 | 0.22009 | 0.26474 | 0.73173 |

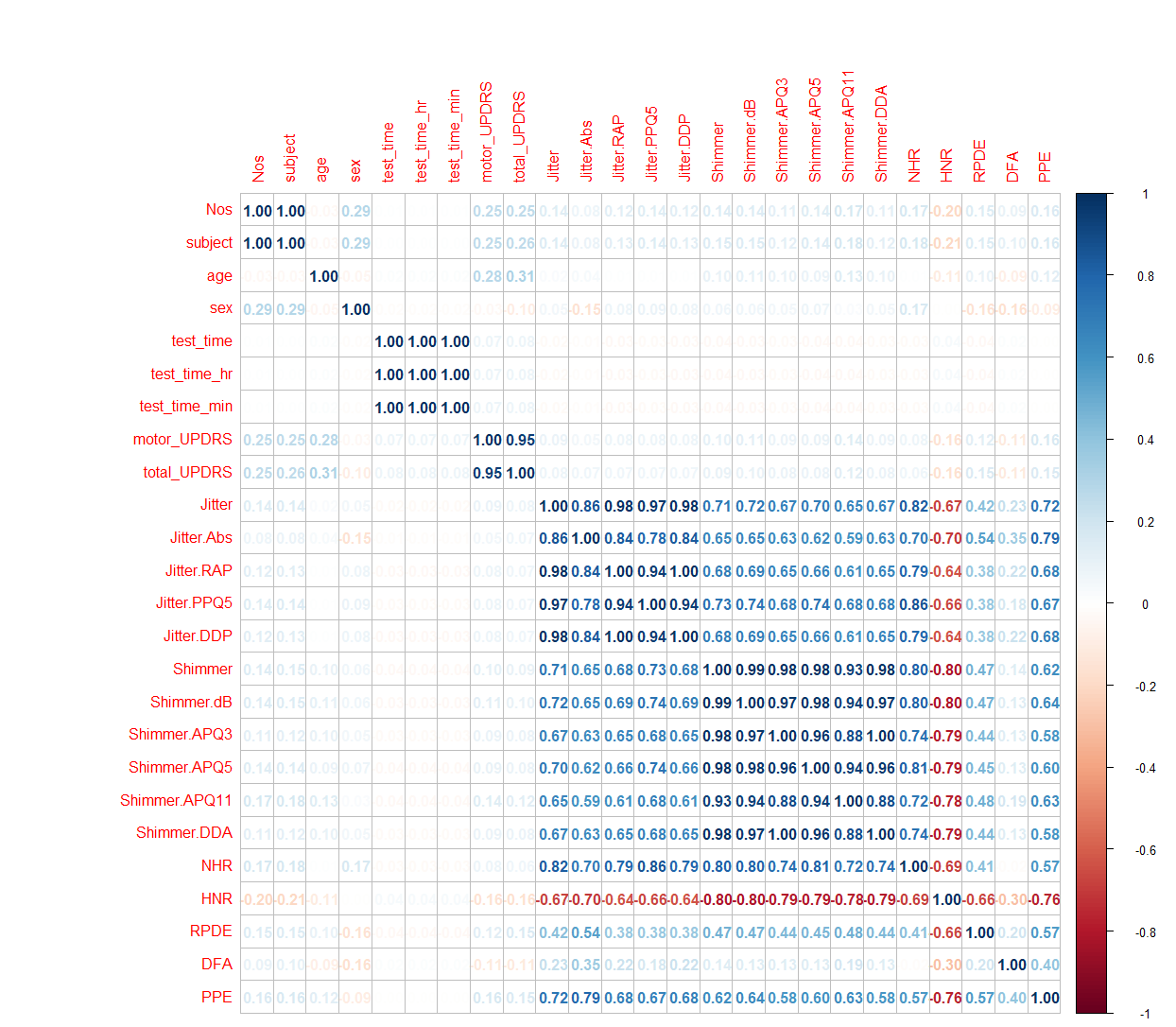
## PDTest summary

PDTest dataset is a subset of the Parkinson dataset, which is applied with the 80/20 rule. 20% subset of data is a clone to PDTrain dataset. There are 1,073 observations with the summary below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| Nos | 1 | 1515 | 2940 | 2949 | 4396 | 5874 |
| Subject | 1 | 11 | 22 | 21.57 | 32 | 42 |
| Age | 36 | 58 | 65 | 64.77 | 72 | 85 |
| Sex | 0 | 0 | 0 | 0.329 | 1 | 1 |
| test\_time | 0.3958 | 46.775 | 90.799 | 92.7311 | 138.4600 | 215.4900 |
| Test\_time\_hr | 9.517 | 1122.6 | 2179.176 | 2225.546 | 3323.04 | 5171.760 |
| Test\_time\_min | 571 | 67356 | 130751 | 133533 | 199382 | 310306 |
| Motor\_UPDRS | 5.038 | 15 | 21.000 | 21.306 | 27.594 | 39.511 |
| Total\_UPDRS | 7 | 21.54 | 27.67 | 28.98 | 36.4 | 54.99 |
| Jitter | 0.00106 | 0.00358 | 0.00487 | 0.005985 | 0.00665 | 0.71830 |
| Jitter.Abs | 3.940e-06 | 2.310e-05 | 3.340e-05 | 4.295e-06 | 5.190e-05 | 3.638e-04 |
| Jitter.RAP | 0.000420 | 0.001580 | 0.002220 | 0.002902 | 0.003230 | 0.043560 |
| Jitter.PPQ5 | 0.00052 | 0.00182 | 0.00245 | 0.00316 | 0.00335 | 0.04917 |
| Jitter.DDP | 0.0001250 | 0.004730 | 0.006670 | 0.008707 | 0.009690 | 0.130680 |
| Shimmer | 0.00306 | 0.01913 | 0.2675 | 0.03361 | 0.03931 | 0.23915 |
| Shimmer.dB | 0.0260 | 0.1760 | 0.2470 | 0.3075 | 0.3680 | 1.97 |
| Shimmer.APQ3 | 0.00161 | 0.00912 | 0.01318 | 0.01690 | 0.02037 | 0.16267 |
| Shimmer.APQ5 | 0.00194 | 0.01075 | 0.01556 | 0.01976 | 0.02341 | 0.16246 |
| Shimmer.APQ11 | 0.00249 | 0.01561 | 0.02221 | 0.02708 | 0.03221 | 0.16991 |
| Shimmer.DDA | 0.00484 | 0.02737 | 0.03954 | 0.05070 | 0.06110 | 0.48802 |
| NHR | 0.000774 | 0.011183 | 0.018116 | 0.0312272 | 0.031574 | 0.72537 |
| HNR | 1.799 | 19.588 | 21.984 | 21.740 | 24.410 | 34.555 |
| RPDE | 0.2588 | 0.4695 | 0.5390 | 0.5403 | 0.6102 | 0.9089 |
| DFA | 0.5147 | 0.5931 | 0.6409 | 0.6514 | 0.7072 | 0.8220 |
| PPE | 0.02634 | 0.15495 | 0.20276 | 0.21737 | 0.26443 | 0.59467 |

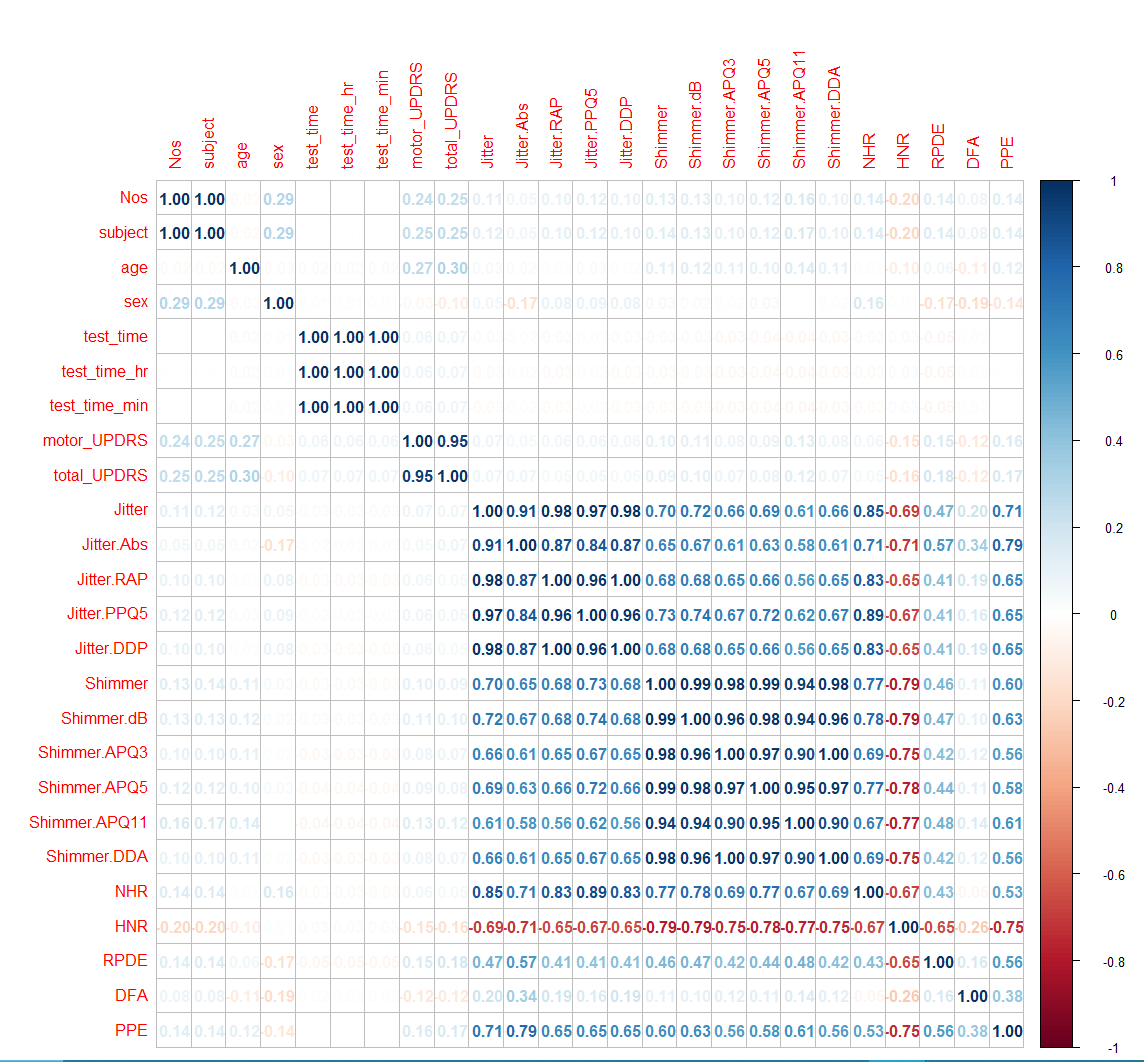
## Correlation Map for PDTrain Dataset

Correlation mapping of all PDTrain variables is below. The number with blue colour is a positive relationship, and the colour with orange is a negative relationship. For example, a dark blue with a value of one on Nos and subject means a solid positive relationship between these two variables. For the high-level summary, negative relationships were found on variables pairs with HNR, and other variables pairs such as Sex with RPDE, Sex with DFA, Sex with PPE, Nos with HNR, Subject with HNR, and so now. Other variables correlation are with a positive relationship.



## Correlation Map for PDTest Dataset

Similar to the PDTrain dataset, the correlation mapping of all PDTest variables are below. The heatmap pattern and the value are identical; the size of observations from the same dataset does not deviate much from the original.

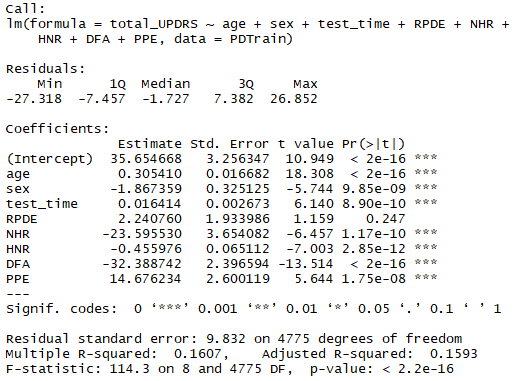


## Linear Regression on Training Set

Linear regression build model to predicts ‘total\_UPDRS’ with selected predictors with ‘age’, ‘sex’, ‘test\_time’, ‘RPDE’, ‘NHR’, ‘HNR’, ‘DFA’, ‘PPE’, the graphical presentation below

|  |  |
| --- | --- |
| Age | Sex |
|  |  |
| Test\_Time | RPDE |
|  |  |
| NHR | HNR |
|  |  |
| DFA | PPE |
|  |  |

Linear Regression Summary – PD Train dataset



The linear equation on total UPDRS below means one unit of total UPDRS increase required the factor of these predictors + 35.6547 intercepts. For RPDE, the p-value is 0.247, which is greater than the 0.05 confident level. Thus we accept the null hypothesis, and there is no relationship between RPDE and Total\_UPDRS. Therefore we remove this predictor from the linear regress model.

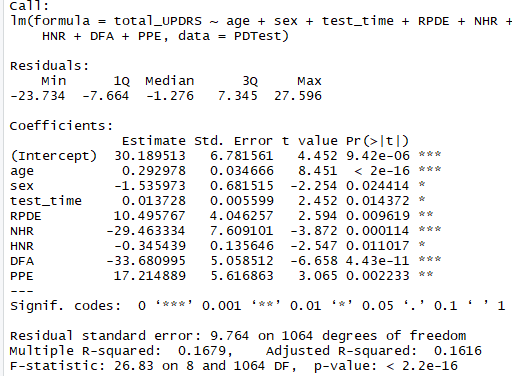
Total UPDRS = 0.3054(Age) – 1.867 (Sex) + 0.0164 (Test\_Time) – 23.5855 (NHR) – 0.4560 (HNR) – 32.3887 (DFA) + 14.6762 (PPE) + 35.6547

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 9.832, which is quite significant as it was calculated with 4,775 degrees of freedom (data point in consideration). R-square interprets how fit the data model is, with 0.1607 (R-square) and 0.1593 (adjusted R-square), it tells around 16% of response variable can be explained by predictor variable. In general, the R-square value close to 1 is better.

F-statistic is another indicator that explains the relationship between predictor and response variable. In regular practice, rejected the null hypothesis when the F-statistic value is larger than one. In this model, the F-statistic is 114.3, which is larger than 1. Same as previously mentioned, we reject the null hypothesis and, statistically, there is a relationship between predictor and response variable[[1]](#footnote-1).

## Linear Regression Summary – PD Test dataset



The linear summary of PD Test data above shows the p-values on all predictors with less than a signification level of 0.05. The null hypothesis has been rejected, and there is a relationship between predictors and observed variables. The linear equation on total UPDRS below means one unit of total UPDRS increase required the factor of these predictors + 30.189513 intercepts.

Total UPDRS = 0.2930(Age) - 1.5360(Sex) + 0.0137 (Test\_Time) + 10.4958 (RPDE) - 29.4633 (NHR) - 0.3454 (HNR) - 33.6810 (DFA) + 5.6169 (PPE) + 30.1895

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, the significant code has the combination of \*, \*\*, \*\*\* indicates we can reject the null hypothesis that there is a relationship between the response variable and predictors.

Residual standard error, similar to the training dataset, with 9.764, interprets the average deviation from the regression line. R-square interprets how fit the data model is, with 0.1679 (R-square) and 0.1616 (adjusted R-square), it tells around 16.79% of response variable can be explained by predictor variable. In this model, the F-statistic is 9.764, which is larger than 1, same as the training dataset; we reject the null hypothesis and, statistically, there is a relationship between predictor and response 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 9.832, and the MAE is 8.144. The residual standard error is calculated based on prediction, and MAE is calculated as the absolute, which gives a more precise (1.7 difference) on error measurement of the regression model. Similar to the Test dataset, the discrepancy of the two measures is 1.784 (9.764 – 7.980). These two regression models have identified differences in absolute error measurement.

On the other hand, Train dataset, the MAE value is 8.144, and in the Test dataset, the MAE value is 7.980; the test dataset has a smaller MAE value, which means the model is better in theory. However, when looking at the size of the dataset, the training dataset is four times bigger than the test dataset; the MAE value may be different or the same when the test dataset size is the same as the training dataset.

|  |  |
| --- | --- |
| Train Dataset | Test Dataset |
|  |  |

## Conclusion

There are several ways to interpret a dataset. This report focuses on the regression model with selected predictors, interpretation of the regression summary, and measurement comparison on MAE and standard error. Experiment results may be the difference when a dataset is split into a different group ratio, and even the summary looks similar to the original dataset; however, the sample size can show information differently. In statistic words, even a complete set dataset is split into two datasets with a 50:50 ratio. All the findings may not have the same.

## Reference

“QUICK GUIDE: INTERPRETING SIMPLE LINEAR MODEL OUTPUT IN R.” Quick guide: Interpreting simple linear model output in R, October 25, 2015. https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R.

# Appendix – R Code

#Set Working Directory

setwd("~/2. Lanagara College/DANA\_4810/Assignment2\_Parkinson")

#read in the data

library("readr")

parkinsons <- read\_csv('ParkinsonsClean2.csv')

#Split 80 20 Ratio

library('caTools')

PDSplit = sample.split(parkinsons$total\_UPDRS, SplitRatio = .8)

PDTrain = subset(parkinsons, PDSplit == TRUE)

PDTest = subset(parkinsons, PDSplit == FALSE)

#Check Row Count and statistic

count(PDTrain)

summary(PDTrain)

#Check Row Count and statistic

count(PDTest)

summary(PDTest)

library(corrplot)

M = cor(PDTrain)

corrplot(M, method = 'number') # colorful number

M = cor(PDTest)

corrplot(M, method = 'number') # colorful number

PDTrain\_lm <- lm(formula = total\_UPDRS ~age + sex + test\_time + RPDE +NHR + HNR + DFA + PPE, data=PDTrain)

summary(PDTrain\_lm)

######### Line graph ################

library('tidyverse')

library('ggplot2')

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = age)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = sex)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = test\_time)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = RPDE)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = NHR)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = HNR)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = DFA)) + geom\_point(color='blue') + geom\_smooth(method="lm")

ggplot(data= PDTrain,aes(x = total\_UPDRS, y = PPE)) + geom\_point(color='blue') + geom\_smooth(method="lm")

#----------- Cross check-----------------------

PDTest\_lm <- lm(formula = total\_UPDRS ~age + sex + test\_time + RPDE +NHR + HNR + DFA + PPE, data=PDTest)

summary(PDTest\_lm)

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

library(tidyverse)

library(modelr)#provides helper functions for computing regression model performance metrics

library(broom)#creates easily a tidy data frame containing the model statistical metrics

#Model 1, - full Dataset

PD\_Model <- lm(formula =total\_UPDRS ~age + sex + test\_time + RPDE +NHR + HNR + DFA + PPE, data = parkinsons)

summary(PD\_Model)

#install.packages('modelr')

library(modelr)

data.frame(

R2 = rsquare(PD\_Model, data = parkinsons),

RMSE = rmse(PD\_Model, data = parkinsons),

MAE = mae(PD\_Model, data = parkinsons)

)

#Model 2, - Train Dataset

PDTrain\_Model <- lm(formula =total\_UPDRS ~age + sex + test\_time + RPDE +NHR + HNR + DFA + PPE, data = PDTrain)

summary(PDTrain\_Model)

#install.packages('modelr')

library(modelr)

data.frame(

R2 = rsquare(PDTrain\_Model, data = PDTrain),

RMSE = rmse(PDTrain\_Model, data = PDTrain),

MAE = mae(PDTrain\_Model, data = PDTrain)

)

#Model 3 - Test Dataset

PDTest\_Model <- lm(formula =total\_UPDRS ~age + sex + test\_time + RPDE +NHR + HNR + DFA + PPE, data = PDTest)

summary(PDTest\_Model)

#install.packages('modelr')

library(modelr)

data.frame(

R2 = rsquare(PDTest\_Model, data = PDTest),

RMSE = rmse(PDTest\_Model, data = PDTest),

MAE = mae(PDTest\_Model, data = PDTest)

)

1. “QUICK GUIDE: INTERPRETING SIMPLE LINEAR MODEL OUTPUT IN R,” Quick guide: Interpreting simple linear model output in R, October 25, 2015, https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R. [↑](#footnote-ref-1)