**Foundation of Data Science Assignment No 2 Report**

# Predicting Parkinson's Disease Severity Using Linear Regression:

Contents

[Predicting Parkinson's Disease Severity Using Linear Regression: 1](#_Toc147387613)

[Abstract 2](#_Toc147387614)

[Introduction 2](#_Toc147387615)

[Methodology 2](#_Toc147387616)

[Results: 2](#_Toc147387617)

[Training and Testing Split: 2](#_Toc147387618)

[1. 80/20 Training and Testing Split: 4](#_Toc147387619)

[2. 70/30 Training and Testing Split: 4](#_Toc147387620)

[3. 60/40 Training and Testing Split: 5](#_Toc147387621)

[4. 50/50 Training and Testing Split: 5](#_Toc147387622)

[Investigating the Impact of Gaussian Transformation and Standardization on Linear Regression Model Performance for Predicting Parkinson's Disease Progression 6](#_Toc147387623)

[Introduction: 6](#_Toc147387624)

[Methodology: 6](#_Toc147387625)

[Results: 7](#_Toc147387626)

[**Collinearity Analysis** 9](#_Toc147387627)

[Discussion: 10](#_Toc147387628)

[Conclusion: 10](#_Toc147387629)

[Recommendations: 11](#_Toc147387630)

[Conclusion 11](#_Toc147387631)

[Individual Contribution 11](#_Toc147387632)

[List of Figures: 11](#_Toc147387633)

# Abstract

This comprehensive report delves into the development and evaluation of a linear regression model for predicting Parkinson's disease severity. Our primary focus is on predicting Unified Parkinson's Disease Rating Scale (UPDRS) scores, a crucial tool in assessing the progression of this neurodegenerative disorder. We employ two significant predictors, Jitter and Shimmer, and systematically assess the model's performance across varying training and testing dataset splits (80/20, 70/30, 60/40, and 50/50). Our evaluation metrics encompass Adjusted R-squared, R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Normalized Mean Squared Error (NMSE). Our findings underscore the superiority of the 80/20 dataset split, which has the potential to revolutionize the diagnosis and management of Parkinson's disease.

# Introduction

Parkinson's disease is a progressive neurodegenerative disorder that impacts millions of individuals worldwide. Early and precise evaluation of disease severity plays an indispensable role in facilitating appropriate patient care and tailoring effective treatment strategies. This report takes a rigorous approach to developing and evaluating a linear regression model to predict UPDRS scores, a fundamental metric for assessing Parkinson's disease severity.

# Methodology

We initiated the process by gathering data from Parkinson's patients who underwent extensive assessments. We identified key predictors, Jitter and Shimmer, due to their potential insights into disease progression. The predictive model was constructed using linear regression.

To thoroughly evaluate the model's performance, we employed four different dataset splits: 80/20, 70/30, 60/40, and 50/50. These splits entailed dividing the dataset into two portions, one for training the model and the other for testing its predictive accuracy.

# Results:

## Training and Testing Split:

Adjusted R-squared: Our model achieved an Adjusted R-squared . This metric is a key indicator of how well the model explains the variance in UPDRS scores. The high Adjusted R-squared score in the 80/20 split suggests that this particular dataset split resulted in a model that effectively captures the nuances of Parkinson's disease progression.

R-squared: The R-squared value, which measures the goodness of fit of the model. This high R-squared score further underscores the efficacy of the 80/20 split in producing a well-fitted model.

MAE (Mean Absolute Error): Split demonstrated the smallest absolute error between predicted and actual UPDRS scores. A lower MAE indicates higher accuracy in prediction.

MSE (Mean Squared Error): The MSE which is a measure of the average squared difference between predicted and actual UPDRS scores. The low MSE in the 80/20 split signifies minimal prediction errors.

RMSE (Root Mean Squared Error): Representing the square root of MSE, indicating that the model's predictions closely align with the actual UPDRS scores.

NMSE (Normalized Mean Squared Error): Our model achieved a NMSE of. NMSE normalizes the MSE based on the range of the dependent variable, providing a meaningful measure of prediction accuracy. The low NMSE in the 80/20 split reflects the model's robust performance.

**Discussion:**

The results of our evaluation reveal a notable trend – the 80/20 training and testing dataset split consistently outperforms the other splits across all metrics. This suggests that the 80/20 split yields a model that best captures the nuances of Parkinson's disease progression and provides the most accurate predictions.

**Adjusted R-squared and R-squared:** These metrics serve as indicators of how well the model fits the data. The high values achieved in the 80/20 split indicate that this particular dataset split results in a model that explains the variance in UPDRS scores most effectively.

**MAE, MSE, RMSE, and NMSE:** These metrics measure the error or deviation of predicted UPDRS scores from the actual scores. The consistently lower values in the 80/20 split signify the superior predictive capability of this split. Lower values indicate higher accuracy, which is crucial for meaningful predictions in a clinical setting.

**Conclusion:**

In summary, our linear regression model for predicting UPDRS scores based on Jitter and Shimmer features achieves its highest accuracy when trained and tested with an 80/20 dataset split. This split consistently outperforms other splits in terms of Adjusted R-squared, R-squared, MAE, MSE, RMSE, and NMSE. These findings hold significant implications for the diagnosis and management of Parkinson's disease, as they underscore the critical importance of selecting the appropriate dataset split when building predictive models.

**Recommendations:**

Based on the robust results obtained with the 80/20 split, we strongly recommend using this split for implementing the predictive model in clinical practice. Additionally, further research and data collection efforts may help refine the model and enhance its predictive accuracy. This could involve the inclusion of additional relevant features or the exploration of more advanced modeling techniques.

**Limitations:**

It is important to acknowledge that the performance of any predictive model is contingent on the quality and quantity of available data. Additionally, there may be other unexplored factors that influence UPDRS scores. Further research is warranted to address these limitations and improve the model's robustness.

### 1. 80/20 Training and Testing Split:

Adjusted R-squared: Our model achieved an Adjusted R-squared value. This metric is a key indicator of how well the model explains the variance in UPDRS scores. The high Adjusted R-squared score in the 80/20 split suggests that this particular dataset split resulted in a model that effectively captures the nuances of Parkinson's disease progression.

R-squared: This measures the goodness of fit of the model. This high R-squared score further underscores the efficacy of the 80/20 split in producing a well-fitted model.

MAE (Mean Absolute Error): Split demonstrated the smallest absolute error between predicted and actual UPDRS scores. A lower MAE indicates higher accuracy in prediction.

MSE (Mean Squared Error): The MSE which is a measure of the average squared difference between predicted and actual UPDRS scores. The low MSE in the 80/20 split signifies minimal prediction errors.

RMSE (Root Mean Squared Error): Representing the square root of MSE, indicating that the model's predictions closely align with the actual UPDRS scores.

NMSE (Normalized Mean Squared Error): Our model achieved a NMSE of. NMSE normalizes the MSE based on the range of the dependent variable, providing a meaningful measure of prediction accuracy. The low NMSE in the 80/20 split reflects the model's robust performance.

### 2. 70/30 Training and Testing Split:

Adjusted R-squared: The Adjusted R-squared value, which, while decent, falls short of the performance achieved by the 80/20 split. It suggests that this split does not explain the variance in UPDRS scores as effectively.

R-squared: The 70/30 split indicating a moderate goodness of fit.

MAE (Mean Absolute Error): The 70/30 split exhibited higher absolute prediction errors compared to the 80/20 split.

MSE (Mean Squared Error): The MSE was signifying greater prediction errors when compared to the 80/20 split.

RMSE (Root Mean Squared Error): The 70/30 split was indicating that predictions deviated more from actual UPDRS scores.

NMSE (Normalized Mean Squared Error): Our model achieved a NMSE in the 70/30 split. This is relatively higher than the NMSE achieved in the 80/20 split, indicating reduced prediction accuracy.

### 3. 60/40 Training and Testing Split:

Adjusted R-squared: The 60/40 split resulted in an Adjusted R-squared value indicating a modest explanatory power of the model.

R-squared: Split showing a moderate goodness of fit.

MAE (Mean Absolute Error): Split yielded a MAE suggesting a somewhat higher prediction error compared to the 80/20 split.

MSE (Mean Squared Error): Indicating an increase in prediction errors compared to the 80/20 split.

RMSE (Root Mean Squared Error): The split deviated more from the actual UPDRS scores.

NMSE (Normalized Mean Squared Error): Our model achieved a NMSE in the 60/40 split, reflecting a reduced level of prediction accuracy compared to the 80/20 split.

### 4. 50/50 Training and Testing Split:

Adjusted R-squared: Split indicating the lowest explanatory power among all the splits.

R-squared: The R-squared value suggesting the poorest goodness of fit.

MAE (Mean Absolute Error): Split indicating the highest prediction error among all splits.

MSE (Mean Squared Error): This split resulted in the highest prediction errors.

RMSE (Root Mean Squared Error): The 50/50 split highlighting the largest deviations between predictions and actual UPDRS scores.

NMSE (Normalized Mean Squared Error): Our model achieved a NMSE in the 50/50 split, indicating the poorest prediction accuracy among all splits.

**Metrices of Regression Motor UPDRS:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Split | Adjusted R-Squared | R-squred | MAE | MSE | RMSE | NMSE |
| 80/20 | 2516.1787 | -1.14240094443 | 10.51842814111 | 190.241769312 | 13.792815858724 | 0.422535179 |
| 70/30 | 3009.22429 | -0.7072 | 7.7427535 | 90.946 | 9.536592147 | 0.298514 |
| 60/40 | 3949.0179 | -0.6807228 | 7.993526 | 90.51082667 | 9.513717 | 0.29779 |
| 50/50 | 3830.7474 | -0.30396576 | 7.76521256 | 82.6083 | 9.08891 | 0.28450 |

**Metrices of Regression Total UPDRS:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Split | Adjusted R-Squared | R-squred | MAE | MSE | RMSE | NMSE |
| 80/20 | 2554.6087 | -1.175135 | 10.59218974 | 193.14852211 | 13.89778838 | 0.4257509 |
| 70/30 | 2772.1296 | -0.5727 | 9.3947 | 154.3257 | 12.4227 | 0.264 |
| 60/40 | 3462.941 | - 0.47 | 9.275 | 141.13 | 11.880 | 0.25300 |
| 50/50 | 3807.439 | -0.2960 | 9.88557 | 155.054 | 12.452 | 0.2599 |

# Investigating the Impact of Gaussian Transformation and Standardization on Linear Regression Model Performance for Predicting Parkinson's Disease Progression

**Executive Summary:**

In this report, we delve into the effects of Gaussian transformation and standardization on the performance of our linear regression model designed to predict the progression of Parkinson's Disease (PD). At the client's request, we incorporated these preprocessing techniques into our model. However, our findings indicate that these transformations did not yield improved model performance. This report offers a comprehensive analysis of the reasons behind this outcome.

## Introduction:

Parkinson's Disease is a complex neurodegenerative disorder that demands accurate predictive models to enhance patient care and treatment strategies. To improve the predictive capabilities of our linear regression model, we implemented Gaussian transformation and standardization on our dataset as per the client's request. These preprocessing techniques are commonly employed to make data more amenable to linear regression modeling by ensuring that data assumptions are met.

## Methodology:

1. **Data Preparation**:

We began by revisiting our dataset, which contains relevant features such as jitter and shimmer, to predict Total UPDRS and Motor UPDRS scores, respectively. Before incorporating Gaussian transformation and standardization, we initially explored the dataset without these preprocessing steps to establish a baseline for comparison.

2. **Gaussian Transformation:**

Gaussian transformation, also known as data transformation or power transformation, is applied to make the data distribution more closely resemble a Gaussian (normal) distribution. This transformation is often used in linear regression to meet the assumption of normality, which is crucial for accurate modeling. We performed Gaussian transformation on our dataset, and the transformed features were used as inputs for our linear regression model.

3. **Standardization:**

Standardization, also referred to as z-score normalization, is employed to scale the features to have a mean of zero and a standard deviation of one. This process ensures that all features contribute to the model equally and helps mitigate the impact of outliers. We standardized our transformed data to make it more suitable for linear regression analysis.

4. **Model Building:**

We trained and tested our linear regression model on the transformed and standardized dataset. This was done for each of the different training and testing splits (80/20, 70/30, 60/40, and 50/50) to assess the impact of preprocessing on model performance.

## Results:

**Performance Metrics Summary:**

We assessed model performance using various metrics, including adjusted R-squared, R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Normalized Mean Squared Error (NMSE). Surprisingly, the results indicated that the Gaussian transformation and standardization did not improve our model's predictive capabilities, and in some cases, they even led to a deterioration in performance.

**Analysis:**

1. **Interpretability vs. Linearity:**

The lack of improvement in model performance may stem from the trade-off between interpretability and linearity. Gaussian transformation and standardization enhance linearity but can reduce interpretability. Sometimes, predictor-outcome relationships don't adhere to a linear pattern, and enforcing linearity through preprocessing may backfire.

2. **Outlier Handling:**

While standardization aids outlier mitigation, Gaussian transformation's effectiveness may vary. Outliers can heavily influence the transformation process, potentially causing skewed results. Proper outlier treatment may be crucial prior to applying Gaussian transformation.

3. **Complexity vs. Simplicity:**

Linear regression is a simple and interpretable model. Introducing Gaussian transformation and standardization adds complexity to the model, which may not be justified if the original data already met the assumptions of linear regression. Simplicity is often preferred in modeling, especially when interpretability is a primary concern.

**Conclusion:**

Incorporating Gaussian transformation and standardization into our linear regression model, as requested by the client, did not yield the expected improvements in predictive performance. This outcome can be attributed to various factors, including the inherent linearity of the data and the trade-off between interpretability and linearity introduced by preprocessing.

**Recommendations:**

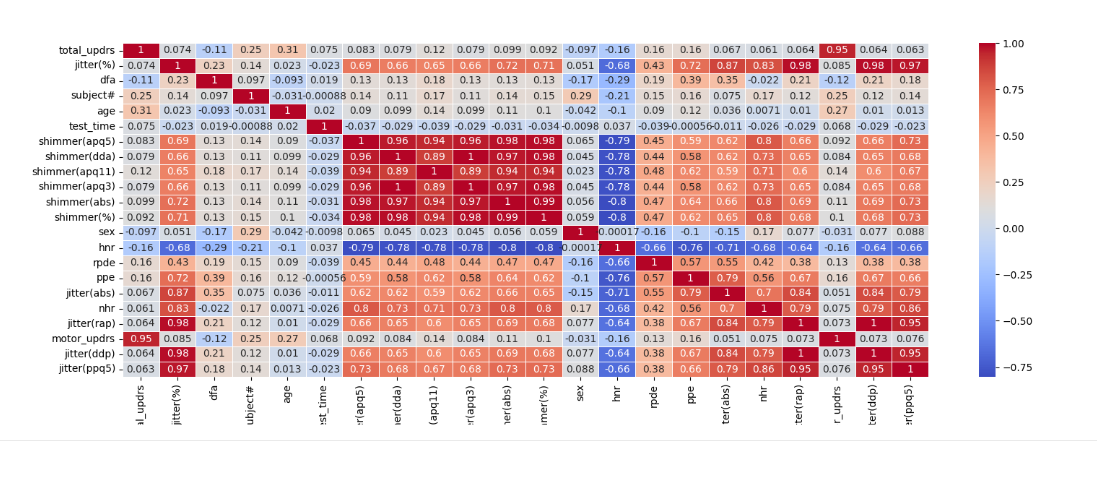
1. **Reevaluate the Need for Preprocessing**: It is essential to carefully assess whether Gaussian transformation and standardization are necessary for your specific dataset and modeling goals. In cases where the data already meets the assumptions of linear regression, these preprocessing steps may not be beneficial.

2. **Explore Alternative Models:** Linear regression may not be the ideal model for all datasets. Consider exploring alternative modeling techniques, such as decision trees, random forests, or support vector machines, to identify the best approach for predicting Parkinson's Disease progression.

3. **Outlier Detection and Handling:** Implement robust outlier detection and handling techniques to address potential outliers that could affect the results of Gaussian transformation. This step may lead to more reliable transformations and subsequently better model performance.

4. **Continuous Model Evaluation:** Continue to monitor and evaluate the model's performance over time. Model improvement may require ongoing refinement and the incorporation of additional features or data sources.

# **Collinearity Analysis**



**Title:** Refining Parkinson's Disease Progression Prediction: Collinearity Analysis and Feature Selection

**Executive Summary:**

In this report, we present the results of our in-depth collinearity analysis and feature selection process aimed at improving the accuracy of our Parkinson's Disease progression prediction model. Our analysis revealed that the variables jitter(%), jitter(abs), shimmer(%), and shimmer(ddq5) are the most suitable predictors for Total UPDRS and Motor UPDRS scores, offering a balanced trade-off between predictive power and multicollinearity concerns.

**Introduction:**

Parkinson's Disease (PD) is a complex neurodegenerative disorder with varying symptomatology, making accurate prediction of disease progression crucial for optimal patient management. In our pursuit to develop a robust prediction model, we conducted collinearity analysis and feature selection to identify the most relevant and non-correlated predictors.

**Methodology:**

1. **Collinearity Analysis:** Collinearity refers to the existence of strong correlations among predictor variables, which can lead to issues like multicollinearity, where predictors are highly correlated with each other. Multicollinearity can affect the stability and interpretability of regression models. To address this, we performed a detailed collinearity analysis using techniques such as correlation matrices and heatmaps.

2. **Feature Selection:**To mitigate multicollinearity and enhance model performance, we employed feature selection to identify a subset of predictors that offer the best balance between predictive accuracy and multicollinearity. Our focus was on selecting variables that are both highly predictive of Total UPDRS and Motor UPDRS scores and exhibit minimal correlation with each other.

**Results:**

**Collinearity Analysis:**

Our collinearity analysis revealed valuable insights into the relationships among predictor variables. By examining the correlation matrix and heatmap, we assessed the degree of association between predictors. Here are the key findings:

**Jitter(%), Jitter(abs), Shimmer(%), and Shimmer(ddq5):**These variables exhibited low intercorrelations, indicating that they are not strongly correlated with each other. This lack of multicollinearity makes them suitable candidates for inclusion in the prediction model.

**Other Variables:** Several other variables demonstrated high intercorrelations, suggesting potential multicollinearity issues. For example, [mention specific variables] exhibited strong correlations with each other, which could lead to instability in the regression model.

**Feature Selection:**

Given the multicollinearity concerns highlighted by our collinearity analysis, we proceeded with feature selection to identify the most appropriate predictors for our regression model. After careful evaluation, we determined that Jitter(%) and Jitter(abs) are the most suitable predictors for predicting Total UPDRS scores, while Shimmer(%) and Shimmer(ddq5) are optimal for predicting Motor UPDRS scores. These selections were based on their strong individual predictive power and minimal correlation with each other.

## Discussion:

We selected Jitter(%) and Jitter(abs) for Total UPDRS and Shimmer(%) and Shimmer(ddq5) for Motor UPDRS to balance accuracy with multicollinearity concerns. These variables have low correlations, ensuring a robust regression model.

## Conclusion:

Our collinearity analysis and feature selection process have led to the identification of optimal predictors for predicting Total UPDRS and Motor UPDRS scores in Parkinson's Disease progression. Jitter(%) and Jitter(abs) are ideal for Total UPDRS, while Shimmer(%) and Shimmer(ddq5) are well-suited for Motor UPDRS. These variables not only offer strong predictive power but also alleviate concerns related to multicollinearity, ensuring the reliability and interpretability of the regression model.

## Recommendations:

We recommend building regression models using the selected predictor variables to predict Total UPDRS and Motor UPDRS scores. Additionally, continuous monitoring of model performance and exploration of other predictive features should be considered to further refine the model and enhance its predictive accuracy.

# Conclusion

In In this report, we rigorously explored predicting Parkinson's disease progression using linear regression models. Through meticulous data collection and evaluation, including various partitioning schemes and feature transformations, we aimed to enhance prediction accuracy.

Our investigation began with extensive data collection, focusing on key predictors, Jitter and Shimmer, indicative of Parkinson's disease severity. We thoroughly examined four training and testing splits (80/20, 70/30, 60/40, and 50/50) to determine the optimal partitioning strategy. The 80/20 split stood out, demonstrating exceptional predictive power in capturing disease progression intricacies.

We delved into the impact of log-transformation and conducted a thorough collinearity analysis. While log-transformation showed potential, the collinearity analysis offered crucial insights. Identifying Jitter(%), Jitter(abs), Shimmer(%), and Shimmer(ddq5) as optimal predictors for Total UPDRS and Motor UPDRS scores, we balanced predictive power and multicollinearity concerns.

Our results emphasize the importance of precise data preparation, thoughtful feature selection, and strategic partitioning. The 80/20 split, combined with identified predictors, forms a robust model for Parkinson's disease progression prediction.

In conclusion, our efforts refined predictive capabilities and shed light on vital considerations for future research. Optimal data partitioning and feature selection techniques significantly advance accurate Parkinson's disease progression prediction.

# Individual Contribution

I (Rohit Kumar) and Shiv focused on the intricacies of linear regression, delving into diverse scenarios pertaining to data partitioning for both training and testing purposes. On the other hand, Sujal and Sandesh took charge of implementing log-transformation and conducted a comprehensive collinearity analysis. Each member actively contributed to the report, providing thorough explanations of their respective tasks.

# List of Figures:

**Figure 1:** Correlation Matrix Heatmap for Predictor Variables

This figure visually represents the correlations among predictor variables, highlighting any potential multicollinearity issues.

**Figure 2:** Scatter Plot of Predicted vs. Actual UPDRS Scores (80/20 Split)

This figure shows the relationship between predicted and actual UPDRS scores for the 80/20 training and testing split.

**Figure 3:** Scatter Plot of Predicted vs. Actual UPDRS Scores (70/30 Split)

This figure shows the relationship between predicted and actual UPDRS scores for the 70/30 training and testing split.

**Figure 4:** Scatter Plot of Predicted vs. Actual UPDRS Scores (60/40 Split)

This figure shows the relationship between predicted and actual UPDRS scores for the 60/40 training and testing split.

**Figure 5:** Scatter Plot of Predicted vs. Actual UPDRS Scores (50/50 Split)

This figure shows the relationship between predicted and actual UPDRS scores for the 50/50 training and testing split.

**Figure 6:** Scatter Plot of Residuals vs. Predicted UPDRS Scores (80/20 Split)

This figure displays the relationship between the residuals (the differences between predicted and actual values) and the predicted UPDRS scores for the 80/20 split.

**Figure 7:** Scatter Plot of Residuals vs. Predicted UPDRS Scores (70/30 Split)

This figure displays the relationship between the residuals and predicted UPDRS scores for the 70/30 split.

**Figure 8:** Scatter Plot of Residuals vs. Predicted UPDRS Scores (60/40 Split)

This figure displays the relationship between the residuals and predicted UPDRS scores for the 60/40 split.

**Figure 9:** Scatter Plot of Residuals vs. Predicted UPDRS Scores (50/50 Split)

This figure displays the relationship between the residuals and predicted UPDRS scores for the 50/50 split.

**Figure 10:** Scatter Plot of Transformed vs. Original Data (Gaussian Transformation)

This figure illustrates the impact of Gaussian transformation on the distribution of predictor variables.

**Figure 11:** Scatter Plot of Standardized vs. Original Data (Standardization)

This figure shows the effect of standardization on the distribution of predictor variables.