AMS572 PROJECT REPORT

Group Members: Lihan Huang, Saransh Surana and Bernard Tenreiro

Linking Writing Processes to Writing Quality | Kaggle

Use typing behavior to predict essay quality

ABSTRACT

In our project, we aim to assess the influence of the writing process on writing quality using R. First we will process the data for testing two different hypotheses. The first hypothesis employs a Z-test to determine significant differences in the mean of variables between individuals with high and low scores. The second hypothesis employs multiple linear regression to evaluate the significance of the coefficients of certain independent variables in the model to predict the score.

INTRODUCTION

Our dataset was found in a Kaggle competition. The objective of the competition is to predict writing quality by examining the impact of typing behavior on essay outcomes. Our project will focus on analyzing the dataset for our project, rather than the competition itself. Leveraging a robust dataset of keystroke logs, the generated datasets describe the intricate relationship between an individual's writing behaviors and performance, which could provide valuable insights for writing instruction, the development of automated writing evaluation techniques, and intelligent tutoring/teaching systems.

DATA DESCRIPTION

Kaggle: Keystroke Data Collection Procedure

The data collection involved recruiting participants via Amazon Mechanical Turk for activities like demographic surveys, typing tests, an argumentative essay task, and a vocabulary test on a dedicated website. To have consistent data, participants were required to use computers with the same keyboard. In the essay portion, the participants had 30 minutes to write a persuasive essay to SAT prompts, with instructions such as a minimum 200-word length and restrictions on external references. The process included monitoring mechanisms to ensure task focus, issuing warnings for extended inactivity or attempts to switch windows. This meticulous approach aimed to capture nuanced writing behaviors during the argumentative task and greatly enriched the raw dataset.

Kaggle: Keystroke Logging Program

The keystroke logging program (written in JavaScript) measures the behavior of the writing participant. It recorded detailed keystroke and mouse events, providing timestamps, cursor positions, and operation types (e.g., input, delete). The output table showcased the chronological indexing of events, initiation and conclusion times, operation durations, cursor positions, word counts, and specific text changes. This comprehensive logging approach facilitated a thorough understanding of participants' writing processes during the argumentative task. Below is an example of what the actual raw dataset looks like for a certain individual.

Event ID	Down Time	Up Time	Action Time	Event	Position	Word Count	Text Change	Activity
1	30185	30395	210	Leftclick	0	0	NoChange	Nonproduction
2	41006	41006	0	Shift	0	0	NoChange	Nonproduction
3	41264	41376	112	I	1	1	I	Input
4	41556	41646	90	Space	2	1		Input
5	41815	41893	78	b	3	2	b	Input
6	42018	42096	78	e	4	2	e	Input
7	42423	42501	78	l	5	2	l	Input
8	42670	42737	67	i	6	2	i	Input
9	42873	42951	78	e	7	2	e	Input
10	43041	43109	68	v	8	2	v	Input
11	43289	43378	89	Space	9	2		Input
12	44560	44605	45	Backspace	8	2		Remove/Cut
13	44661	44762	101	e	9	2	e	Input
14	44954	45032	78	Space	10	2		Input
15	45325	45381	56	t	11	3	t	Input
16	45460	45538	78	h	12	3	h	Input
17	45640	45730	90	a	13	3	a	Input
18	45741	45808	67	t	14	3	t	Input
19	45933	46011	78	Space	15	3		Input

Kaggle: Raw Dataset

The raw dataset consists of two files, named train_scores.csv, train_logs.csv. These files contain the following information as shown below.

train_scores.csv

- id - The unique ID of the essay
- score - The score the essay received out of 6 (Dependent Variable)

train_logs.csv

Input logs to be used as training data. To prevent reproduction of the essay text, all alphanumeric character inputs have been replaced with the "anonymous" character q; punctuation and other special characters have not been anonymized. From the Kaggle page:

- id - The unique ID of the essay
- event_id - The index of the event, ordered chronologically
- down_time - The time of the down event in milliseconds
- up_time - The time of the up event in milliseconds

- action_time - The duration of the event (the difference between down_time and up_time)
- activity - The category of activity which the event belongs to
    - Nonproduction - The event does not alter the text in any way
    - Input - The event adds text to the essay
    - Remove/Cut - The event removes text from the essay
    - Paste - The event changes the text through a paste input
    - Replace - The event replaces a section of text with another string
    - Move From [x1, y1] To [x2, y2] - The event moves a section of text spanning character index x1, y1 to a new location x2, y2
- down_event - The name of the event when the key/mouse is pressed
- up_event - The name of the event when the key/mouse is released
- text_change - The text that changed as a result of the event (if any)
- cursor_position - The character index of the text cursor after the event
- word_count - The word count of the essay after the event

Data Processing and Keystroke Measures

We cannot compute our hypothesis on raw data, but with the help of the ID available in both the raw data files, we computed multiple variables. There are some keystroke measures already given in the Kaggle page. Below is the description of these variables and how we can compute these variables in R.

Production Rate

The rate of written language production can be measured by counting the number of characters, words, clauses, sentences, or text change in the writing process or written product generated per unit of time. Example measures are as follows.

- Number of characters per minute : We compute the characters per minute for typing-related activities like input, replace, paste, extracting start and end times of each activity, determining total typing time, and tallying characters processed.
- Number of words per minute : We compute the words per minute by taking the word_count of each ID, and then divide the total typing time, expressed in minutes.

Pause

Pauses are generally defined as inter-keystroke intervals (IKI) above a certain threshold (e.g. 2000 milliseconds). To elaborate, IKI is the time between consecutive key presses. Pausing is measured by duration and frequency from various dimensions. Examples are as follows:

- Number of pauses: The number of pauses is identified during typing sessions by detecting intervals exceeding a specified threshold (2000 milliseconds)
- Proportion of pause time : The proportion_of_pause_time can be calculated by identifying the intervals exceeding a specified threshold, then by measuring the total pause time, and calculating the proportion relative to the total typing duration.
- Mean pause length : To pause length can be measured by the duration of each pause for that event and then find the mean of all lengths

- Pause lengths or frequencies within words : Counts the number of pauses then divided by total writing time and computes the pause frequency per minute.

Revision

Revisions are the event of either deletions or insertions. A deletion is defined as the removal of any stretch of characters from a text whereas an insertion refers to a sequence of activities to add characters to a growing text (except the end). Below are some measure present in the dataset:

- Number of deletions: Deletion count is determined by identifying remove/cut activities, providing a quantitative measure of the number of deletions.
- Number of insertions: Total insertions are computed by identifying instances of 'input,' 'paste,' or 'replace' activities, resulting in the count of insertions.
- Length of deletions: Length of deletions is determined by identifying each entry related to 'Remove/Cut' activities and summing the character lengths for accurate measurement.
- Length of insertions: Length of insertions determined by identifying each entry related to 'Input,' 'Paste,' or 'Replace' activities and summing the character lengths for accurate measurement.
- Proportion of deletions: The proportion of deletions is computed by taking the total time dedicated to deletions and expressing it as a percentage of the overall typing duration.
- Proportion of insertions :The proportion of insertions is computed by taking the total time dedicated to insertions and expressing it as a percentage of the overall typing duration.

Burst

Bursts refer to the periods in text production in which stretches of texts were continuously produced with no pauses and/or revisions. There are mainly two types of bursts: P-bursts that refer to the written segments terminated by pauses, and R-bursts that describe the segments terminated by an evaluation, revision or other grammatical discontinuity

- Number of P-bursts : The number of p bursts can be done by identifying activities which are inputs, without pauses. For each p-burst, it accumulates character length and time duration, updating the count.
- Number of R-bursts : The number of r-bursts can be done by identifying activities which are non-inputs, without pauses. For each r-burst, it accumulates character length and time duration, updating the count
- Proportion of P-bursts :To calculate the proportion of p-bursts, we determine the start and end times of p-bursts relative to the overall typing duration; it measures the percentage of time dedicated to uninterrupted input.
- Proportion of R-bursts: To calculate the proportion of r bursts, we determine the start and end times of r-bursts relative to the overall typing duration, it measures the percentage of time dedicated to Non-Input type activities.
- Length of P-bursts: The length of p-bursts can be calculated by identifying the p-bursts based on intervals without pauses and 'Input' activities, accumulating the character length for each burst.

- Length of R-bursts : The length of r-bursts can be calculated by identifying the r-bursts based on intervals without pauses and Non 'Input' activities, accumulating the character length for each burst.

Process Variance

Process variance examines writing fluency dynamics over time. It involves dividing the writing process into intervals and calculating characters produced, normalized to average characters per minute. The standard deviation of characters per interval indicates process variance.

Subset Selection and Missing Data

1. Our generated dataset consists of more than 2000 rows, so to do our hypothesis testing we will select 200 rows and perform testing.
2. The tests will be performed on the different versions of the dataset including no missing data, random missing data, non-random missing data to differentiate and compare the results

Complete Data (V0)

1. In this dataset, we would have our complete data, with no missing values.
2. This is our original dataset.

10% Random Missing Data (R10)

1. While processing the raw dataset, we would randomly remove 10% of the values and then compute our dataset.
2. After computing our dataset, we would choose a subset to perform our testing.

20% Random Missing Data (R20)

1. While processing the raw dataset, we would randomly remove 20% of the values and then compute our dataset.
2. After computing our dataset, we would choose a subset to perform our testing.

30% Random Missing Data (R30)

1. While processing the raw dataset, we would randomly remove 30% of the values and then compute our dataset.
2. After computing our dataset, we would choose a subset to perform our testing.

Non-Random Missing Data (NR)

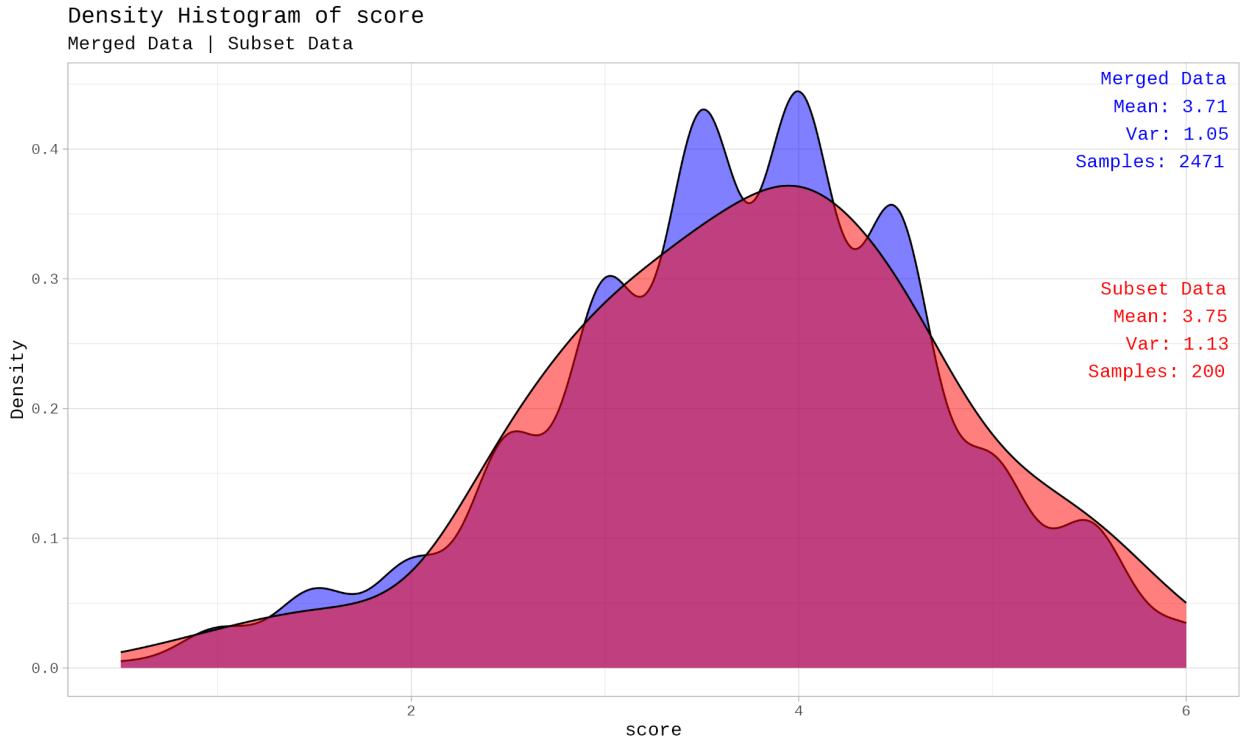
1. To simulate non-random missing data, we drop activities “Nonproduction”, “Replace” and “Paste”, assuming they are not recorded due to some issues/errors in the collection procedure.
2. Later we choose a subset to perform our hypothesis testing.

Note: All 5 versions of data available in the google drive found here: processed_data

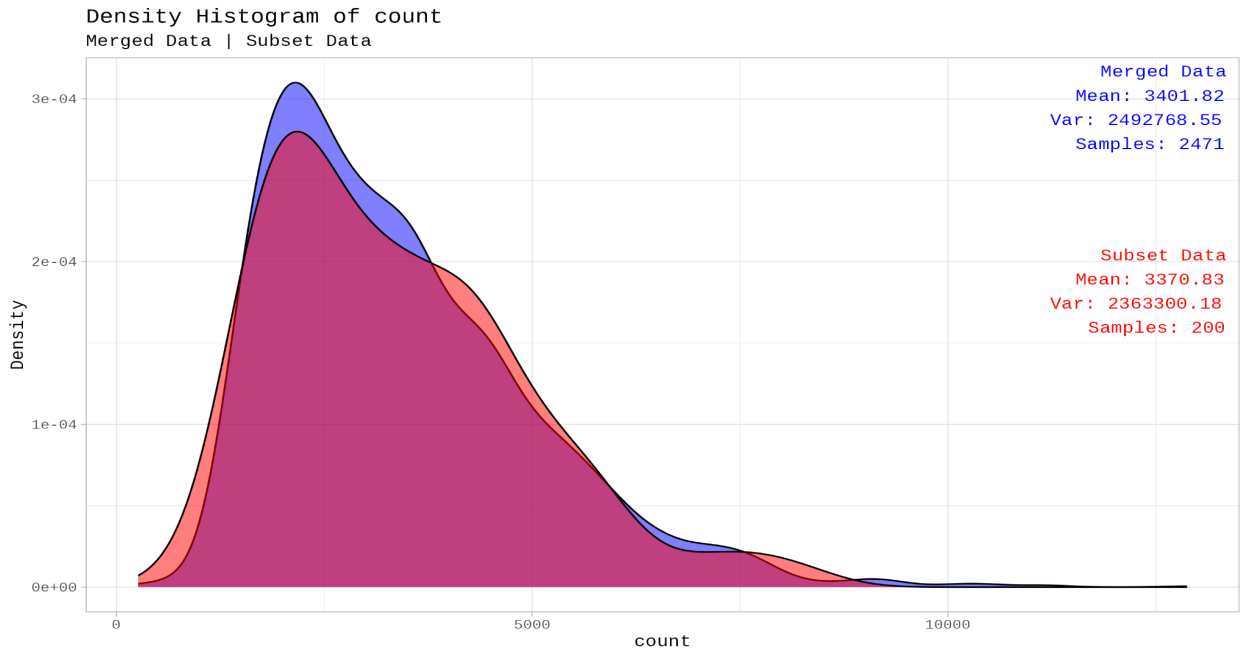
Exploratory Data Analysis

After we merge our dataset with the help of the individual's unique ID, the new dataset contains more than 2000 rows. For the subset we are going to take about 200 rows. Here are some visualizations with respect to score.

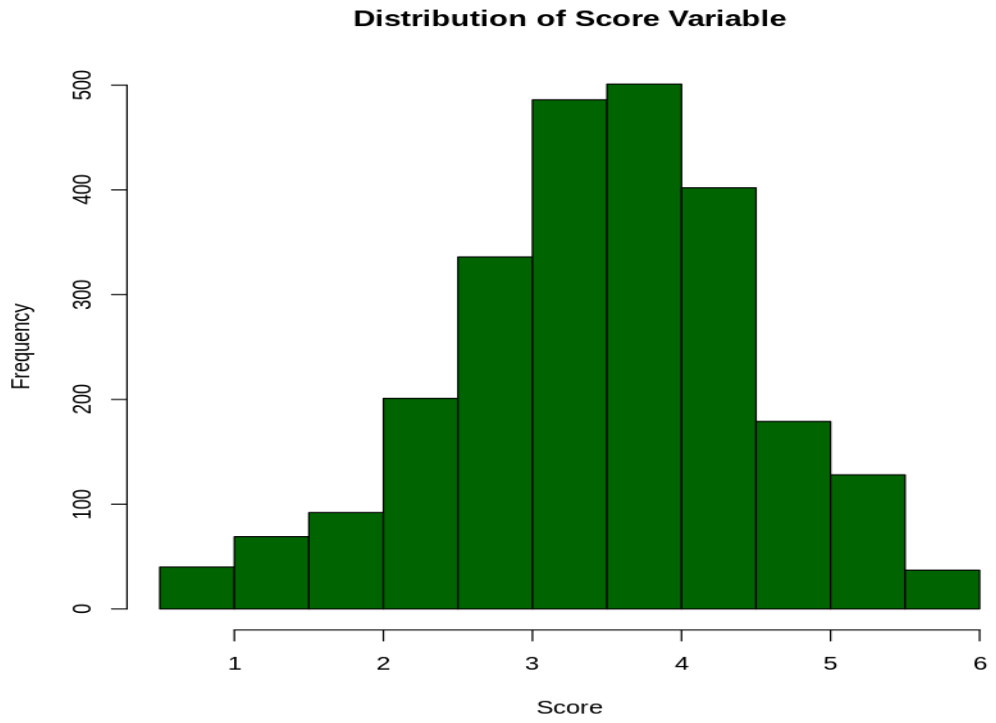
1. Below is the density histogram of score between merged data and subset data.



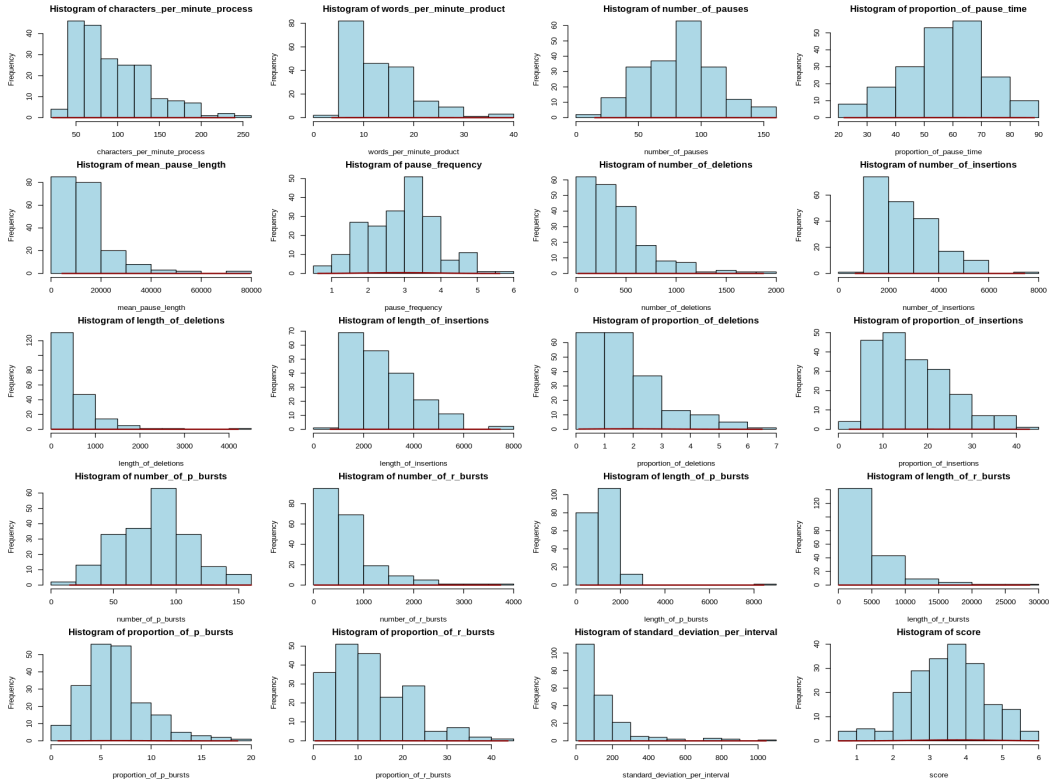
2. Below is the density histogram of count between Merged data and subset data



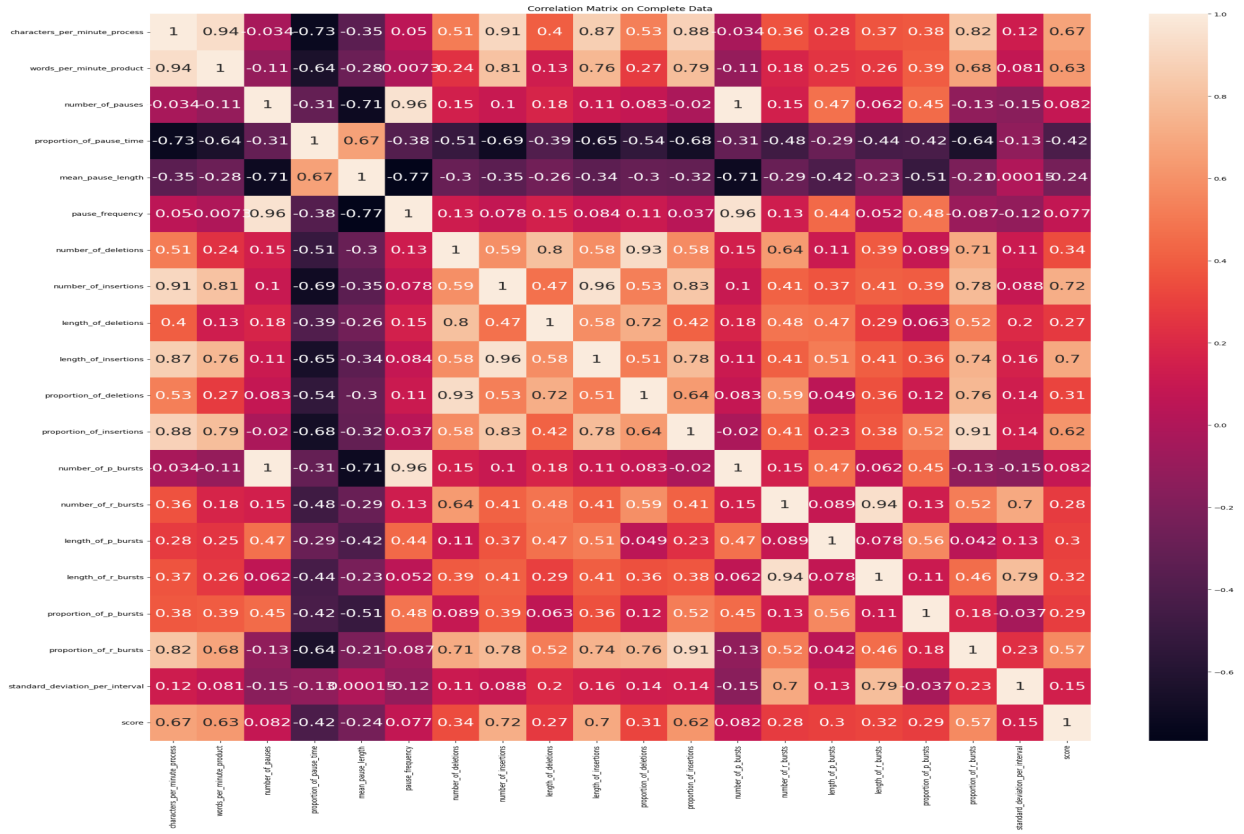
3. Distribution of score variable using Histograms



4. Visualization on complete data



5. Correlation Matrix on Complete data



Hypothesis 1: Z-Test

Introduction

The goal of this hypothesis test is to determine whether there is a significant difference in the mean of each variable in our dataset between individuals with high and low scores. Two groups are defined based on a threshold based on the mean score, and a Z-test is performed to assess the statistical significance of the observed differences.

Data Preparation

In order to perform a hypothesis test to investigate potential differences in the sample means of each variable, a classification threshold needed to be established. We set the threshold by dividing the sample into two different categories by the mean. Since the mean value is 3.79, we take 4 as a threshold. Two groups were created, where the high_score_group consists of individuals with score values above the threshold, and the low_score_group consists of individuals with scores below the threshold. This classification will give an idea of the variable values for individuals with higher and lower scores. We can now calculate sample means,

sample variances, and sample sizes for each variable for both the high and low score groups with a given dataset to perform hypothesis testing.

Assumptions

1. Normal Distribution: We assume the means of the two groups to be normally distributed. This is inferred by the central limit theorem (CLT) along with Slutsky's Theorem states that with sufficiently large sample sizes, the sampling distribution of the sample mean tends to a normal distribution, irrespective of the original distribution.
2. Convergence Assumption: For large sample sizes, the sample mean and sample variances can be used to estimate the population mean and population variances.
3. Independent Samples: Each value of both samples are from one individual, and each individual is independent of all other individuals, so these two samples are independent and identically distributed by their own distribution.

These assumptions are all true for our dataset, and thus satisfy the assumptions for a Z-Test.

Z-Test

This z-test used to compare the difference between two independent samples with unknown variances with a large sample size. The formula for the Z statistic is:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For the dataset, the following values are defined as:
$\bar{X}$: Sample mean in the high scores group
$\bar{Y}$: Sample mean in the low scores group
μ_1: Unknown population mean in the high scores group
μ_2: Unknown population mean in the low scores group
s_1^2 : Sample variance in the high score group
s_2^2 : Sample variance in the low score group
n_1 : sample size of the high score group
n_2 : sample size of the low score group

Hypothesis:

H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$ or equivalently, H_0: $\mu_1 - \mu_2 = 0$ vs H_1: $\mu_1 - \mu_2 \neq 0$
The null hypothesis is that the mean of a certain variable between individuals with a high score group is equal than the low scores group, while the alternative hypothesis states that the mean number of pauses in the high scores group are not equal. The significance level alpha (α) is set at 0.05.

Decision Rule

The p-value of this test will be $2Pr(Z > |z|)$, where z is the calculated test statistic. If the p-value from the z-test is less than the significance level (α), we will reject the null hypothesis, and if the p-value from the z-test is greater than the significance level (α), we fail to reject the null hypothesis and draw the appreciate conclusion.

Hypothesis Test Example: Complete Data Subset (V0)

Here we will give an example of the hypothesis test for the mean number of pauses. We will test if the mean differs in the higher score group and the lower score group.

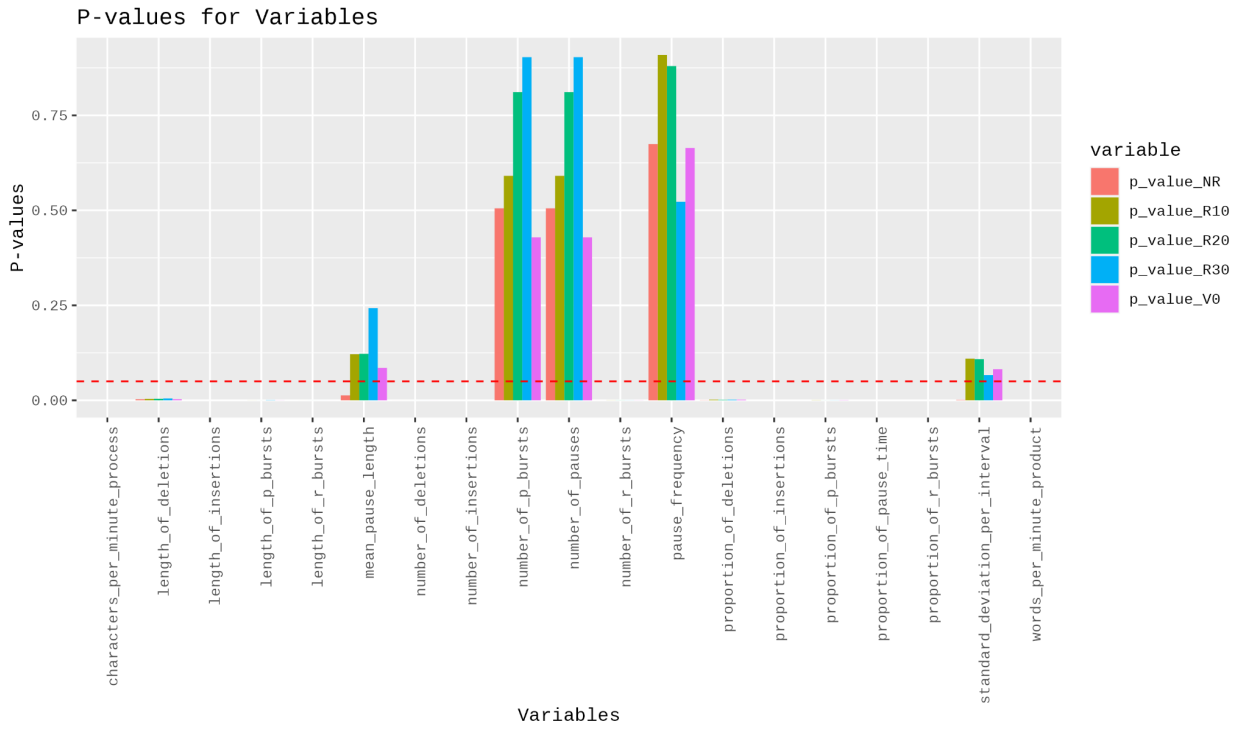
Calculation Example:

$\bar{X} = 85.52884615$, $\bar{Y} = 82.2708$
$s_1^2 = 26.98145955$, $s_2^2 = 30.90919780$
$n_1 = 104$, $n_2 = 96$
$Z = 0.791305$
p-value = $2Pr(Z > 0.791305) = 0.4287$

Conclusion Example:

Since the p-value is greater than α=0.05, we fail to reject the null hypothesis. We can conclude that there is no evidence that the number of pauses between individuals with a high score is different from the mean number of pauses between individuals with a low score in the V0 dataset.

These calculations were run for all variables in all the subsets, V0, R10, R20, R30, and NR. Below is the plot showing the p-value of the z-test for all the variables in each of the datasets. The horizontal red dashed line is the line for alpha=0.05, any p-value below this line is statistically significant and we can reject the null hypothesis.



Final Conclusion

Visually from the plot we can see that there is no significant difference between the means of the high and low scored groups for the variables number_of_p_bursts, number_of_pauses, and pause_frequency for any of the subsets. In addition, there is no significant difference between the means for the variables mean_pause_length and Standard_deviation, except for the NR subset. This may be due to the fact that the nonrandom missing data changes the variables mean_pause_length and Standard_deviation significantly for the low and high scoring group, so a significant difference arises. For all other variables not mentioned above, the means are significantly different between the two samples of the high score and low score groups for all datasets.

Hypothesis 2: Multiple Linear Regression

Introduction

Multiple Linear Regression (MLR) is a statistical technique that extends the principles of simple linear regression to model the relationship between a dependent variable and multiple independent variables. MLR provides a structured approach to disentangling the influences of various factors on a target variable. Within the confines of this project, our focus is centered on the computation of the impact of multiple independent variables on a dependent variable, represented as the "score." By employing MLR, we aim to dissect and quantify the individual and collective contributions of these variables to the observed scores.

Assumptions

1. **Independence**: The observations should be independent from one another. In other words, the dependent variable values for one observation should not be changed by the dependent variable values for other observations.
2. **Linearity**: There should be a linear relationship between the independent variables and the dependent variable. This means that a one-unit change in any of the independent variables results in a constant change in the dependent variable while holding the other variables constant.
3. **Homoscedasticity**: The error variance (residuals) should be constant across all levels of the independent variables. This assumption guarantees that the spread of the residuals remains constant across the range of independent variables.

Breusch-Pagan Test for Homoscedasticity Assumption

The Breusch-Pagan test is a statistical test used in econometrics to assess the presence of heteroscedasticity in a regression model. Heteroscedasticity refers to the situation where the variability of the errors (residuals) in a regression model is not constant across all levels of the independent variables. The formula for test-statistic is given as:

$LM = n \times R^2$

Here n is the size of the sample

R^2 : It refers to the coefficient of determination from the auxiliary regression of squared residuals on the independent variables.

Hypothesis Test for Breusch-Pagan Test

Null Hypothesis: H_0 : The variance of the errors (residuals) in the regression model is constant across all levels of the independent variables.

Alternate Hypothesis: H_a : The variance of the errors is not constant across all levels of the independent variables.

While the test explicitly focuses on heteroscedasticity, if you fail to reject the null hypothesis, it indirectly supports the assumption of homoscedasticity.

Durbin Watson Test for Independence among variables

The Durbin-Watson test is a statistical test used to detect the presence of autocorrelation in the residuals of a regression analysis. Autocorrelation occurs when the residuals are correlated with each other, indicating a lack of independence among the observations. The formula for test-statistic is given as:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Here, e_i is the i-th residual (the difference between the observed and predicted values) and n is the number of observations.

Hypothesis Test for Durbin Watson Test

Null Hypothesis: H_0 : The residuals exhibit no first-order autocorrelation, indicating independence among the observations.
Alternate Hypothesis: H_a :The residuals exhibit first-order autocorrelation, indicating no independence among the observations.

Multicollinearity

Multicollinearity refers to the condition in a multiple regression analysis where two or more predictor variables are highly correlated, making it challenging to distinguish their individual effects on the dependent variable. Checking for multicollinearity is important because it impacts precision of predictions, efficiency of the model, analysis, and violates assumptions of independence. So, to check multicollinearity we use the variation inflation factor.

Variation Inflation Factor

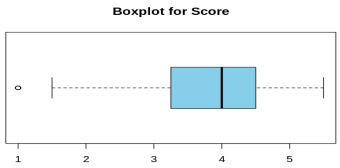
The Variance Inflation Factor (VIF) is a statistical metric used in multiple regression analysis to identify multicollinearity. The formula for VIF is

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Here, R_i^2 is the squared multiple correlation coefficient for the i-th variable. The higher the VIF, the higher the possibility that multicollinearity exists. Generally, a VIF value greater than 6 is assumed to be highly correlated.

Data Preparation and Outlier Removal

1. Based on the results of VIF, the variables with the least multicollinearity were the following: proportion of pause time, proportion of deletions, length of p bursts, proportion of r bursts, length of deletions, mean pause length, and standard deviation per interval. These were selected to perform MLR.
2. After plotting the boxplot for the target variable 'score,' it was observed that outliers existed below the first quartile (Q1). To address this, scores below 2 were considered outliers and were subsequently removed, resulting in a revised dataset with a minimum score of 2.



Significance Level (α)

The significance level (α) will be set to 0.05 for our tests.

Hypothesis

H_o: $\beta_i = 0$, the coefficient is not significant in the MLR model
H_1: $\beta_i \neq 0$, the coefficient is significant in the MLR model

The null hypothesis in a multiple linear regression model is that a predictor coefficient parameter is equal to zero, indicating that the corresponding predictor variable has no effect on the response variable. The alternative hypothesis is that the parameter is not equal to zero, suggesting that the predictor variable is significant in predicting the response variable. The necessary assumptions for multiple linear regression, such as linearity, independence, homoscedasticity have been tested and met and will be discussed later, now we continue on with our hypothesis.

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between a dependent variable and multiple independent variables. Unlike simple linear regression, which involves only one predictor, MLR considers the simultaneous influence of several factors on the outcome. In the context of multiple linear regression, t-tests are commonly used to assess the statistical significance of individual coefficients (slope parameters) associated with each independent variable. The t-test evaluates whether each variable has a significant impact on the dependent variable.

$$t_i = \frac{\widehat{\beta}_i - \beta_{0i}}{\sqrt{MSE / \sum_j (X_{ij} - \bar{X}_i))}}$$

Here it is a t-test since σ^2 is unknown so an approximation is required.
β_{0i} = 0 for all coefficients for this test.
MSE is the mean squared sum of errors, i.e. the mean square of residuals, of our MLR.
$\widehat{\beta}_i$ is the fitted β coefficient of the predictor.
The X_i variable in the equation corresponds to the data values in predictor i, with X_{ij} being the jth element of predictor i.

Formula and Calculation of Multiple Linear Regression

Multiple Linear Regression works as the extension of the OLS (Ordinary Least Squares) Regression because it involves more than one explanatory variable. The formula of MLR is given as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

Here,
Y_i = Dependent variable
X_i = Explanatory Variable
β_0 = y-intercept (constant term)
β_p = slope coefficients for each explanatory variable.
Ɛ = Models' error term

The estimates for β can be found with the following formula:

$$\begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

V0 Dataset Hypothesis Test Results

Variance Inflation Factor:

```
[1] "The VIF results on processed data"
        proportion_of_deletions        proportion_of_pause_time
                       5.004952                        3.381520
             length_of_p_bursts          proportion_of_r_bursts
                       2.230016                        3.595017
            length_of_deletions               mean_pause_length
                       3.776497                        2.446065
standard_deviation_per_interval
                       1.096313
```

Result: proportion_of_deletions is the only variable which is slightly highly correlated. The rest are low or moderately correlated.

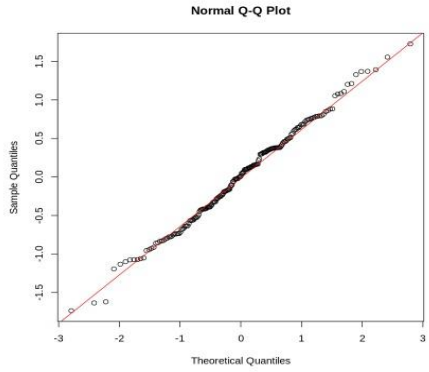
Multiple Linear Regression Model Summary

```
[1] "MLR on processed data"

Call:
lm(formula = score ~ proportion_of_deletions + proportion_of_pause_time +
    length_of_p_bursts + proportion_of_r_bursts + length_of_deletions +
    mean_pause_length + standard_deviation_per_interval, data = data_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73609 -0.43849  0.01115  0.40912  1.72692

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.494e+00  4.390e-01   5.680 5.21e-08 ***
proportion_of_deletions        -2.420e-01  8.281e-02  -2.922  0.00391 **
proportion_of_pause_time        1.257e-03  6.510e-03   0.193  0.84708
length_of_p_bursts              4.532e-04  1.005e-04   4.508 1.17e-05 ***
proportion_of_r_bursts          1.016e-01  1.090e-02   9.324  < 2e-16 ***
length_of_deletions            -2.787e-04  1.834e-04  -1.519  0.13037
mean_pause_length              -1.190e-06  7.711e-06  -0.154  0.87751
standard_deviation_per_interval -6.235e-05  3.214e-04  -0.194  0.84642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6646 on 183 degrees of freedom
Multiple R-squared:  0.5052,    Adjusted R-squared:  0.4863
F-statistic:  26.7 on 7 and 183 DF,  p-value: < 2.2e-16
```

Result: The adjusted R-square value is 0.486, which suggests a moderate level of explanatory power. About 48.6% provides a moderate level of explanation for the variance in the dependent variable. Here, the variables proportion_of_deletions, length_of_p_bursts, proportion_of_r_bursts, are statistically significant among all the variables.

Residual test on processed Data

```
Shapiro-Wilk test for normality of residuals:

        Shapiro-Wilk normality test

data:  residuals
W = 0.99438, p-value = 0.6908

The residuals appear to be normally distributed (fail to reject the null hypothesis).
```



Result: After performing the shapiro test for residuals, we found that the p-value is 0.09 which is greater than the significance level (0.05), so we fail to reject the null hypothesis. We can conclude that there is no evidence that residuals are not normally distributed.

R10 Dataset Hypothesis Test Results

Variance Inflation Factor

```
[1] "The VIF results on 10% random missing data"
        proportion_of_deletions        proportion_of_pause_time
                       4.846030                        3.459304
             length_of_p_bursts          proportion_of_r_bursts
                       1.876621                        3.653109
           length_of_deletions               mean_pause_length
                       3.341671                        2.462587
standard_deviation_per_interval
                       1.078895
```

Result: All the variables have VIF less than 5, which suggests they are low or moderately correlated.

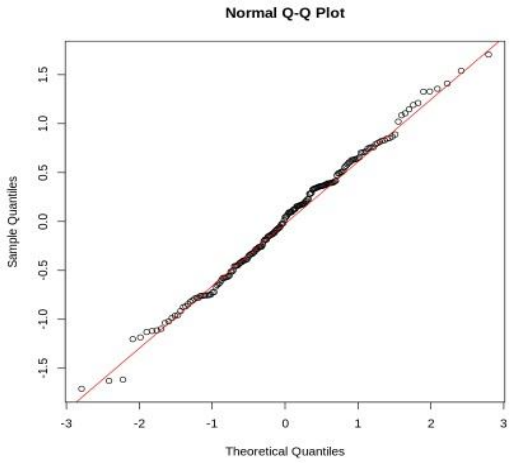
Multiple Linear Regression Model Summary

```
[1] "MLR on 10% random missing data"

Call:
lm(formula = score ~ proportion_of_deletions + proportion_of_pause_time +
    length_of_p_bursts + proportion_of_r_bursts + length_of_deletions +
    mean_pause_length + standard_deviation_per_interval, data = data_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70960 -0.45339  0.04491  0.40359  1.70460

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.597e+00  4.499e-01   5.771 3.31e-08 ***
proportion_of_deletions       -2.661e-01  9.031e-02  -2.947  0.00363 **
proportion_of_pause_time      -8.348e-04  6.641e-03  -0.126  0.90010
length_of_p_bursts             5.050e-04  1.036e-04   4.875 2.35e-06 ***
proportion_of_r_bursts         1.099e-01  1.230e-02   8.938 4.26e-16 ***
length_of_deletions           -2.697e-04  1.988e-04  -1.357  0.17652
mean_pause_length              1.246e-06  7.888e-06   0.158  0.87467
standard_deviation_per_interval -1.087e-04  3.501e-04  -0.310  0.75667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6654 on 183 degrees of freedom
Multiple R-squared:  0.504,     Adjusted R-squared:  0.4851
F-statistic: 26.57 on 7 and 183 DF,  p-value: < 2.2e-16
```

Result: The adjusted R-square value is 0.485, which suggests a moderate level of explanatory power. About 48.5% provides a moderate level of explanation for the variance in the dependent variable. Here, the variables proportion_of_deletions, length_of_p_bursts, proportion_of_r_bursts, are statistically significant among all the variables.

Normality of Residual test on 10% randomly missing Data

```
Shapiro-Wilk test for normality of residuals:

        Shapiro-Wilk normality test

data:  residuals
W = 0.99485, p-value = 0.7573

The residuals appear to be normally distributed (fail to reject the null hypothesis).
```



Result: After performing the shapiro test for residuals, we found that the p-value is 0.7573 which is much greater than the significance level (0.05), so we fail to reject the null hypothesis. We can conclude that there is no evidence that residuals are not normally distributed.

R20 Dataset Hypothesis Test Results

Variation Inflation Factor

```
[1] "The VIF results on 20% random missing data"
        proportion_of_deletions        proportion_of_pause_time
                       4.771866                        3.580620
            length_of_p_bursts          proportion_of_r_bursts
                       2.433649                        3.929884
          length_of_deletions               mean_pause_length
                       3.598684                        2.466418
standard_deviation_per_interval
                       1.109882
```

Result: All the variables have VIF less than 5, which suggests they are low or moderately correlated.

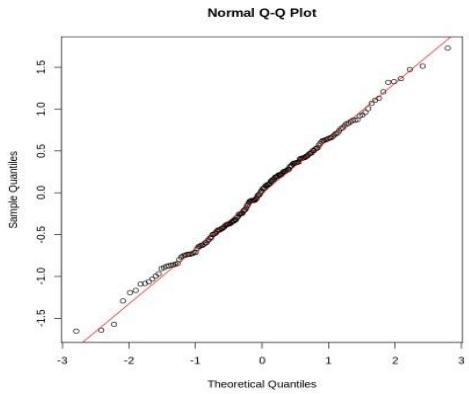
Multiple Linear Regression Model Summary

```
[1] "MLR on 20% random missing data"

Call:
lm(formula = score ~ proportion_of_deletions + proportion_of_pause_time +
    length_of_p_bursts + proportion_of_r_bursts + length_of_deletions +
    mean_pause_length + standard_deviation_per_interval, data = data_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65324 -0.45383  0.02085  0.43616  1.72867

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.601e+00  4.844e-01   5.370 2.37e-07 ***
proportion_of_deletions        -2.454e-01  9.960e-02  -2.464   0.0147 *
proportion_of_pause_time       -3.219e-03  6.884e-03  -0.468   0.6406
length_of_p_bursts              6.387e-04  1.230e-04   5.191 5.53e-07 ***
proportion_of_r_bursts          1.192e-01  1.430e-02   8.337 1.78e-14 ***
length_of_deletions            -4.322e-04  2.019e-04  -2.140   0.0336 *
mean_pause_length               6.627e-06  8.298e-06   0.799   0.4255
standard_deviation_per_interval -9.711e-05  3.933e-04  -0.247   0.8052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6618 on 183 degrees of freedom
Multiple R-squared:  0.5093,    Adjusted R-squared:  0.4906
F-statistic: 27.14 on 7 and 183 DF,  p-value: < 2.2e-16
```

Result: An adjusted R-squared of 0.4906 suggests a moderate level of explanatory power. About 49.06% of the variance in the dependent variable is captured by the independent variables. Here from the p-value, we can say that the variables proportion_of_deletions, length_of_p_bursts, proportion_of_r_bursts, length_of_deletions are statistically significant.

Normality of Residual test on 20% randomly missing Data

```
Shapiro-Wilk test for normality of residuals:

        Shapiro-Wilk normality test

data:  residuals
W = 0.99543, p-value = 0.8344

The residuals appear to be normally distributed (fail to reject the null hypothesis).
```



Result: After performing the shapiro test for residuals, we found that the p-value is 0.8344 which is much greater than the significance level (0.05), so we fail to reject the null hypothesis. We can conclude that there is no evidence that residuals are not normally distributed.

R30 Dataset Hypothesis Test Results

Variation Inflation Factor

```
[1] "The VIF results on 30% random missing data"
        proportion_of_deletions        proportion_of_pause_time
                       4.970792                        3.723215
             length_of_p_bursts          proportion_of_r_bursts
                       3.652977                        3.770128
           length_of_deletions               mean_pause_length
                       5.175580                        2.101819
standard_deviation_per_interval
                       1.202026
```

Result: All the variables have VIF less than 5, which suggests they are low or moderately correlated.

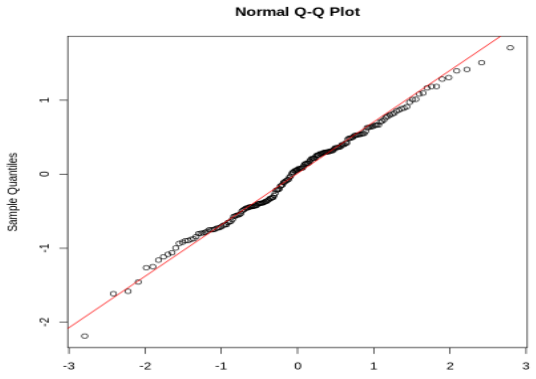
Multiple Linear Regression Model Summary

```
[1] "MLR on 30% random missing data"

Call:
lm(formula = score ~ proportion_of_deletions + proportion_of_pause_time +
    length_of_p_bursts + proportion_of_r_bursts + length_of_deletions +
    mean_pause_length + standard_deviation_per_interval, data = data_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-2.18662 -0.45719  0.06359  0.48064  1.70853

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.702e+00  5.449e-01   4.959 1.61e-06 ***
proportion_of_deletions        -1.172e-01  1.204e-01  -0.974  0.33133
proportion_of_pause_time       -4.347e-03  7.099e-03  -0.612  0.54110
length_of_p_bursts              7.386e-04  1.449e-04   5.098 8.54e-07 ***
proportion_of_r_bursts          1.161e-01  1.585e-02   7.323 7.44e-12 ***
length_of_deletions            -8.252e-04  2.867e-04  -2.878  0.00448 **
mean_pause_length               8.927e-06  7.131e-06   1.252  0.21217
standard_deviation_per_interval -2.082e-04  4.648e-04  -0.448  0.65468
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6833 on 183 degrees of freedom
Multiple R-squared:  0.4769,    Adjusted R-squared:  0.4569
F-statistic: 23.83 on 7 and 183 DF,  p-value: < 2.2e-16
```

Result: An adjusted R-squared of 0.4566 suggests a moderate level of explanatory power. About 45.69% of the variance in the dependent variable is captured by the independent variables. Here from the p-value, we can say that the variables length_of_deletions, length_of_p_bursts, proportion_of_r_bursts, length_of_deletions are statistically significant.

Normality of Residual test on 30% randomly missing Data

```
Shapiro-Wilk test for normality of residuals:

        Shapiro-Wilk normality test

data:  residuals
W = 0.99344, p-value = 0.5569

The residuals appear to be normally distributed (fail to reject the null hypothesis).
```



Result: After performing the shapiro test for residuals, we found that the p-value is 0.5569 which is much greater than the significance level (0.05), so we fail to reject the null hypothesis. We can conclude that there is no evidence that residuals are not normally distributed.

NR Dataset Hypothesis Test Results

Variation Inflation Factor

```
[1] "The VIF results on non random missing data"
        proportion_of_deletions        proportion_of_pause_time
                       4.417308                        4.135216
             length_of_p_bursts          proportion_of_r_bursts
                       1.615725                        5.555985
           length_of_deletions               mean_pause_length
                       3.104668                        2.645164
standard_deviation_per_interval
                       2.329514
```

Result: All the variables have VIF less than 5, except proportion_of_r_bursts which suggests they are slightly highly correlated.

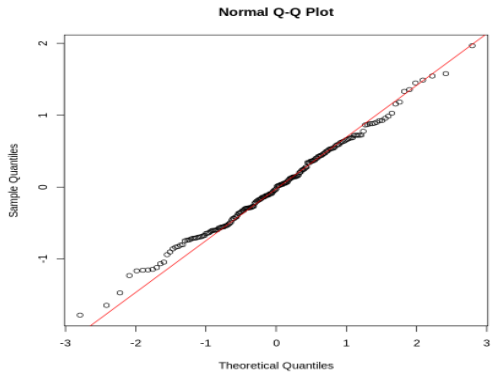
Multiple Linear Regression Model Summary

```
[1] "MLR on non random missing data"

Call:
lm(formula = score ~ proportion_of_deletions + proportion_of_pause_time +
    length_of_p_bursts + proportion_of_r_bursts + length_of_deletions +
    mean_pause_length + standard_deviation_per_interval, data = data_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77949 -0.50992 -0.02332  0.46113  1.96676

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.187e+00  4.804e-01   4.554 9.60e-06 ***
proportion_of_deletions       -8.547e-02  7.248e-02  -1.179   0.2399
proportion_of_pause_time      -1.807e-04  7.183e-03  -0.025   0.9800
length_of_p_bursts             7.532e-04  9.655e-05   7.801 4.54e-13 ***
proportion_of_r_bursts         6.035e-02  1.156e-02   5.219 4.85e-07 ***
length_of_deletions           -2.053e-04  1.672e-04  -1.228   0.2211
mean_pause_length             -2.178e-06  8.946e-06  -0.243   0.8079
standard_deviation_per_interval 3.848e-03  1.907e-03   2.018   0.0451 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.668 on 183 degrees of freedom
Multiple R-squared:    0.5,     Adjusted R-squared:  0.4809
F-statistic: 26.15 on 7 and 183 DF,  p-value: < 2.2e-16
```

Result: An adjusted R-squared of 0.4809 suggests a moderate level of explanatory power. About 48.09% of the variance in the dependent variable is captured by the independent variables. Here from the p-value, we can say that the variables length_of_p_bursts, proportion_of_r_bursts, standard_deviation_per_interval are statistically significant.

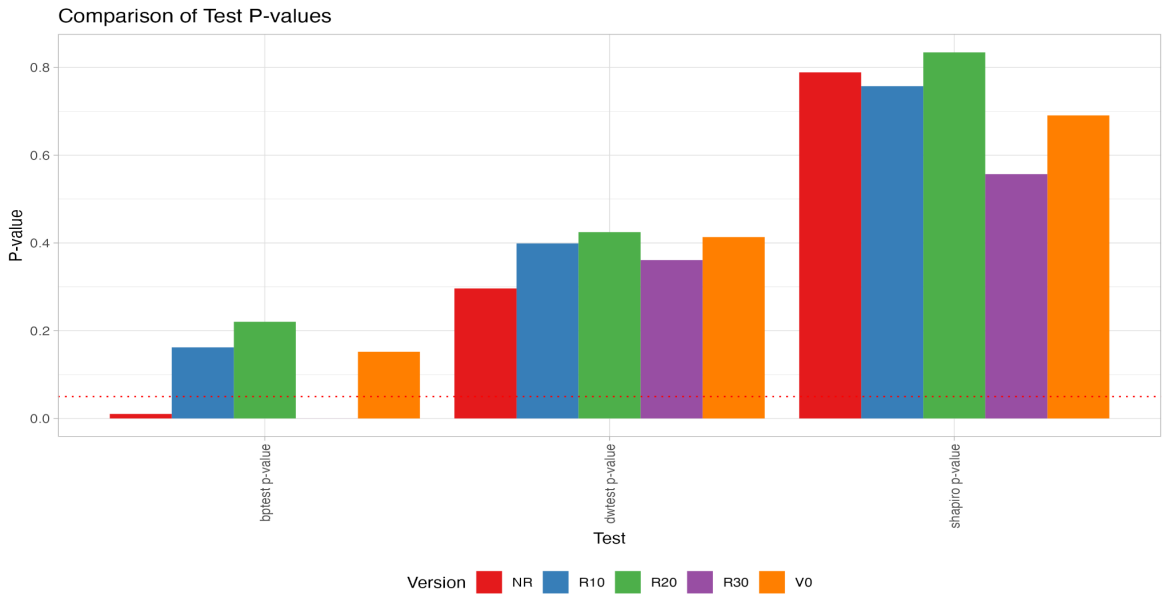
Normality of Residual test on Non random missing Data

```
Shapiro-Wilk test for normality of residuals:

        Shapiro-Wilk normality test

data:  residuals
W = 0.99508, p-value = 0.7887

The residuals appear to be normally distributed (fail to reject the null hypothesis).
```



Result: After performing the shapiro test for residuals, we found that the p-value is 0.7887 which is greater than the significance level(0.05). So, we can conclude that the residuals are normally distributed. Hence, we fail to reject the null hypothesis.

Assumptions Conclusion



Conclusion on Breusch-Pagan Test

The test for homoscedasticity has failed for non-random missing data and 30% randomly missing dataset, it can be because of the impact of the missing data on the dataset. Varied imputation methods and non-random missingness can affect residuals, impacting the assumption of constant variability. The rest of the versions of data, that is, complete data(V0), 10% randomly missing data, 20% missing data pass the test for homoscedasticity. The versions for non random missing data, and also 30% randomly missing data do not pass the test for homoscedasticity.

Conclusion on Durbin Watson Test

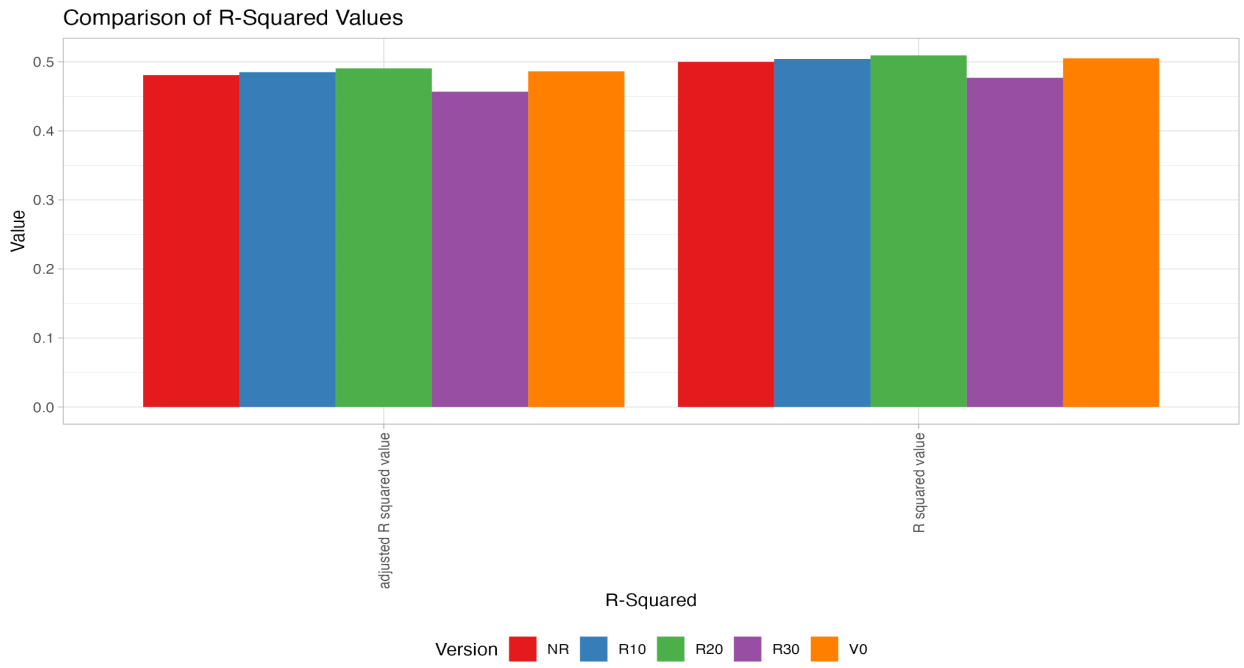
The Durbin Watson Test has failed to reject the null hypothesis stating that the data is independent with respect to the dependent variable. So we can conclude that our data is independent in all 5 different types of datasets, that is, complete data(V0), 10% random missing data(R10), 20% random missing data(R20), 30% random missing data(R30), Non-random missing data.

Conclusion on Shapiro test for residuals to check normality

The results of the Shapiro-Wilk test suggest that the residuals of the regression model can be reasonably assumed to follow a normal distribution. The null hypothesis of the Shapiro-Wilk test is that the data is normally distributed. Since all the p-values are greater than 0.05 (horizontal dashed line), we fail to reject the null hypothesis and conclude that there is not significant evidence that the residuals of all our models with the different subsets are not normal. This

means that all the assumptions for MLR are met for all our cases. Surprisingly, the dataset with 10% and 20% missing data and non random missing data had a higher p-value than the dataset with no missing data. This suggests that these residuals are more normal, which is surprising since we expect a better fit with more data, but the converse is true for these cases. This may be due to random chance, since the p-values overall do not differ that much.This finding enhances our confidence in the reliability of statistical inferences drawn from the model.

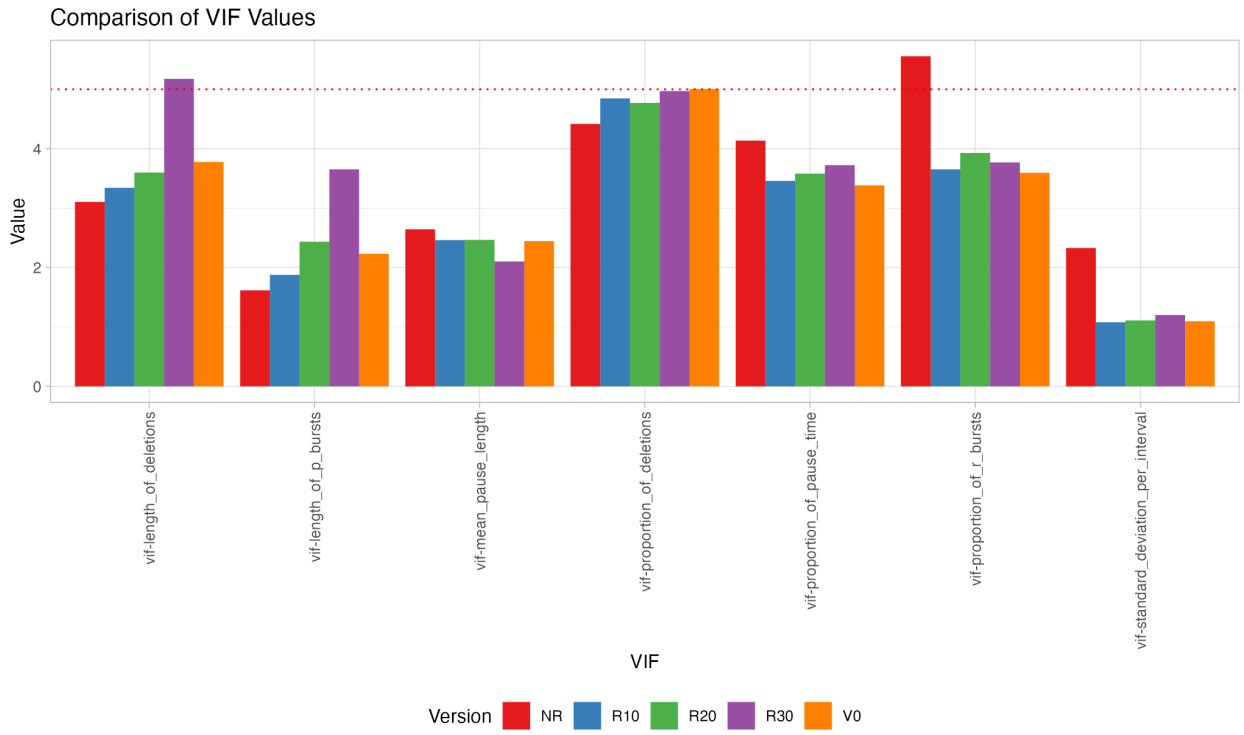
Conclusion on R-Squared Values



The R-squared values across all five versions of the dataset hover consistently between 0.48 and 0.52, indicating marginal differences. Specifically, for the non-random version, the R-squared is approximately 0.5. Versions with 10%, 20%, and 30% random missing data yield R-squared values of 0.51, 0.52, and 0.48, respectively. Version V0 records an R-squared of 0.51. Interestingly, Version V0 records an R-squared of 0.51. Notably, the R-squared values for 10% missing data, 20% missing data, and V0 are nearly identical, hinting at the impact of missing data on higher percentages.

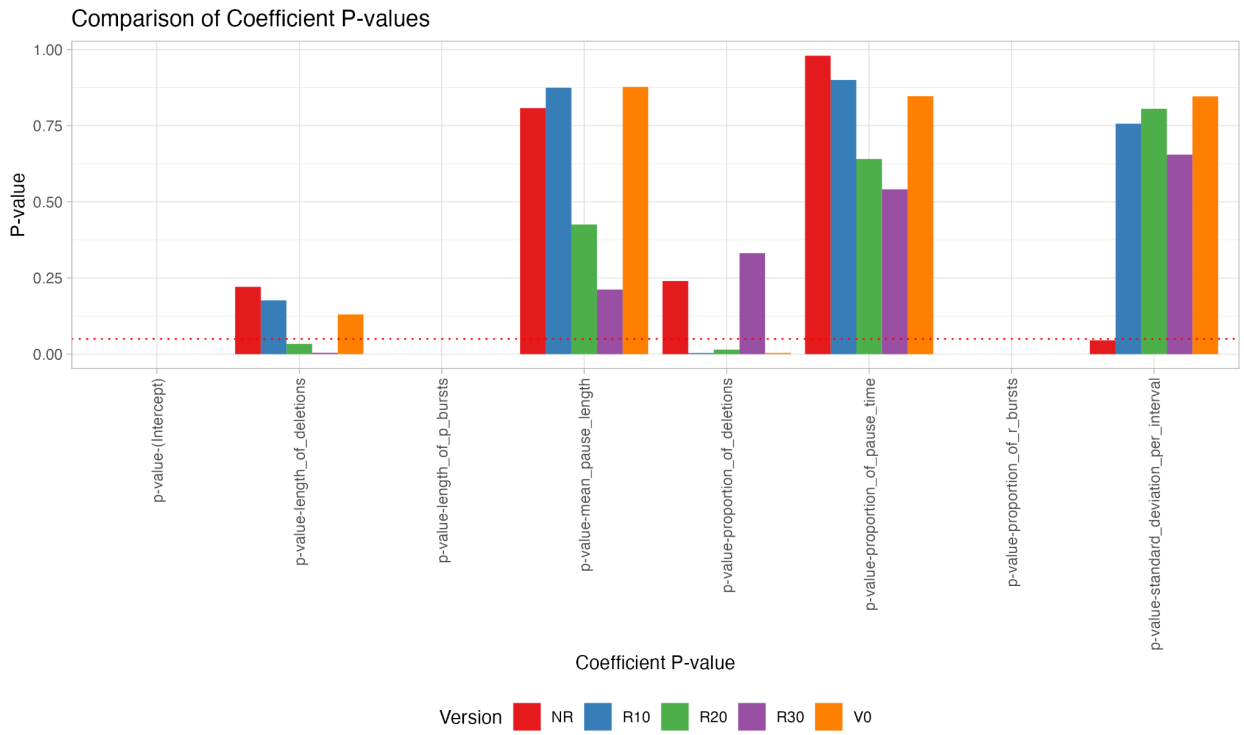
For the adjusted R-squared, it's the same order as R-squared. The adjusted R-squared values demonstrate consistency across different dataset versions. For V0, the adjusted R-squared is 0.486, while for 10% missing data, it is 0.485, and for 20% missing data, it slightly improves to 0.4906. However, the version with 30% missing data shows a decrease to 0.4569, and the non-random version is at 0.4809. Despite subtle variations, the adjusted R-squared values suggest a degree of uniformity in explanatory power among the dataset versions. Notably, the decrease in adjusted R-squared for 30% missing data implies a potential impact on model fit with higher levels of missing data.

Conclusion on VIF values



The majority of the variation inflation factor (VIF) values imply independence among variables. In the case of 30% random missing data, the length of deletions exhibits a slightly elevated correlation (VIF value = 5.009), suggesting potential influence from the missing data on this variable. Similarly, for the non-random missing data, the proportion_of_r_bursts shows a modestly elevated correlation (VIF value = 5.55), The observed deviations in the correlation values (VIF) for the length of deletions (in the case of 30% random missing data) and the proportion_of_r_bursts (for non-random missing data) strongly suggest that these variables have been influenced or impacted by the patterns of missing data.

Conclusion on Coefficient P-values



1. The null hypothesis indicates that the corresponding independent variable has no statistically significant effect on the dependent variable. The alternate hypothesis indicates that there is a statistically significant relationship between the independent variable and the dependent variable.

2. The p-values associated with the coefficient for the variable 'length_of_deletions' in the MLR models for 20% randomly missing data (R20) and 30% randomly missing data (R30) are both below the significance level (0.05), indicating a statistically significant relationship. This suggests that despite the missing data 'length_of_deletions' has a noteworthy impact on the dependent variable in both scenarios. The findings emphasize the robustness of this relationship, even in the presence of missing data at varying levels.

3. Interestingly, in all the 5 versions of data, the variable length_of_p_bursts has a significant relationship with the dependent variable score. This shows that the variable is highly important for our dataset. The continuing significance of this variable despite differences in dataset versions highlights its dependability and significance in determining the outcome, offering useful insights for our analysis.

4. The coefficient values attached with mean_pause_length have all failed to reject the null hypothesis, suggesting that this predictor does not have any significance with the predicting the score.

5. The coefficient values of proportion_of_deletions in the 30% randomly missing data is not statistically significant, but in the rest of the subsets it is statistically significant. This may suggest that when more data is missing the predictor proportion_of_deletions becomes insignificant, perhaps due to variation from increased randomness.

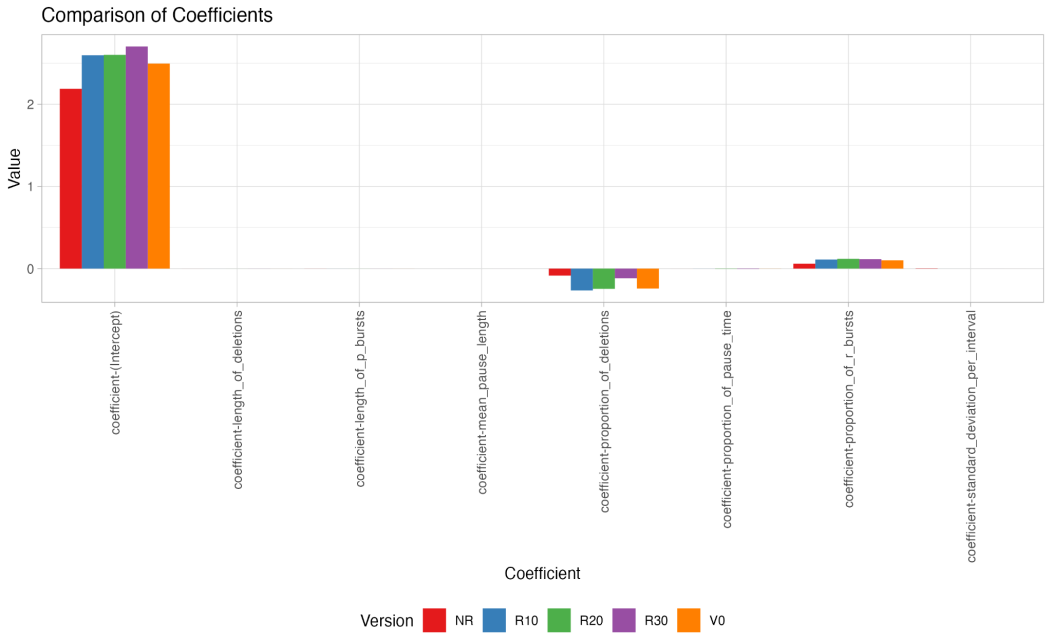
6. The coefficient values attached with proportion_of_pause_time have all failed to reject the null hypothesis, suggesting that this predictor does not have any significance with predicting the score.

7. It is noteworthy that the variable proportion_of_r_bursts rejects the null hypothesis across all five datasets. This consistent rejection indicates a statistically significant relationship with the dependent variable. The constant significance of proportion_of_r_bursts emphasizes its crucial role and impact for score prediction.

8. Across dataset versions, the variable standard_deviation_per_interval shows varied contributions to the dependent variable. Surprisingly, with the exception of the non-random missing data version, the variable has no significant effect on the dependent variable in the other versions of the dataset. This disparity reveals possible complexities in the relationship between standard_deviation_per_interval and the dependent variable, stressing the need of taking dataset-specific dynamics into account when analyzing variable impact.

9. In conclusion, our thorough examination of the dataset reveals complex relationships in the interactions between variables and the dependent variable across different versions. Notably, length_of_p_bursts, and proportion_of_r_bursts is a consistent contributor, highlighting its critical importance. standard_deviation_per_interval, on the other hand, has a variable impact, with significance detected only in the non-random missing data variant. Moreover, variables like proportion_of_pause_time, mean_pause_length have no significance on the output data. Also, proportion_of_deletions, length_of_deletions have varied impact depending on the type of dataset versions. These findings highlight the importance of a comprehensive understanding of variable connections that takes into consideration the context of the dataset versions. Such insights improve the precision and reliability of our regression models, allowing for a more accurate assessment of the dataset dynamics.

Conclusion on Coefficients



The plot above displays the coefficients associated with each variable/predictor in the regression model. These coefficients act as weights, indicating the magnitude and direction of the relationship between each independent variable and the dependent variable. In the context of the plot, the negative fitted values for "proportion_of_deletions" suggest a negative relationship with the dependent variable, the score. This implies that an increase in the "proportion_of_deletions" is associated with a decrease in the score. Similarly, the proportion_of_r_bursts coefficients suggest a positive relationship indicating increase in the variable also increases the predicted value in the variable. The rest of the variables have a coefficient value extremely close to 0. An important note, this graph does not say anything about the statistical significance of each coefficient, but rather the value of that coefficient in relation to other coefficients.

Summary

We generated 5 datasets: Complete data, 10% randomly missing data, 20% randomly missing data, 30% randomly missing data, and non-random missing values. The first hypothesis, which was multiple Z-tests, showed no significant differences in the mean of the variables number_of_p_bursts, number_of_pauses, and pause_frequency between high and low scored groups. Exceptions are observed in mean_pause_length and Standard_deviation for the NR subset, possibly due to complications arising from non-random missing data. Surprisingly, many variables do not exhibit statistically significant differences in the means, which suggests that not every variable may be useful for fitting in order to make a prediction of the score. For the second hypothesis, we choose to do a MLR, running multiple T-tests. The null hypothesis posits no statistically significant relationship, while the alternative suggests a significant relationship

between the independent variable score and a dependent variable. 'Length_of_deletions' exhibits consistent significance in regression models, even with varying degrees of missing data. 'Length_of_p_bursts' consistently proves significant across all dataset versions, indicating its robust importance. Conversely, 'Mean_pause_length' lacks statistical significance in all scenarios, suggesting it may not be significant in our model and can be removed. The impact of 'Proportion_of_deletions' varies with the degree of missing data, becoming insignificant with higher levels. 'Proportion_of_pause_time' lacks statistical significance across the board. 'Proportion_of_r_bursts' consistently shows significance across all dataset versions. 'Standard_deviation_per_interval' demonstrates varied impact, proving significant only in the non-random missing data variant. Complex relationships emerge, with 'Length_of_p_bursts' and 'Proportion_of_r_bursts' consistently crucial, while 'Standard_deviation_per_interval,' 'Proportion_of_pause_time,' and 'Mean_pause_length' lacking significance. The impact of 'Proportion_of_deletions' and 'Length_of_deletions' varies with dataset versions, further testing may be required. Assumptions were also shown to often hold for all of the data subsets. The datasets V0, R10, and R20 have extremely similar outcomes and conclusions, but for R30 and NR, the results have changed significantly. This is most likely due to higher randomness.

.