



Linking Writing Process to Writing Quality

Predicting an essay's quality by typing behavior

AMS 572 Final Presentation

Group Members: Lihan Huang, Saransh
Surana and Bernard Tenreiro



Overview

01

Introduction

The goal of this project.

02

Data

A brief explanation of the data preparation.

03

Hypothesis 1

Is there a difference in the typing behavior between higher and lower scoring groups?

04

Hypothesis 2

Can we fit a linear model to predict score?

05

Summary

A summary of our results and final conclusions.

Kaggle: Linking Writing Processes to Writing Quality

Goal: study the relationship between writing behaviors and writing performance.

- Performance: score of the essay
- Behaviors: keystroke data recorded during the argumentative writing tasks

Collection interface (Kaggle): Participants saw the image on the right.

More information found [here](#).

Prompt

While some people promote competition as the only way to achieve success, others emphasize the power of cooperation. Intense rivalry at work or play or engaging in competition involving ideas or skills may indeed drive people either to avoid failure or to achieve important victories. In a complex world, however, cooperation is much more likely to produce significant, lasting accomplishments.

Do people achieve more success by cooperation or by competition?

- Write independently for 30 minutes.
- Write at least 200 words.
- Write at least three paragraphs.
- Do not leave this page while writing.

Caution: Bonus (\$11.75) will not be paid if plagiarism is found in your essay or your essay does not address the prompt question.

I believe that


Word Count: 3

Submit


Keystroke: Raw Data

Raw data:

- 2471 essays in train_logs.csv & train_scores.csv
- Each has sequential events recorded, example: see image on right (from Kaggle)



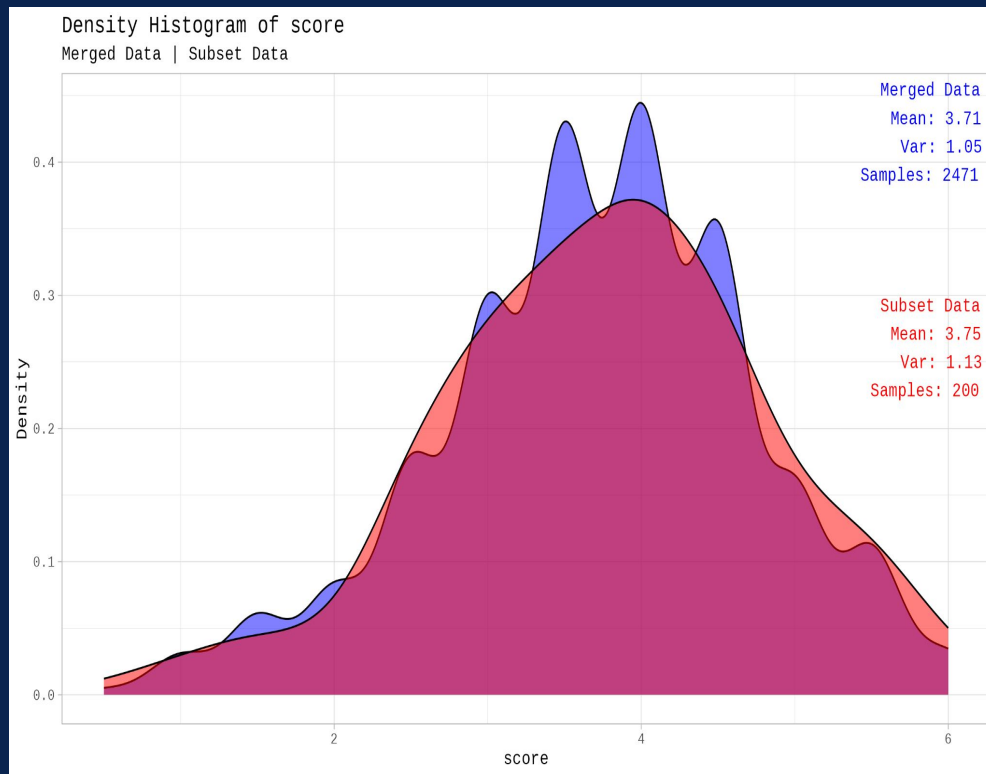
Event ID	Down Time	Up Time	Action Time	Event	Position	Word Count	Text Change	Activity
1	30185	30395	210	Leftclick	0	0	NoChange	Nonproduction
2	41006	41006	0	Shift	0	0	NoChange	Nonproduction
3	41264	41376	112	I	1	1	I	Input
4	41556	41646	90	Space	2	1		Input
5	41815	41893	78	b	3	2	b	Input
6	42018	42096	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



Keystroke: Subset Selection

Subset selection: A sample of 200 was taken from the population of 2471 for fast downstream experiments

The sample is shown to be a good representation of the population.



Keystroke: Variables and Measures

Variables can be computed from recording:

- Measures about production rate
- Measures about pause
- Measures about revision
- Measures about burst
- Measures about process variance

Each has multiple measures, with detailed definitions on [Kaggle](#). We computed 19 measures and score to have a total of 20 variables.

Most variables are continuous, but we use score differently in the two hypotheses:

- In hypothesis 1, we use score as the criteria to split data into 2 groups (categorical)
- In hypothesis 2, we use score as the regression target (continuous)



Missing Data Creation and Handling

Original subset will be referred to as the dataset V0

Randomly missing data: drop events in a recording at different percentage levels

- Dataset R10: randomly drop 10%
- Dataset R20: randomly drop 20%
- Dataset R30: randomly drop 30%

Non-random (dataset NR): drop events with activity nonproduction, replace & paste

- Assume these kinds of events are not recorded during collection because of logging issues
- Handle missing data: ignore the missing event, only compute variables based on events recorded



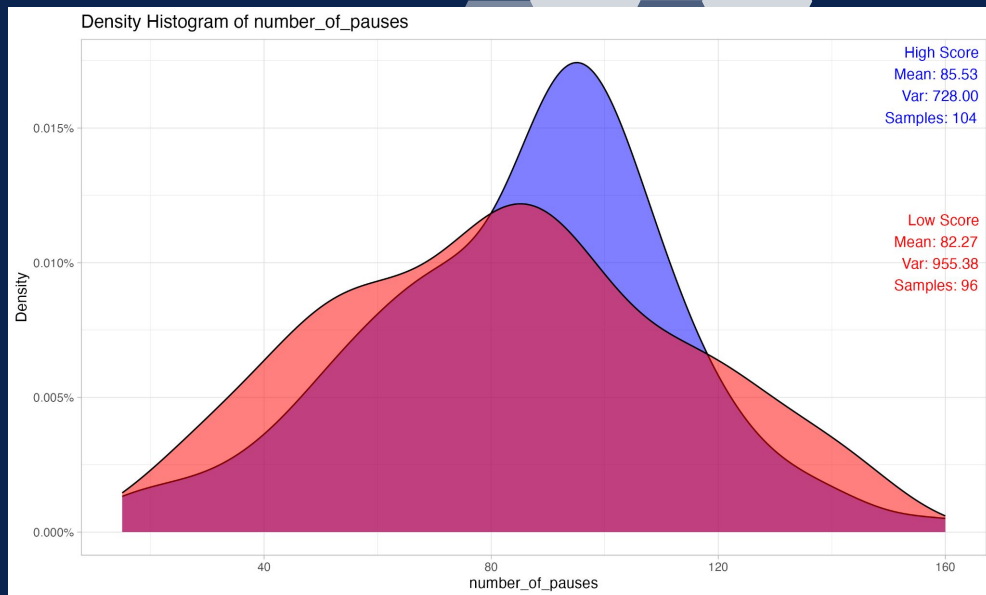
Hypothesis 1: Difference in Behaviors Between Groups (High vs Low Score)

Score will be the threshold for grouping:

- Mean: 3.79
- High score: ≥ 4
- Low score: < 4

EDA example:

- Data: V0
- Variable: Number of pauses
- High score: 104 samples
- Low score: 96 samples
- Testing means between samples
- Z-test since both samples large



Hypothesis 1: Continued

H_0 : Sample means are the same in both groups: $\mu_1 = \mu_2$

H_1 : Sample means are different in both groups: $\mu_1 \neq \mu_2$

5 versions of data (each subset V0, R10, R20, R30 and NR)

- 19 variables considered
- Dotted red line is $\alpha=0.05$

$H_0 : \mu_1 = \mu_2$ (Null Hypothesis)

$H_1 : \mu_1 \neq \mu_2$ (Alternative Hypothesis)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where

\bar{x}_1 : Sample mean of the first group

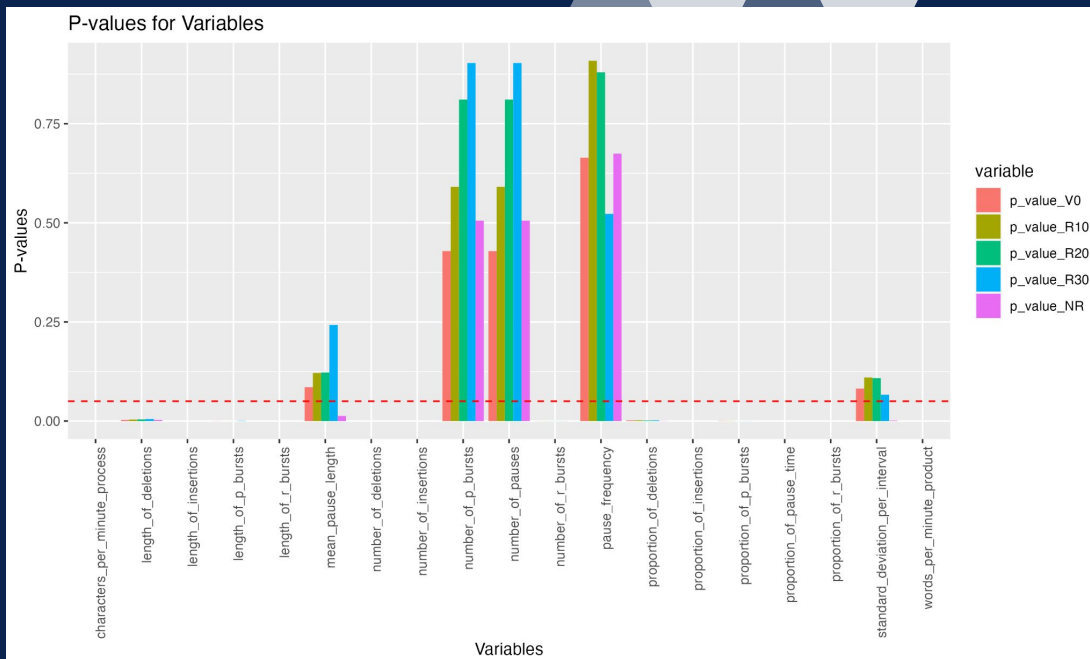
\bar{x}_2 : Sample mean of the second group

s_1 : Sample standard deviation of the first group

s_2 : Sample standard deviation of the second group

n_1 : Sample size of the first group

n_2 : Sample size of the second group



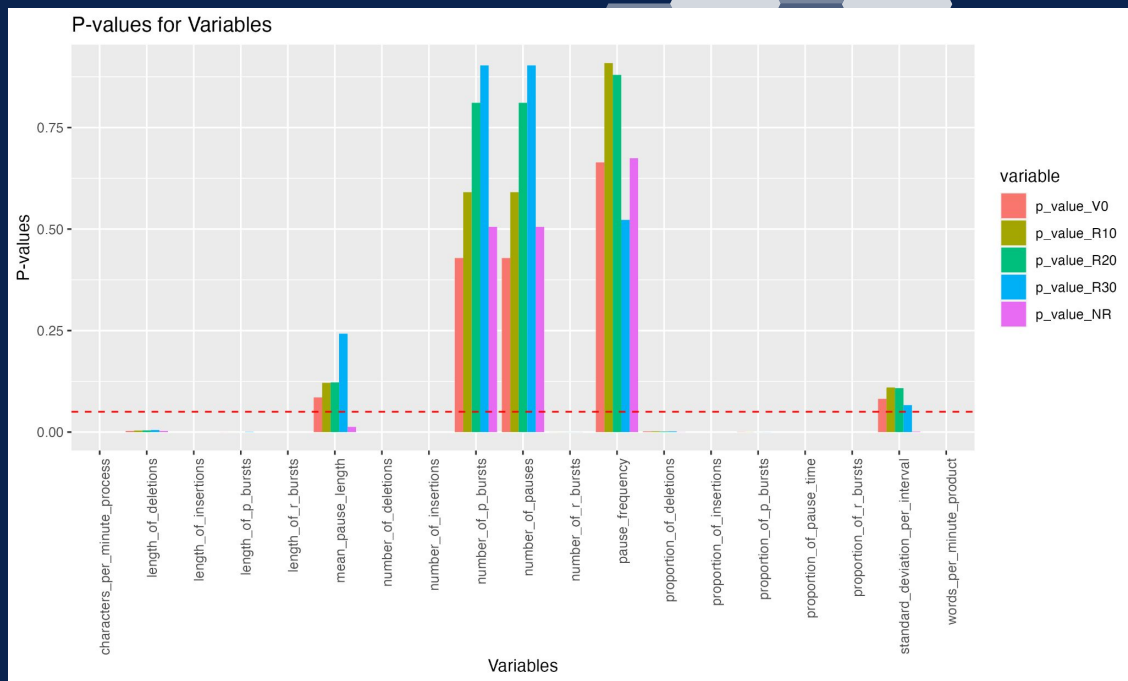
Hypothesis 1: Conclusion

Conclusions:

No significant difference for:

- mean_pause_length (Except NR)
- number_of_p_bursts
- number_of_pauses
- Pause_frequency
- Standard_deviation (Except NR)

Important to note: No assumptions are necessary since sample size is large (CLT and Slutsky)



Hypothesis 2: Multiple Linear Regression to Predict Score

Subset (columns) used for Multiple Linear Regression (MLR):

- 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

Outlier removal: remove score < 2 , sample of 191 remaining

- This gives much more significant results

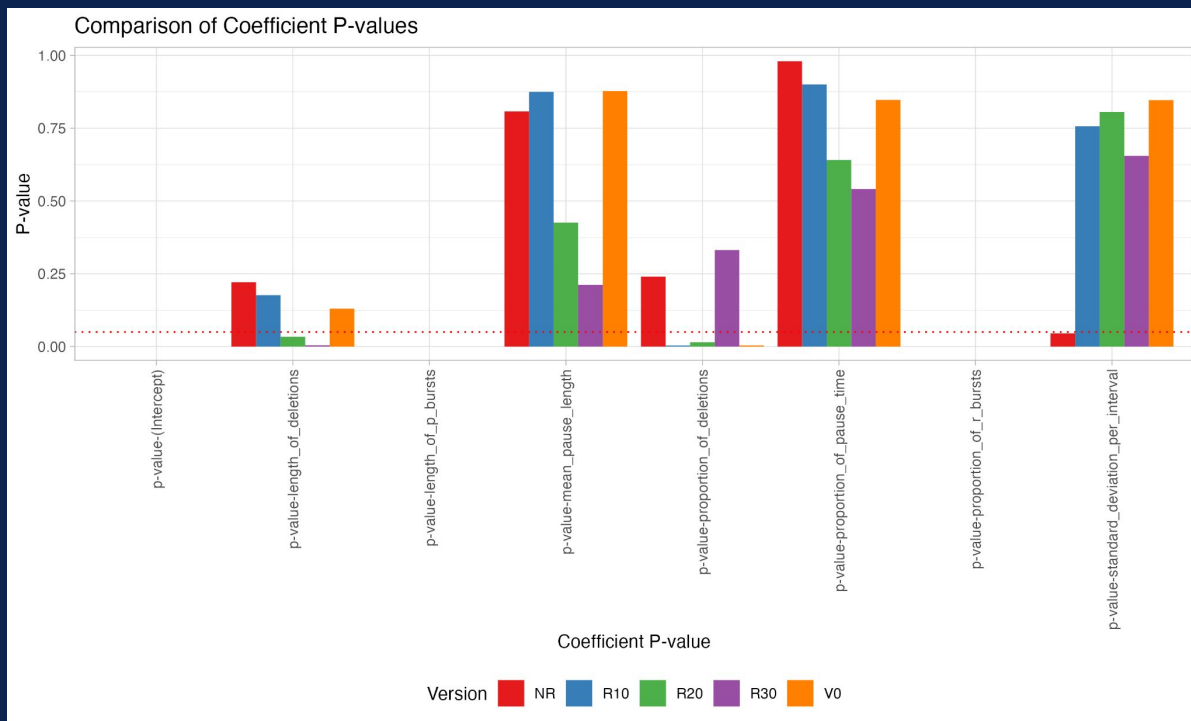


Hypothesis 2: Significance of Coefficients

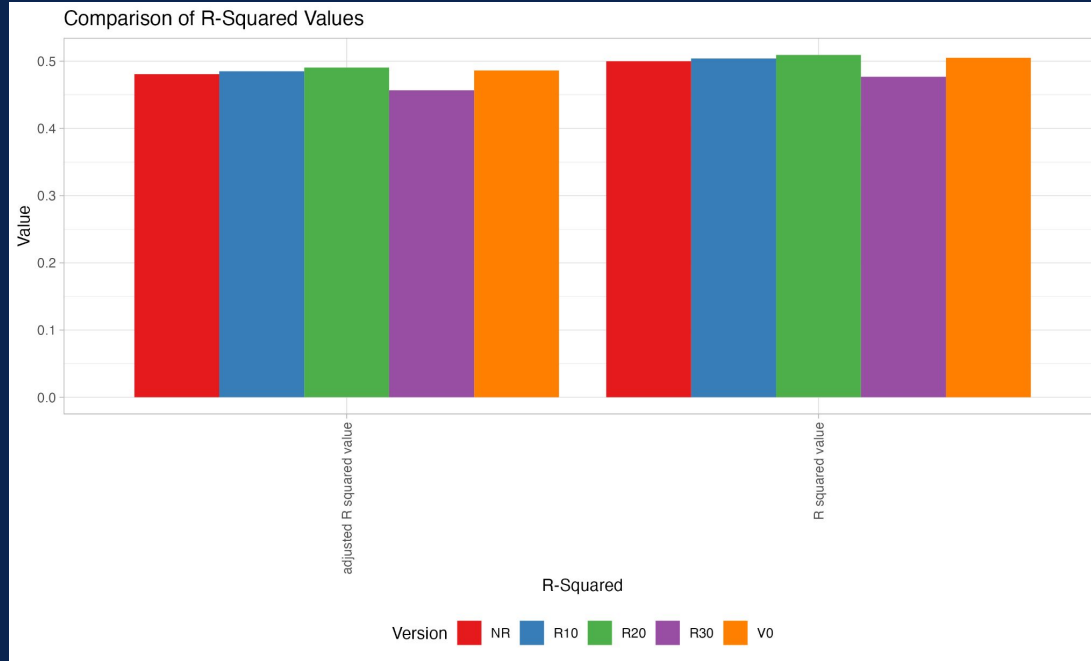
$H_0: \beta_i = 0$, the coefficient is not significant in the MLR model

$H_1: \beta_i \neq 0$, the coefficient is significant in the MLR model

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



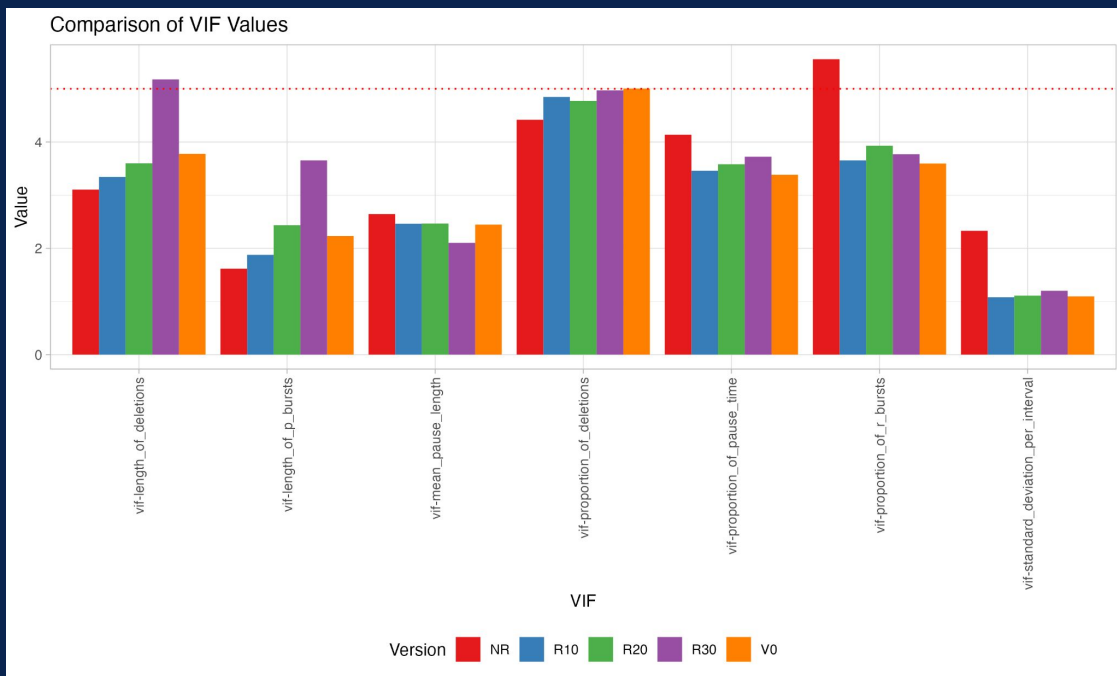
MLR: Goodness of Fit



MLR: Assumptions Check

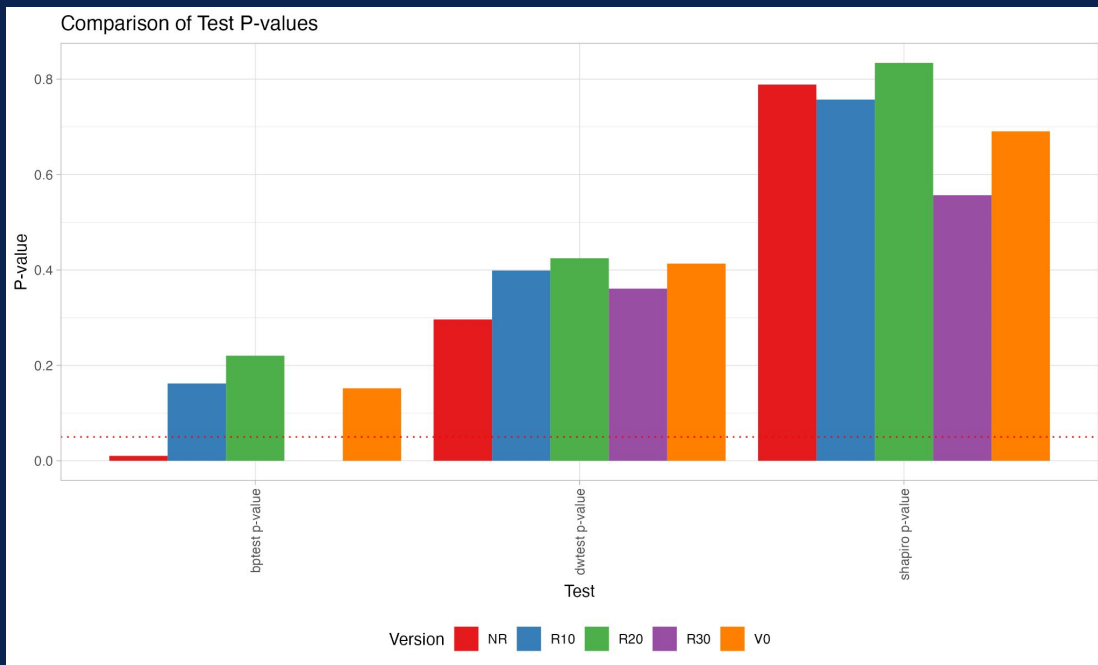
Testing for multicollinearity: VIF Criteria

- Vertical line is $VIF = 5$, below or near this line we can conclude there is no multicollinearity



MLR: Assumptions Check

Testing residual normality, homoscedasticity and no autocorrelation



MLR: Assumption Conclusions

- VIF: Although $VIF > 5$ for in two cases, it is close enough to say there is low multicollinearity
- Shapiro Test: Test for normality was successful, we can conclude there is no evidence that residuals are not normal
- Breusch-Pagan Test: Test for homoscedasticity has failed for non-random missing data and 30% randomly missing data
- Durbin Watson Test: We can conclude that our data is independent in all 5 different datasets

Our assumptions hold for a majority of the cases. When they fail it is commonly in a dataset with high randomness (R30 or NV), so that can be an explanation for a failed assumption.



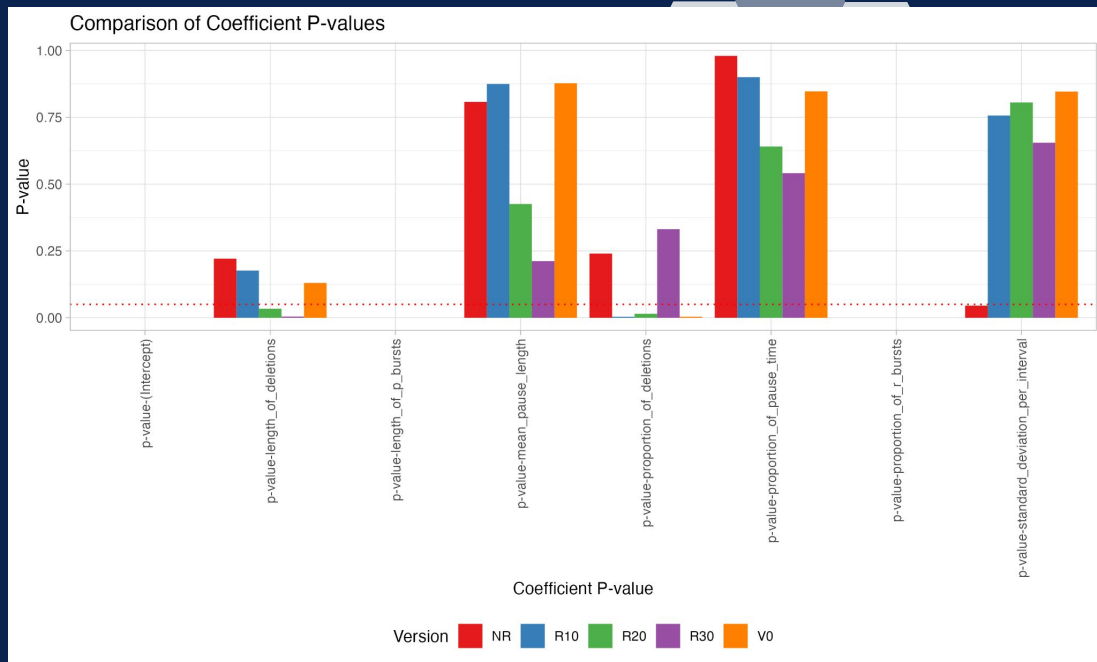
Hypothesis 2: Conclusions

Our assumptions are held for most cases, so we can continue with our hypothesis testing.

Intercept, length of p bursts, and proportion of r bursts is always significantly different from 0, thus significant in the models

Mean pause length and proportion of pause time are never significantly different from 0, thus not significant in the models

Other variables have varying p-values in each subset, more testing necessary



SUMMARY

Hypothesis 1:

- Conducted Z-tests to find that there is no significant mean differences between high and low-scored groups for variables number_of_p_bursts, number_of_pauses, and pause_frequency
- Mean_pause_length and Standard_deviation show differences in the NR subset, likely due to these values vastly changing.
- All other variables show a significant difference in all subsets, a subset of these variables were used for MLR to predict score

Hypothesis 2:

- Ran a MLR, conducted a T-test for predictor coefficients
- Intercept, length of p bursts, and proportion of r bursts are always significantly different from 0, thus significant in the MLR model
- Mean pause length and proportion of pause time are never significantly different from 0, thus not significant in the MLR model
- Other predictors need more testing to make further overall conclusions

Thank you