Statistical Analysis of Reaction Time and its Potential Significant Predictors Based on Survey Data

Group Members:

Ziqi Xu

- Contributions: Exploratory data pre-processing and analysis; model building

Zheer Wang

- Contributions: Exploratory data pre-processing and analysis; model building

Mingqian Wang

- Contributions: R code review; drafting and editing final report; data pre-processing

Abstract:

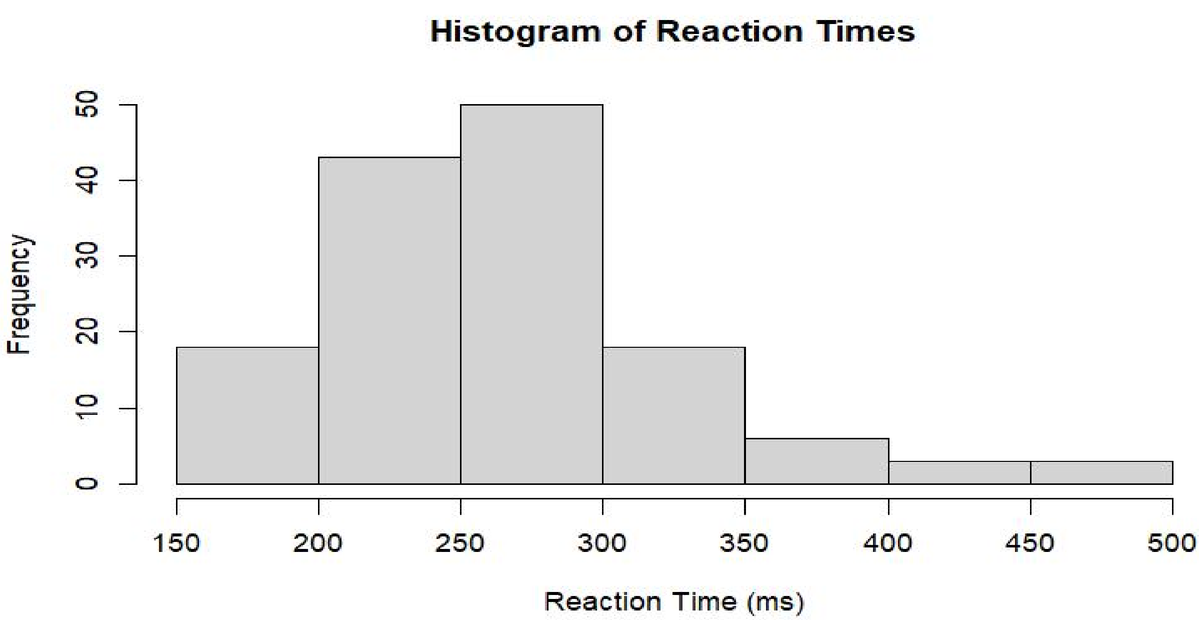
This report explores the Reaction Time Survey data collected at the beginning of the semester. We want to explore what kind of factors will potentially affect the reaction time of students, and want to use regression analysis to fit models that describe the relation between predictor variables and the reaction time. The dataset includes various predictors such as class year, age, average sleep time, caffeine intake, visual acuity, etc.

The analysis focuses on constructing and refining multiple regression models. Initial exploratory data analysis was conducted to understand the distribution of variables and their potential relationships. After that, we pre-process the raw data by dividing and combing some categories of categorial predictors before model building. Subsequent model building involved fitting multiple linear regression models, non-parametric models, one-way ANOVA, variable selection, ANCOVA, collinearity check, lack of fit test check, diagnosing potential issues through residual plots and leverage statistics, and enhancing model accuracy by transforming variables as necessary.

Key findings suggest that variables such as WiFi.stable, Visual.acuity, Input.device, Avg. hours. exercise, Sport.freq, Temp. level, Stress. level and Class significantly impacts reaction time.

Exploratory Data Analysis:

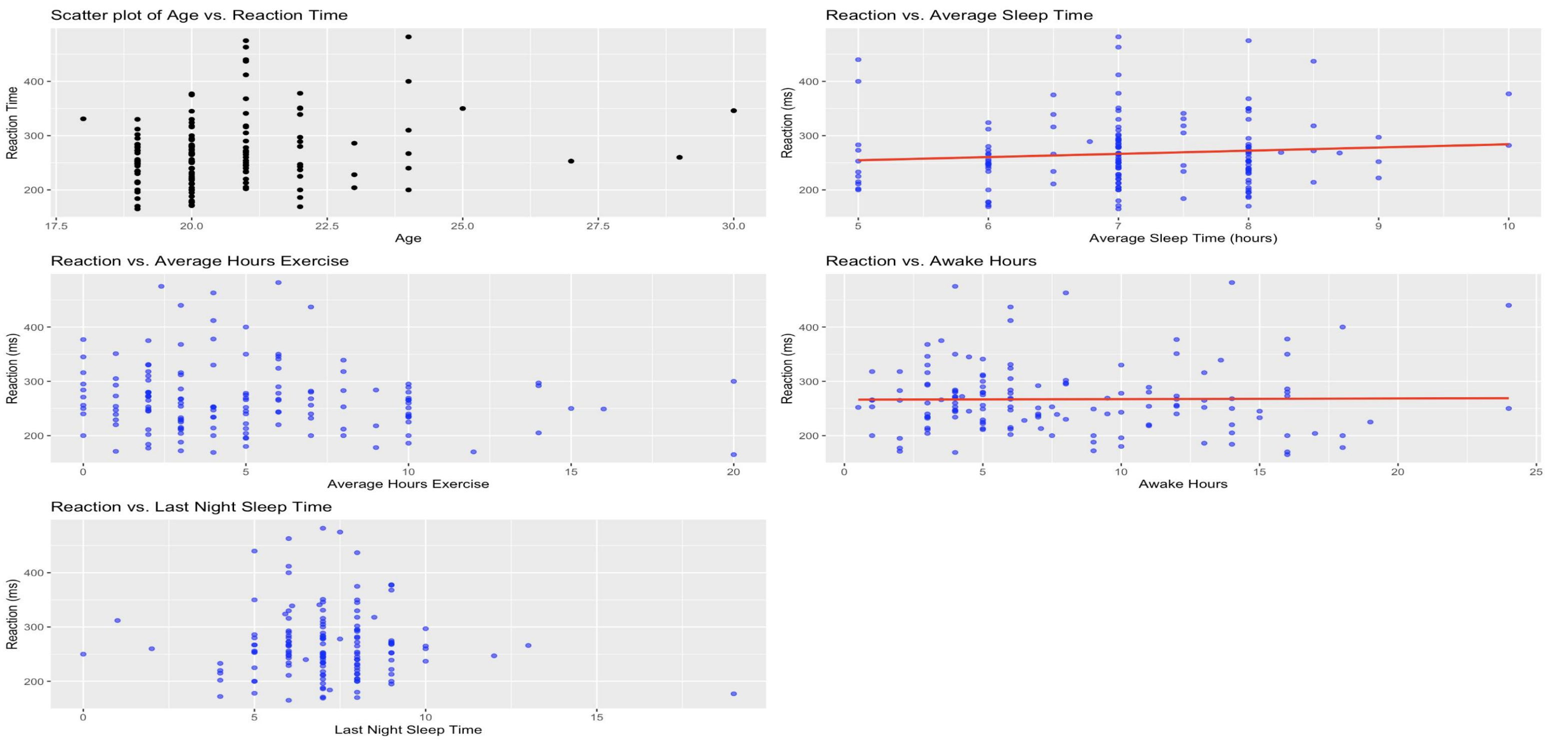
Data Understanding: The Reaction Time Survey dataset, sourced from a class survey, includes 23 variables aimed at exploring factors affecting students' reaction times and 141 collected observations from the survey. The variables encompass demographics (age, class year), lifestyle choices (average sleep time, caffeine intake), and environmental factors (noise level, visual acuity). Initial analyses focus on response variable reaction time:



From the Histogram of Reaction Time, we can tell most of the reaction time is within 200ms-300ms. And from the R output, we know:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 165     228     256     267     292     482
```

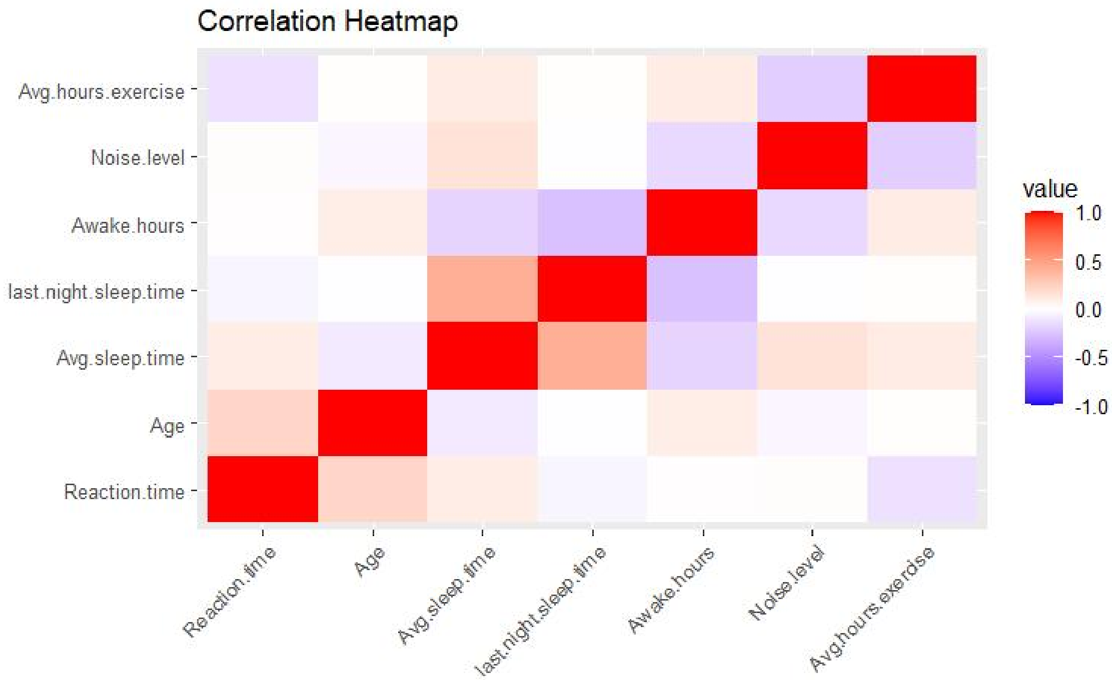
For numerical variables (last.night.sleep.time, age, average sleep time, awake hours, Avg.hours.exercise), we create a scatter plot for numerical data relationships: From the graph, we can see the approximate positive linear relation between reaction time and average sleep time.



For categorical variables, we will explore them in the pre-process part.

Data Insightfulness:

We tried to figure out if there exists some potential collinearity between numerical variables via correlation heatmap in R:



From the plot, we can see for most of the variables, their absolute value of correlation is low. Relatively, last.night.sleep.time and Avg. sleep. time; Awake. Hours and last.night.sleep.time have a higher absolute correlation. This might cause collinearity when we fit the model. We will check collinearity when the model is built.

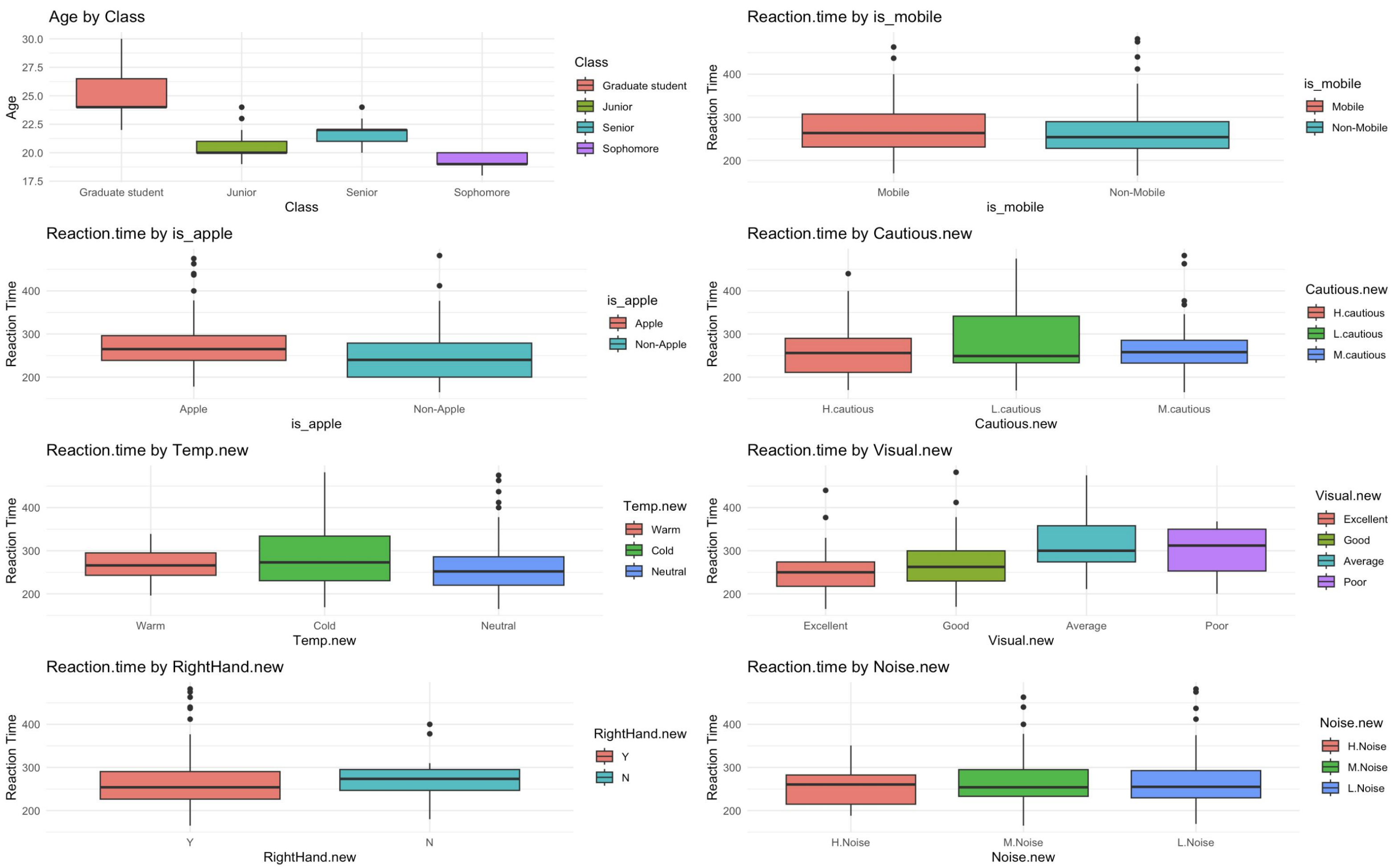
Data Pre-processing:

One feature of this survey is that it contains lots of categorical variables. So we will examine them first. We noticed the following category predictors have categories that only contain a few observations, we process these categorical predictors by dividing and combining some categories to make sure each category will not have too few observations. Based on the original data, we create the following new categorical variables:

1. Input Device(Inputdevide.new):
   Click-based: 43
   Tap-based: 98
2. Fatigue Levels(Fatigue.new):
   H.Fatigue (High Fatigue): 18
   M.Fatigue (Moderate Fatigue): 50
   L.Fatigue (Low Fatigue): 73
3. Noise Levels(Noise.new):
   H.Noise (High Noise): 24
   M.Noise (Moderate Noise): 49
   L.Noise (Low Noise): 68
4. Primary Hand Usage(RightHand.new):
   Y (Right Hand): 127
   N (Not Right Hand): 14
5. Visual Acuity(Visual.new):
   Excellent: 67
   Good: 54
   Average: 15
   Poor: 5
6. Temperature Levels(Temp.new):
   Warm: 25
   Cold: 15
   Neutral: 101
7. Cautiousness Levels(Cautious.new):
   H.cautious (High Cautiousness): 41
   L.cautious (Low Cautiousness): 22
   M.cautious (Moderate Cautiousness): 78
8. Device OS Type (Mobile vs. Non-Mobile):
   Mobile: 20
   Non-Mobile: 121
9. Device OS (Apple vs. Non-Apple):
   Apple: 96
   Non-Apple: 45

We also remove the alcohol.intake, since there is only observation with “yes.

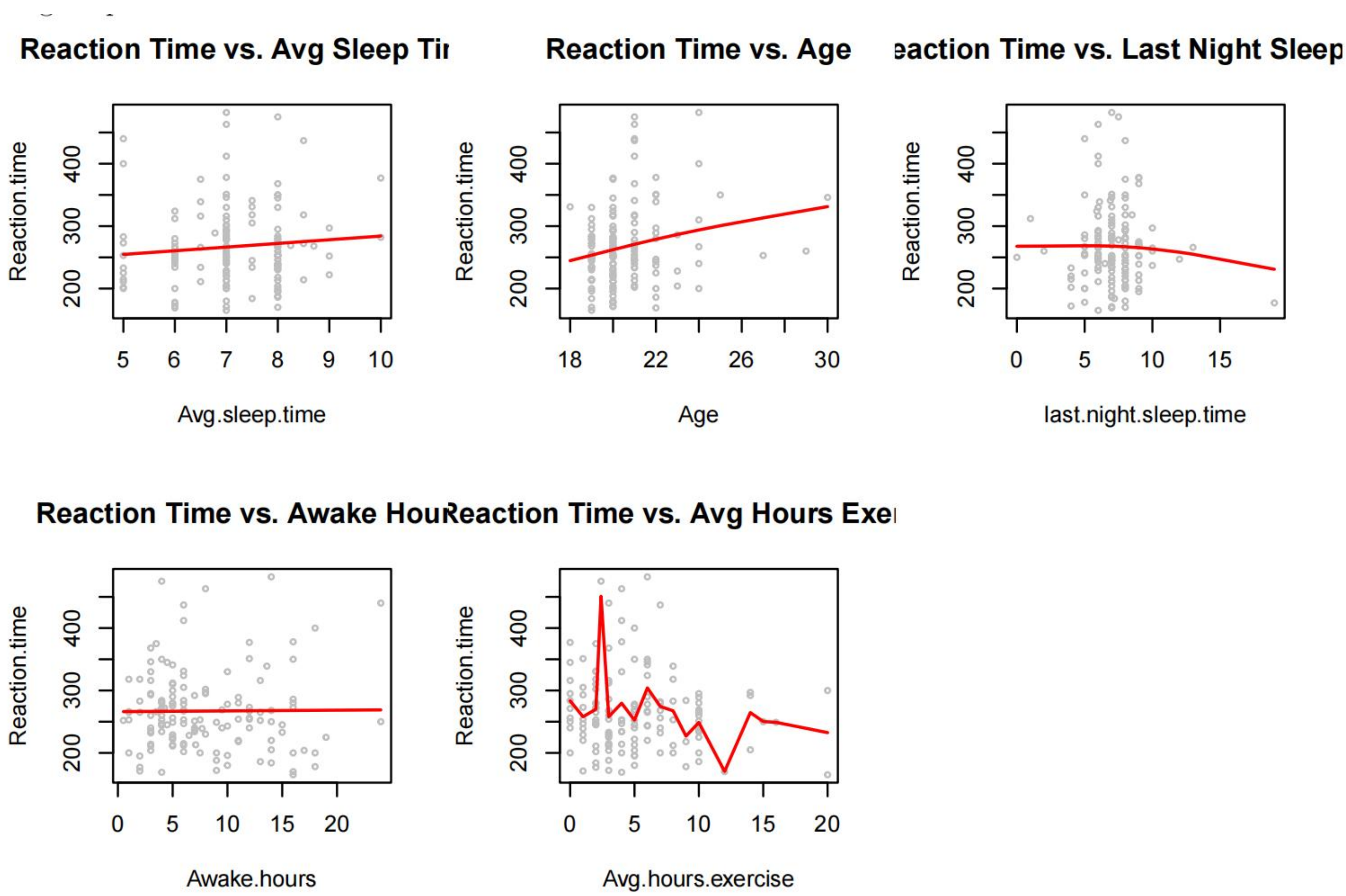
After we process our categorical variables, we can explore them now. We first draw box plots for them:



From the boxplots, we notice that it seems "visual.new" and "inputdevice.new" has a more obvious influence on reaction time, which implies their significance in later model building.

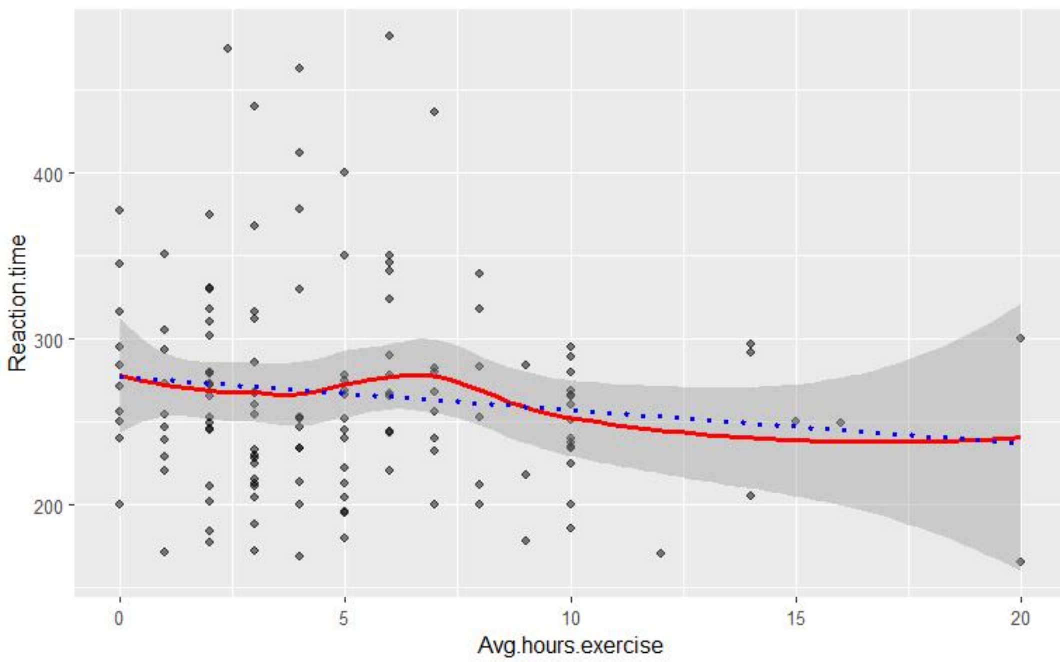
Model Building:

We begin with the general method: The non-parametric model. There are 5 numerical variables in our dataset, we want to first use smoothing splines to determine their relationship with reaction time. Here is the plot of the model we fit:



We find except for Avg.hours.exercise and Reaction time, all of the others appear to have a linear relation with Reaction time in the graph.

To further explore the relation between Avg.hours.exercise and Reaction time, we apply the local polynomial to fit the pattern here. We draw the fitted local polynomial model and its corresponding confidence interval. From the plot, we can tell the linear model is inside the confidence interval of the local polynimial, so we can say linear regression is plausible for Avg.hours.exercise.

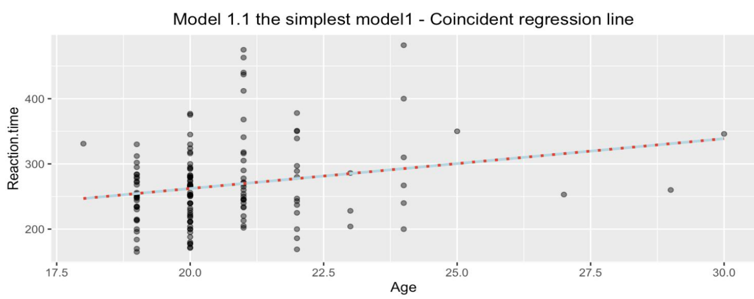


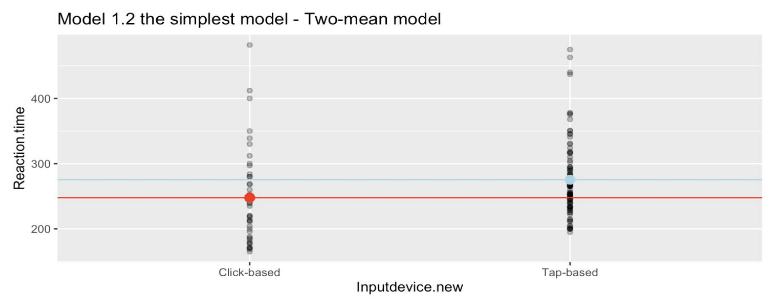
From the above output, we can reasonably say that a linear model is plausible. Thus, for these numerical variables, linear regression is plausible. So we fit g_numerical = Reaction.time ~ Age + Avg.sleep.time + last.night.sleep.time + Awake.hours + Avg.hours.exercise.

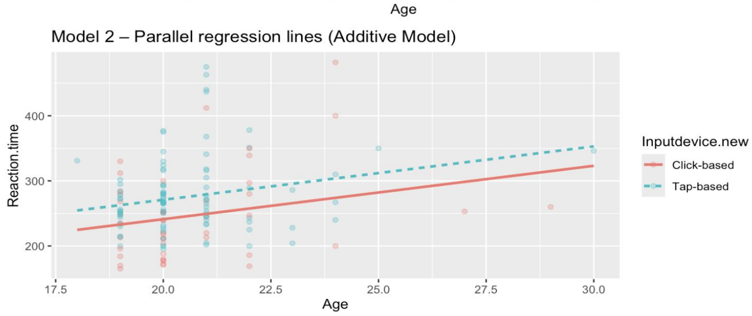
Since the above model only contains numerical variables. The next step is to include the categorical variables. From the previous boxplot of Inputdevice.new, we find that this categorical variable might be significant. We apply A/B test on Inputdevice.new. We first check for assumptions. From Shapiro-Wilk test, its p-value < 0.05, thus the Inputdevice.new is not normal. From Levene's Test for Equality of Variances, since the p-value for the F -test is greater than 1% level, then we conclude that there is no evidence of a non-constant variance. From the Mann-Whitney U test, its p-value = 0.002503 < 0.05, the reaction time of Click-based and Tap-based are significantly different.

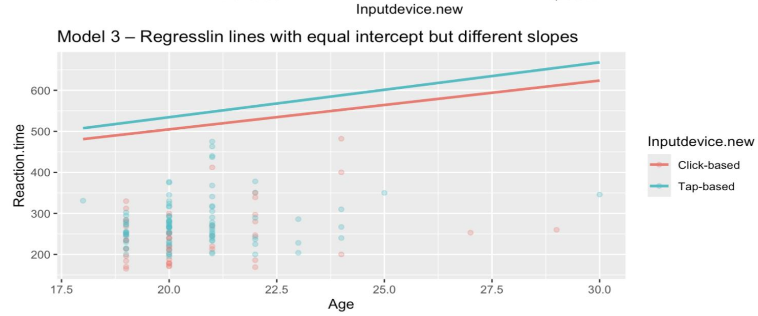
To further explore on Inputdevice.new, we want to explore the relationship between Age & Inputdevice & Reaction Time. Since Age in g_numerical is the only significant predictor. We treat Click-based as our reference level and try the following models and plot them:

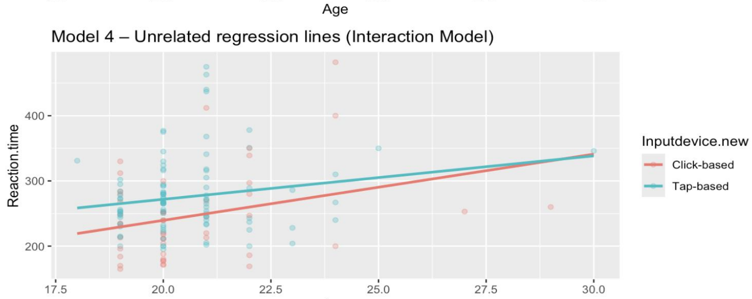
Model 1.1 the simplest model1 - Coincident regression line; Model 1.2 the simplest model - Two-mean model; Model 2 – Parallel regression lines (Additive Model); Model 3 – Regresslin lines with equal intercept but different slopes; Model 4 – Unrelated regression lines (Interaction Model)











Note: for model 3, Since we require an equal intercept for model 3, so from the plot, we can clearly tell this model fits the data poorly.

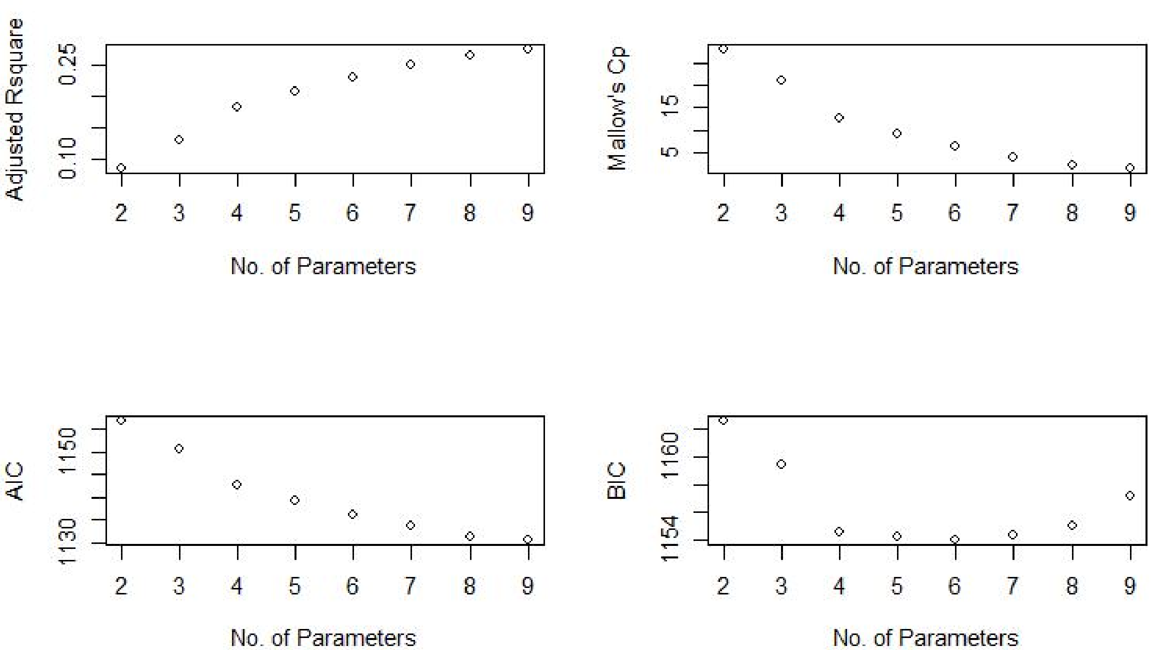
Then we use a sequential ANOVA table to compare these models, from the result, we should use the additive model.

```
Analysis of Variance Table

Response: Reaction.time
                     Df Sum Sq Mean Sq F value   Pr(>F)
Age                   1  25918 25917.9  7.1414 0.008447 **
Inputdevice.new       1  26461 26460.9  7.2911 0.007804 **
Age:Inputdevice.new   1   1322  1322.5  0.3644 0.547070
Residuals           137 497205  3629.2
---
```

So far, we have explored the relationship between Age & Inputdevice & Reaction Time, and fit models to describe the relation between them, and find the additive model is the best one. We are interested to explore if there are other variables related to the reaction time, so we continue to do the following:

To help us better select predictors, we apply variable selection with Criterion-based procedures.

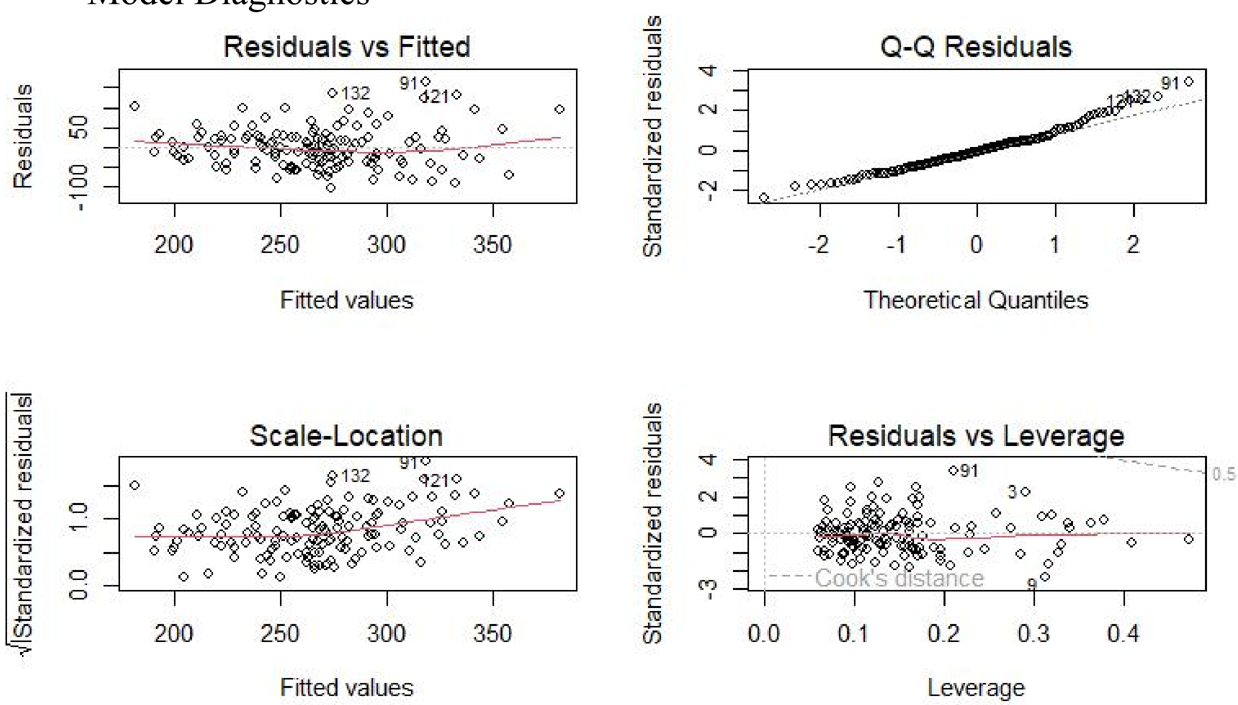


From the plot, except for BIC, they suggest that we need 9 parameters for our model. Then, we have myfit = Reaction.time ~WiFi.stable + Visual.new + Inputdevice.new +Avg.hours.exercise+ Sport.freq+ Temp.new+Stress.level+Class. Compare with full_model = Reaction.time ~ Class + Avg.sleep.time + last.night.sleep.time + Awake.hours + Stress.level + Distraction + Noise.new + Temp.new + Game.freq + Sport.freq + Avg.hours.exercise +Caffein.intake + is_mobile + is_apple + Age + Cautious.new + Inputdevice.new + Visual.new + RightHand.new + Fatigue.new + WiFi.stable + Use.primary.hand, we use F-test and p-value is 0.7301 > 0.05, so we conclude that we prefer myfit model.Then we check for the potential collinearity for myfit model.

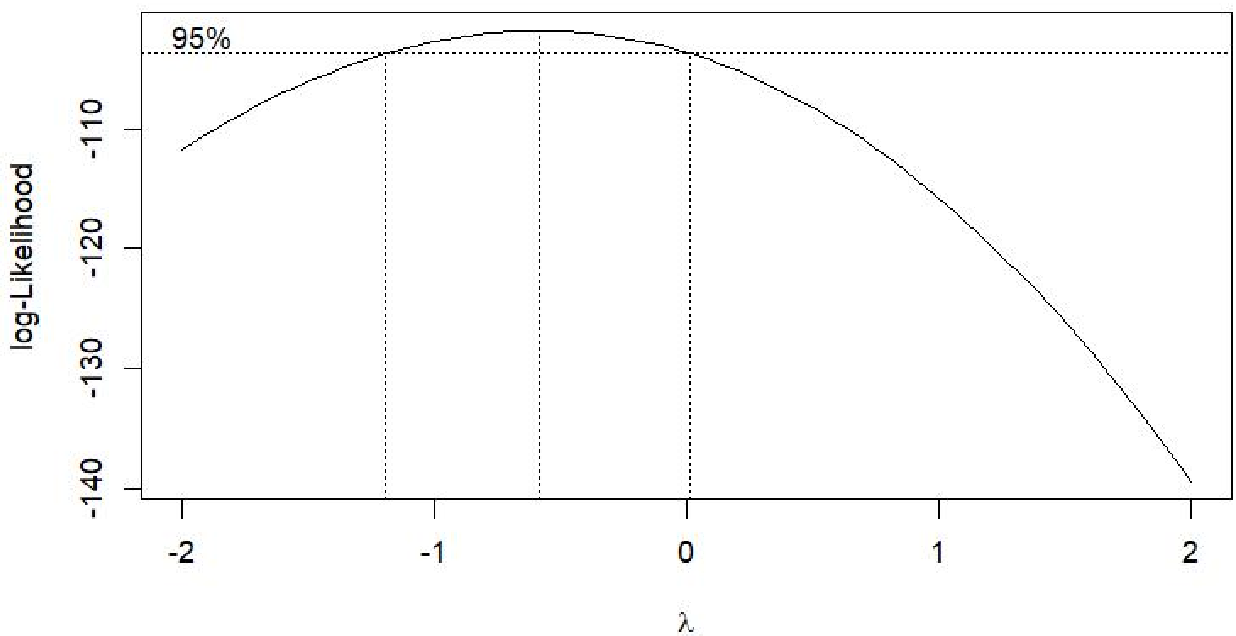
```
       WiFi.stableUnstable              Visual.newGood           Visual.newAverage              Visual.newPoor       Inputdevice.newTap-based
                  1.124852                    1.230512                    1.349735                    1.103892                    1.161219
        Avg.hours.exercise             Sport.freqNever        Sport.freqOnce a week          Sport.freqRarely Sport.freqSeveral times a month
                  1.495930                    4.863149                    5.119844                    8.692810                    5.426370
Sport.freqSeveral times a week            Temp.newCold             Temp.newNeutral              Stress.levelLow         Stress.levelModerate
                  6.053217                    1.717980                    1.694210                    2.217254                    2.427799
     Stress.levelVery High       Stress.levelVery Low                 ClassJunior                 ClassSenior              ClassSophomore
                  1.400713                    1.636937                    4.339715                    2.573094                    4.365765
```

Since there is no vif > 10 and in the previous heatmap, we don't see any high correlation, so there is no collinearity.

Model Diagnostics

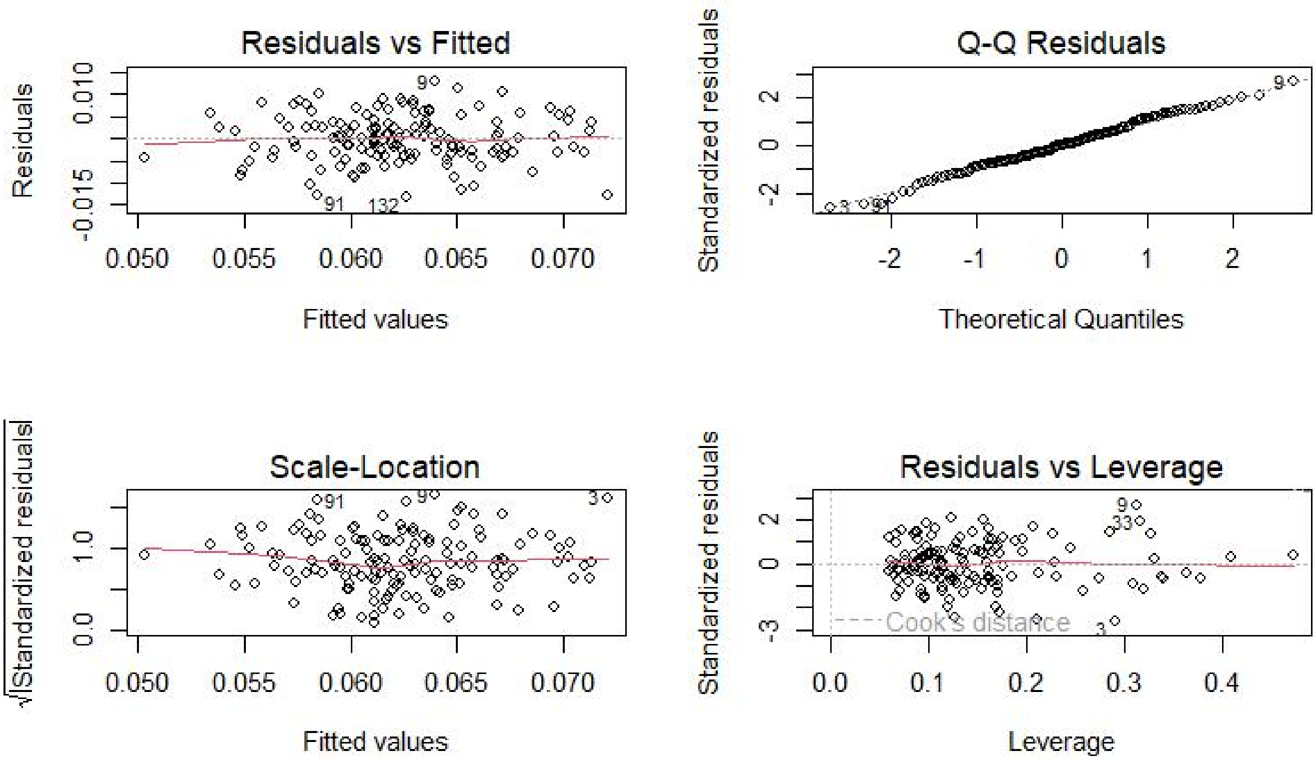


For myfit model, we use Breusch-Pagan Test (BP Test) to check for constant variance. Since p-value = 0.04997 < 0.05, combined with the plot, we conclude no constant variance. For myfit model, we use the Shapiro-Wilk test (SW test) to check if residuals follow normal distribution. Since p-value = 0.009174 < 0.05, we conclude residuals don't follow the normal distribution. For myfit model, we use Durbin-Watson statistic (DW test) to check if residuals residuals are uncorrelated. Since p-value 0.87 > 0.05, we conclude residuals are uncorrelated.

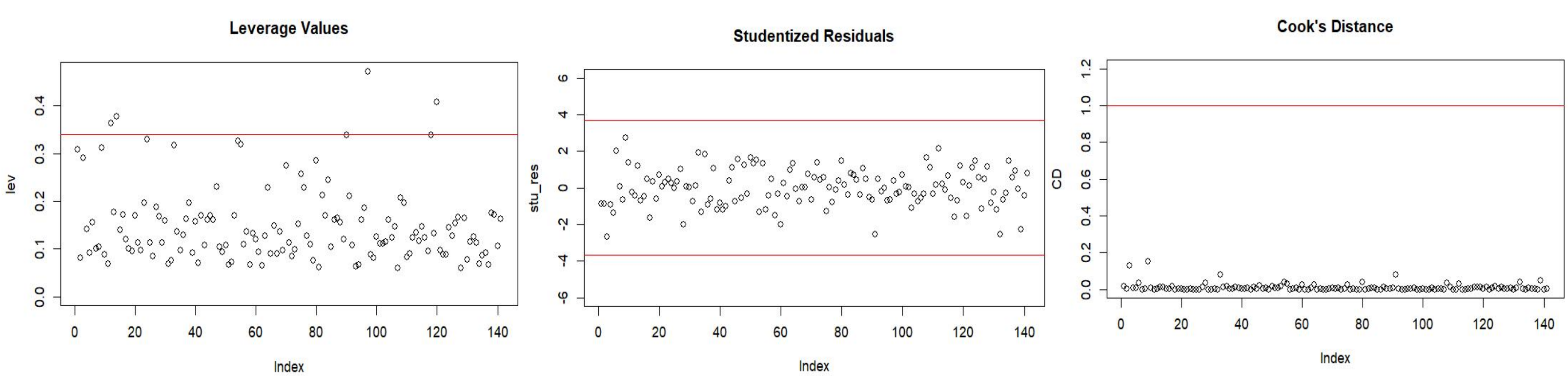


From the above result, assumptions for the MLR model fail. So we consider remediation of transforming the response variable Reaction. time. We draw a box cos plot, and we select the optimal λ = -0.05. And use (Reaction. time)^-0.5 as the response variable to the refit myfit as transformed_myfit. transformed_myfit has smaller RSE(0.005824) than myfit(53.99) and it has a larger Adjusted R-squared:(0.2647) than myfit(0.2592), so we prefer transformed_myfit.

Then we repeat the above model diagnostics on transformed_myfit. The test results of the BP test, SW test, DW test, and plots show that all assumption for MLR holds.



Furthermore, we check for leverage points, outliers, and influential observations. From the following graph, we can see there are a few high-leverage points, no outliers, and no influential observations. Since none of the high-leverage points is an outlier or influential observation, we keep these high-leverage points.



In addition, we perform a lack of fit test here:

```
Analysis of Variance Table

Model 1: Transformed_Reaction.time ~ WiFi.stable + Visual.new + Inputdevice.new +
    Avg.hours.exercise + Sport.freq + Temp.new + Stress.level +
    Class
Model 2: Transformed_Reaction.time ~ factor(WiFi.stable) + factor(Visual.new) +
    factor(Inputdevice.new) + factor(Avg.hours.exercise) + factor(Sport.freq) +
    factor(Temp.new) + factor(Stress.level) + factor(Class)
  Res.Df       RSS Df  Sum of Sq      F Pr(>F)
1    120 0.0040699
2    105 0.0036184 15 0.00045148 0.8734 0.5951
```

Since p-value > 0.05, we conclude that there is no lack of fit.

Discussion of Results and Conclusions:

Finding Summary:

WiFi Stability: Reaction time improves with stable WiFi conditions, emphasizing the importance of reliable internet access for reaction time surveys.

Input Device Usage: The type of input device used (Click-based vs. Tap-based) significantly affects reaction time, with click-based devices associated with quicker responses.

Sports Activity: There is a positive association between sports frequency and reaction time, suggesting that regular physical activity can enhance cognitive speed.

Stress Levels: Moderate stress and Low stress levels correlate with faster reaction times, pointing to the impact of stress on cognitive functions.

Temperature Effects: Temperature affects reaction times, with cold temperatures potentially slightly fostering better cognitive performance.

No Collinearity Concerns: After adjustments, the final model displayed no significant multicollinearity among predictors, ensuring the reliability of the regression coefficients.

Assumption Validation: Model diagnostics confirmed that transformed data met the assumptions of linear regression, validating the analysis results.

Reflect on Challenges and Next Steps:

The largest challenge we met in this survey data contains too many categorical variables, which need us to further consideration of handling them. In the beginning, we tried to convert some of the categorical variables to numerical variables. We failed since we could not find an appropriate scale to convert. For example, for stress level, we cannot assume the relation between each level is linear. That's why we begin with numerical variables first. Another challenge is that some of the categories of categorical variables have too few observations, such as Alcohol. Intake. We solved this by combining some categories and removing some variables.

In addition to reaction time, further research could explore other variables such as stress level, fatigue level, and game frequency as response variables. And maybe if we can divide the current data into training and testing groups, to evaluate our fitted model.

Reflect on Lessons Learned or Study Modifications:

For survey design, we suggest that we should standardize these "level" variables. For example, for stress levels, different people have different criteria for their own stress. We may attach a standardized psychological stress table so that every participant can evaluate their stress under the same measure. Similar to WIFI stability, we may need to provide a quantitive definition of stable WIFI. We may define the network speed > 10Mbps as stable.

In addition, for data collection, we may collect more data from other sources since the backgrounds of students in STAT 425 are pretty similar. We may collect data from their family members so that the age will greatly vary. Since age is always considered significant for reaction time, in our model age is a significant predictor. One reason might be students in this class have very similar ages.