Rigorous Modeling Techniques for Estimating Student Reaction Times

Outline

1. Abstract

The following project aims to use rigorous modeling techniques to estimate student reaction times. To achieve this goal, we first performed data pre-processing which involved removing low-information and highly correlated variables and re-leveling some categorical variables. Following that we used the forward variable selection method to highlight the variables most significant to estimating the reaction time. These variables are “Vision”, “Age”, “Temperature”, “Input.device”, “Avg.sleep.time”, “Distraction”, “Sport.freq”, and “Noise.level”. We further constructed a linear additive model with these variables (baseline model). The resulting Adjusted R2 was not very high (=0.2009). We also conducted diagnostics to assess model assumptions, both graphically and through formal tests. While our model satisfied constant variance and uncorrelated errors, the normality assumption was violated. To address this, we implemented a Box-Cox transformation as a remedial measure. In the final analysis stage, we compared our model to a model with interaction terms and a model whose variables were selected by LASSO Regression. Based on Adjusted R2 (=0.324), the model with interaction terms outperformed both the baseline and LASSO-selected variable models.

2. Exploratory Data Analysis (EDA)

2.1. Data Understanding

In this project, we used a "Reaction Time Survey" dataset which was collected manually by STAT 425 students. This dataset contains the results of the Human Benchmark reaction test taken by each student and some additional information about the students' backgrounds. The raw dataset contains 141 observations and 23 variables (7 num, 16 cat). There were no missing values detected. For this project, we chose our response variable to be "Reaction.time" and the remaining 22 variables were chosen to be predictors.

	Reaction.time <int>	Class <chr>	Age <int>	Avg.sleep.time <dbl>	last.night.sleep.time <dbl>	Awake.hours <dbl>	Fatigue.level <chr>	Stress.level <chr>	Distraction <chr>	Noise.level <int>	Temp.level <chr>
1	368	Junior	21	8	9	3	Slightly Fatigued	Low	Yes	6	Cold
2	280	Sophomore	19	7	6	11	Moderately fatigued	Moderate	No	2	Neutral
3	284	Sophomore	19	8	7	4	Slightly Fatigued	Very Low	Yes	6	Neutral
4	297	Junior	22	9	10	8	Slightly Fatigued	Moderate	No	5	Cold
5	269	Junior	21	8	7	6	Moderately fatigued	Moderate	Yes	5	Neutral
6	165	Sophomore	19	7	6	16	Moderately fatigued	Moderate	No	4	Neutral

Game.freq <chr>	Sport.freq <chr>	Avg.hours.exercise <dbl>	Caffein.intake <chr>	Alcohol.intake <chr>	Visual.acuity <chr>	Primary.hand <chr>	Use.primary.hand <chr>
Several times a week	Several times a month	3	No	No	Poor	Right hand	Yes
Rarely	Several times a week	10	Yes	No	Excellent	Right hand	Yes
Several times a week	Daily	9	No	No	Excellent	Right hand	Yes
Rarely	Several times a week	14	Yes	No	Excellent	Left hand	No
Several times a week	Rarely	10	No	No	Excellent	Right hand	Yes
Daily	Several times a week	20	No	No	Excellent	Right hand	Yes

Cautious.level <chr>	Input.device <chr>	Device.OS <chr>	WiFi.stable <chr>
Moderately cautious	Touch screen	iPhone(Smartphone-IOS)	Stable
Moderately cautious	Mouse	Laptop-Windows	Stable
Moderately cautious	Mouse	Laptop-Windows	Stable
Very cautious	Mouse	Laptop-Windows	Stable
Extremely cautious	Keyboard	Desktopcomputer-Windows	Stable
Moderately cautious	Mouse	Desktopcomputer-Windows	Stable

2.2. Data Insights

Summary Statistic (numerical variables)

```
 Reaction.time       Age         Avg.sleep.time  last.night.sleep.time  Awake.hours      Noise.level     Avg.hours.exercise
 Min.   :165   Min.   :18.00   Min.   : 5.000   Min.   : 0.0          Min.   : 0.500   Min.   :1.000   Min.   : 0.000
 1st Qu.:228   1st Qu.:20.00   1st Qu.: 6.500   1st Qu.: 6.0          1st Qu.: 4.000   1st Qu.:2.000   1st Qu.: 2.000
 Median :256   Median :20.00   Median : 7.000   Median : 7.0          Median : 6.000   Median :4.000   Median : 4.000
 Mean   :267   Mean   :20.62   Mean   : 7.108   Mean   : 7.1          Mean   : 7.799   Mean   :4.007   Mean   : 5.081
 3rd Qu.:292   3rd Qu.:21.00   3rd Qu.: 8.000   3rd Qu.: 8.0          3rd Qu.:11.000   3rd Qu.:6.000   3rd Qu.: 7.000
 Max.   :482   Max.   :30.00   Max.   :10.000   Max.   :19.0          Max.   :24.000   Max.   :9.000   Max.   :20.000
```

From the summary statistic table, we can observe that "Avg.sleep.time" and "Noise.level" have relatively small variability and a small number or no outliers. Whereas, the remaining variables have larger variability and more outliers. Let's verify this from the box plots.

Outliers (numerical variables)

From the box plots (Fig. 1), we can see that all numerical variables except "Noise.level" have outliers, and "Avg.sleep.time" has only one outlier. At this stage, we decided to leave them. We address unusual observations in **Section 3.2**.

Relationship (numerical variables)

From the scatter plots (Fig. 2), we can see that there is no distinct linear relationship between "Reaction.time" and any of the predictors. We also see that "Age" and "last.night.sleep.time" have many outliers.

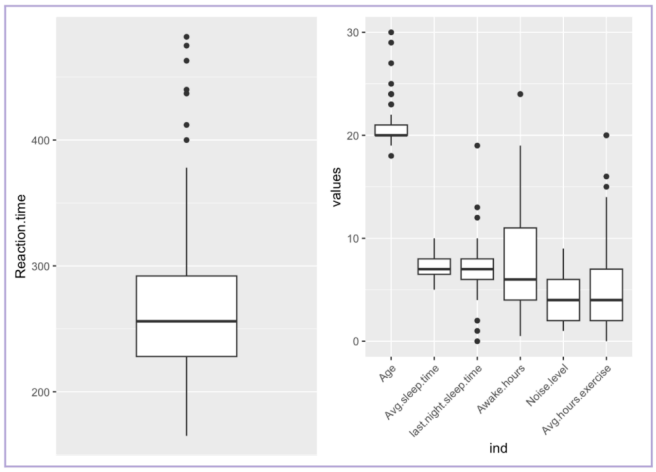

Fig. 1

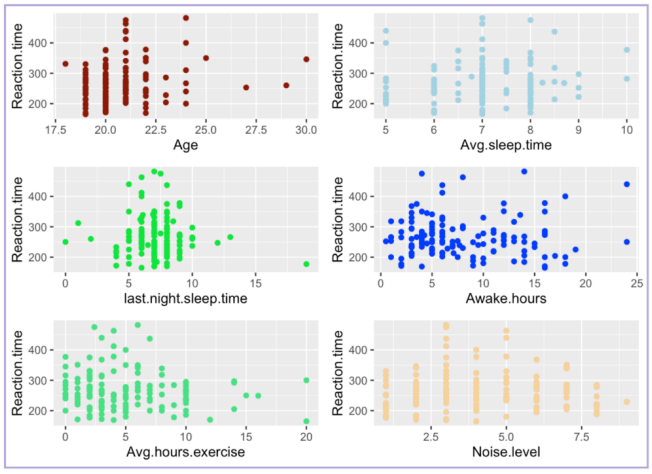

Fig. 2

2.3. Data Pre-processing

After pre-processing, our dataset was reduced from 23 (7 num, 16 cat) to 17 (7 num, 10 cat) variables. The number of observations stayed the same (=141). Let's now go into some details on how we performed data pre-processing.

a) Re-leveling redundant columns

Certain columns were heavily skewed in the sense that a few groups contained most of the data points, while the frequency of occurrence in the rest was very low. Thus, the number of groups could be reduced, by merging such low-frequency classes.

Applying this re-leveling procedure, we created the following new columns:

1. **"Fatigues**" with levels *H.Fatigue*, *M.Fatigue* and *L.Fatigue* (instead of the corresponding 5 levels in "Fatigue.level")
2. **"Temperature"** with levels *Neutral* and *Non Neutral* (instead of the corresponding 5 levels in "Temp.level")

3. **"Vision"** with levels *Excellent*, *Good* and *Suboptimal* (instead of the corresponding 5 levels in "Visual.acuity")

b) Removing low information columns

Some of the categorical variables had just a few groups of data, with a very high percentage (greater than 90%) of data points belonging to a single category or class. Such columns do not provide much information when it comes to checking their impact on predicting the response, and thus they can be dropped. Hence, the columns "WiFi.stable", "Alcohol.intake", "Primary.hand", and "Use.primary.hand" were removed.

c) Removing correlated columns

Removing correlated columns is one of the most important parts of the data pre-processing because it would prevent errors during the model-building process. This part was a bit tricky because the Pearson correlation cannot be used for categorical and mixed variables (ref: The Search for Categorical Correlation). Therefore, we have decided to use **Pearson correlation** for numerical-numerical, **Cramer's V association** for categorical-categorical, and **ANOVA** for numerical-categorical pairs (ref: Stack OverFlow). Here is a list of pairs that showed the highest correlation/association.

x <chr>	y <chr>	assoc <dbl>	type <chr>
Age	Class	0.8325782	anova
Device.OS	Input.device	0.7093000	cramersV
last.night.sleep.time	Avg.sleep.time	0.6085257	correlation
Avg.hours.exercise	Sport.freq	0.4698023	anova
Noise.level	Distraction	0.4226731	anova

We set the correlation threshold to be 0.7 therefore we had to remove either "Age" or "Class", and either "Device.OS" or "Input.device". We decided to remove "Class" (because "Age" was numerical) and "Device.OS" (because has many levels).

3. Model Building

3.1. Variable Selection and Model Fitting

Let's start the model building from a simple additive model that takes all predictors to estimate our response variable "Reaction.time". From the summary table (ref: R code), we may notice that some of the variables are significant and some of them are not. The Adjusted R-squared for the full model was **0.2009** which is a pretty bad goodness of fit.

a) Checking for multicollinearity: VIF

Next, we need to check if there is a multicollinearity between predictors. We utilized the VIF score (GVIF for categorical variables) for that purpose.

Our threshold for VIF score was **6** and as we can see from the table on the right none of the variables showed VIF score higher than the set threshold. Therefore we conclude that there is no multicollinearity and that we can safely proceed further with the model building.

	GVIF
Age	1.307113
Avg.sleep.time	1.631561
last.night.sleep.time	1.684050
Awake.hours	1.661229
Stress.level	3.785597
Distraction	1.810210
Noise.level	1.619208
Game.freq	3.760633
Sport.freq	5.521550
Avg.hours.exercise	1.730667
Caffein.intake	1.272957
Cautious.level	2.954558
Input.device	3.160536
Fatigues	1.926047
Temperature	1.163176
Vision	1.942672

b) Variable selection: forward

The next step in model building was selecting the best subset of variables. We used a test-based procedure for that, specifically the **forward selection** method with p=0.15. The following method identified "Vision", "Age", "Temperature", "Input.device", "Avg.sleep.time", "Distraction", "Sport.freq", and "Noise.level" to be the most important variables. We fitted our next reduced linear model with these variables only and the Adjusted R-squared **increased to 0.2362**. We then proceeded to checking our dataset for unusual observations and to the assumptions diagnostics.

3.2. Diagnostics and Remedies

a) Unusual observations

High Leverage Points: were checked based on the leverage values. From Fig. 3, we see that observation 112 has leverage = 1. This is probably because person 112 was the only one using a *game controller* as an "Input.device". We have decided to remove this point and refit our model on 140 observations again. The Adjusted R-squared **decreased to 0.2283**.

Outliers: were checked based on the Studentized Residual Test with Bonferroni correction. From Fig. 4, we do not see points outside the red line meaning there must be no significant outliers.

Highly influential points: were checked based on the Cook's Distance. We don't see distances larger than 1, nor do we see distances that are significantly larger than others. Therefore we conclude that there are no highly influential points.

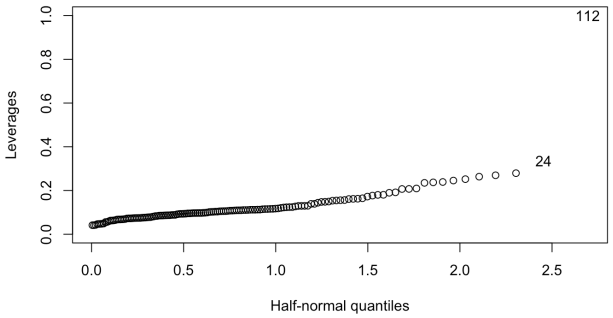

Fig. 3

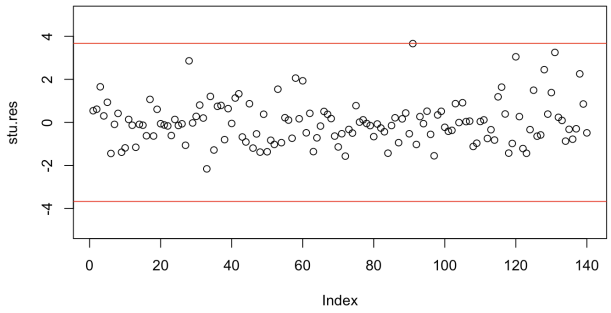

Fig. 4

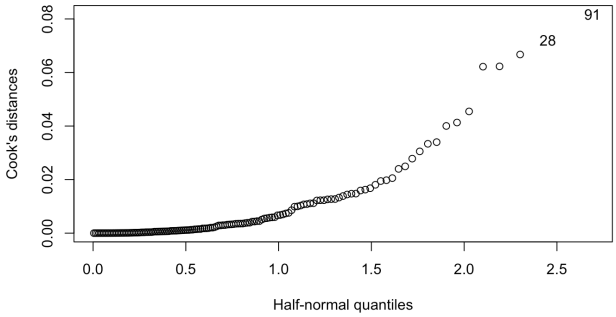

Fig. 5

b) Error assumptions

Normality:

- From *Q-Q plot* (Fig. 6a), we see points at the tail significantly deviating from the red line which suggests that residuals are not normally distributed.

- From *Shapiro-Wilk Test* (ref: R code, line 328), we can see that p-value = 4.003e-05 < 0.05. It means that we reject the null hypothesis and conclude that residuals indeed do not follow Normal distribution.

Let's apply Box-Cox transformation to **y** as a remedial measure. The optimal value for lambda showed to be = -0.6240602 and we will use it to make an appropriate transformation for **y** (ref: equation on the right). Now, let's check Normality again.

$$\frac{y^{-0.6240602}-1}{-0.6240602}$$

- From *Q-Q plot* (Fig. 6b), we see points at the tail become closer to the red line suggesting that residuals became normal after transforming **y**.

- From *Shapiro-Wilk Test* (ref: R code, line 361), we can see that p-value = 0.3276 > 0.05. It means we fail to reject the null hypothesis and can conclude that residuals now follow Normal distribution. This result complies with what we have concluded when assessing the normality graphically.

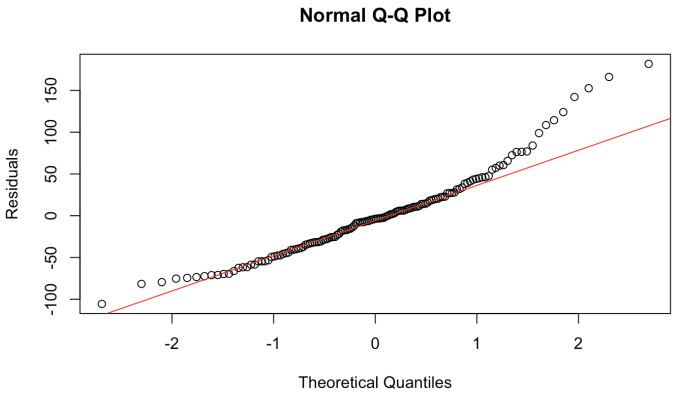

Fig. 6a

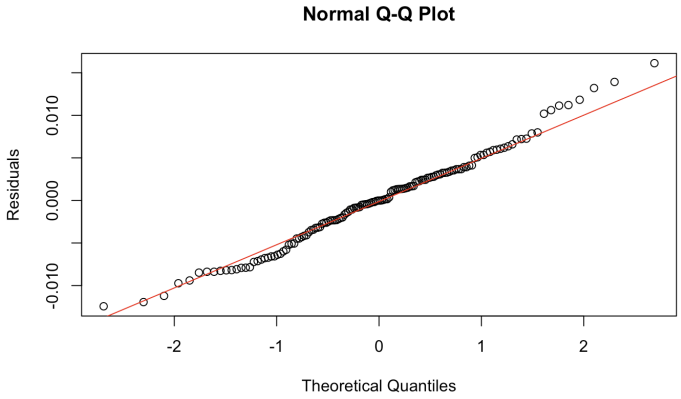

Fig. 6b

Constant variance:

From *Residuals vs Fitted plot* (Fig. 7), we can see that points are distributed randomly across y=0 (red line) with no particular pattern. This suggests that residuals have constant variance.

From *Breusch-Pagan Test* (ref: R code, line 384), we can see that p-value = 0.1627 > 0.05 meaning we fail to reject the null hypothesis and can conclude that residuals have constant variance. This result complies with what we have concluded when assessing the variance graphically.

Correlated errors:

From *Residuals vs Index plot* (Fig. 8), we points distributed randomly across y=0 (blue line) with no particular patter. Therefore we can conclude that errors are not correlated.

From *Durbin-Watson Test* (ref: R code, line 404), we can see that p-value = 0.4601 > 0.05 meaning we fail to reject the null hypothesis and can conclude that errors are not correlated. This result complies with what we have concluded when assessing the correlation between errors graphically.

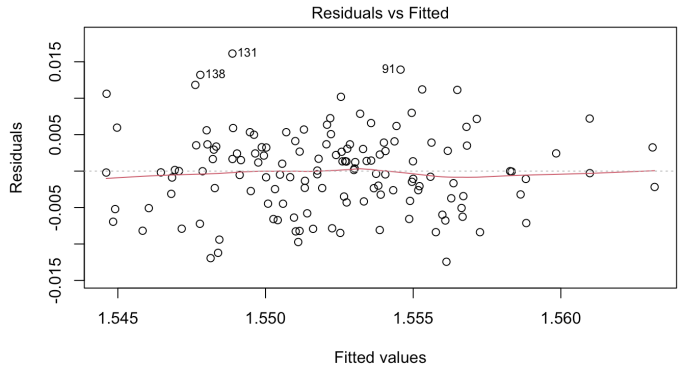

Fig. 7

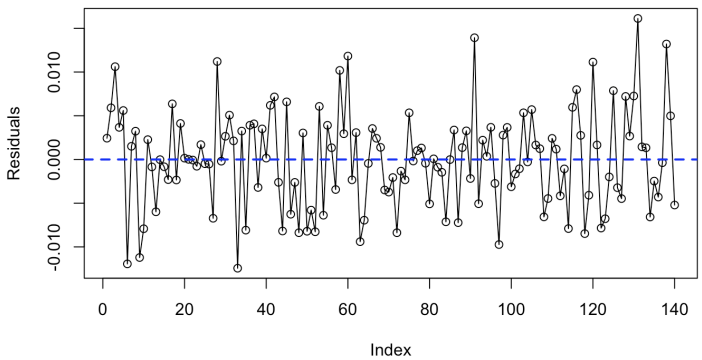

Fig. 8

c) Structure assumptions

Linearity:

From the *Added-Variable plots* (Fig. 9), we may suggest that the linearity assumption is not held. This is because in most of the cases points are not distributed randomly across the blue line (except "Noise.level", "Avg.sleep.time", "DistracitonYes" and "VisionGood"). We also do not see a nonlinear relationship between variables therefore we might assume that there is no need for higher-order terms in our model. However, there might be instances where interaction terms are needed.

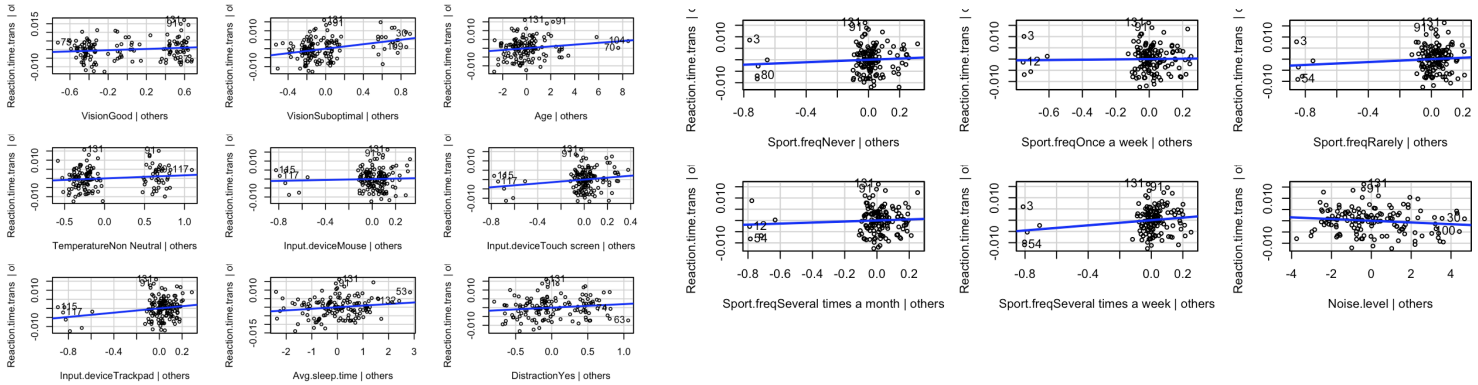

Fig. 9

4. Model Comparison and Selection

Our final dataset is **df_reduced2** with 140 rows and 19 columns (including "Reaction.time" and "Reaction.time.trans") and the final model is **reduced_lm3** (baseline model):

Reaction.time.trans = Vision + Age + Temperature + Input.device + Avg.sleep.time + Distraction + Sport.freq + Noise.level + **Error**

4.1. A model with an interaction term

We want to compare our baseline model to a model with interaction terms. The interaction terms were chosen based on domain knowledge: interact predictors that might depend on each other. These are "Distraction" and "Noise.level", "Sport.freq" and "Avg.sleep.time", "Sport.freq" and "Distraction". We added these interaction terms to **reduced_lm3** to get **interact_model**. We also checked interact_model for assumptions and every assumption was satisfied, meaning we now can proceed to comparing these two models. To compare the two models we first used Adjusted R2:

Adjusted R2 (reduced_lm3) = **0.2375**
Adjusted R2 (interact_model) = **0.324**

Based on the Adjusted R2 we see that the model with interactions performed better at explaining our response variable as its Adjusted R2 is higher than in **reduced_lm3**. We also tried to compare the two models using the Partial F-test as **reduced_lm3** and **interact_model** are nested. Based on this test, we got $p=0.009017 < 0.05$ which means that we reject the null hypothesis and conclude that **interact_model** provides a significantly better fit to the data compared to **reduced_lm3.**

4.2. LASSO Regression

From the Shrinkage Methods lecture, we know that LASSO Regression can also be considered as a variable selection method. We were interested in how a linear model with the variables selected by LASSO Regression would perform compared to our baseline model. For that, we run LASSO Regression to "Reaction.time" vs all predictors. After that, we performed Cross-Validation to identify lambda (=0.04040404) that would result in the best LASSO performance. Based on this value of lambda, LASSO Regression identified "Age", "Stress.level", "Sport.freq", and "Vision" as the most important variables in predicting `Reaction.time`. We then fit these variables into a linear model (**lasso_model**). We also wanted to conduct assumptions diagnostics (and apply appropriate remedies) on this model to ensure a fair and correct comparison. The diagnostics results showed that only Normality assumption was violated, therefore, we applied a Box-Cox transformation as in Section 3.2b, and successfully remedied the Normality Assumption. After finishing all these steps we were good to compare **lasso_model** with **reduced_lm3** baseline model.

To compare two models we used Adjusted R2 because models are not nested.

Adjusted R2 (reduced_lm3) = **0.2375**
Adjusted R2 (lasso_model) = **0.1505**

We see that **reduced_lm3** has a higher Adjusted R2 therefore we may conclude that it should work better in predicting "Reaction.time".

5. Discussion of Results and Conclusion

5.1. Summary

- **Dataset:**
  - Pre-processing: 6 columns removed, 3 re-modeled, 1 observation was removed.
  - Final dataset: 140 observations, 17 (7 numerical, 10 categorical) variables.
  - VIF: no multicollinearity between variables.

- **Forward variable selection:** Vision, Age, Temperature, Input.device, Avg.sleep.time, Distraction, Sport.freq, and Noise.level are the most important variables.

- **Diagnostics:**
  - ➢ Unusual observations: One high leverage point (was removed), no outliers, and no highly influential points.
  - ➢ Error assumptions: Normality violated (remedied with Box-Cox), residuals have constant variance, and errors are not correlated.
  - ➢ Structure assumptions: Linearity violated.

- **Model Comparison:**
  - ➢ A model with interaction vs baseline: The model with interaction performed better than the baseline model based on both Adjusted R2 and Partial F-test.
  - ➢ LASSO Regression: The linear model with variables selected by LASSO Regression performs worse than the baseline model based on Adjusted R2.

5.2. Challenges and Next Steps

Overall, the major challenges faced were related to the presence of categorical variables in our dataset. These challenges are:

1) **Analyzing the correlation**. We initially tried to calculate the Pearson correlation score for numerical and categorical variables together but got very uninformative and weird results. We then realized that Pearson correlation is not intended for categorical variables and finding a better way to calculate correlation was tricky. We ended up using Pearson correlation for numerical, Cramer's V association for categorical, and ANOVA for mixed variables. However, we are not sure if it was the best way to calculate correlation. To address this limitation, the next step would be to investigate other more reliable ways to calculate the correlation between mixed variables.

2) **Improving the model performance.** Our initial Adjusted R2 was 0.2009 and after multiple attempts to improve the model our best Adjusted R2 increased only to 0.324 which is still a poor result. This might be because:
   a) Linearity assumption was initially not satisfied (based on Added-Variable plots).
   b) Model was too complicated (too many variables were included).

   To address these limitations we could further reduce the model complexity by using either backward selection or BIC. Another option would be to add more observations to the data.

3) **Revealing the relationship between categorical and response variables.** While trying to asses the linearity assumption of the model we faced challenges in interpreting the Added-Variable plots. It was difficult to say if the variables had linear or nonlinear relationships because

categorical variables are essentially just 0's and 1's. This caused some confusion when deciding whether to add the higher-order or interaction terms. In the future, we could try other methods to investigate the underlying relationship between categorical variables.

5.3. Reflection on Lessons Learned

The first thing that we found interesting about our results is that the variables that we assumed to be important for estimating "Reaction.Time" ended up being not significant. Initially, our intuition was that "last.night.sleep.time", "Awake.hours", "Game.freq", "Cautious.level", and "Fatigues" will be significant for explaining "Reaction.Time", however, after performing forward variable selection none of these variables were selected as important ones.

Another thing that we learned from this project is that categorical variables have to be treated very carefully during analysis and interpretation. We have to be careful with calculating correlations, interpreting the scatter plots, and converting categorical variables to dummy variables so that their inherent meaning won't change. We also learned how to work with real-world data that is not as perfect as in the textbooks. Specifically, we learned how to pre-process such datasets and how to apply remedial measures if some of the linear model assumptions were not met.

For further data collection improvements, we would say that it would be great if the dataset was larger and more diverse. For example, adding more people being drunk the day before the reaction time to the sample space would be very interesting. For further analysis improvements, we could evaluate models more rigorously using Cross-Validation and try other variable selection methods.