**A Multimodal Interface for Solving Equations**

Xinyue (Katherine) Tan

UNI: Xt2213

July 3, 2019

HUDM 5026: Introduction to Data Analysis and Graphics in R

Instructor: Bryan Keller

**1. INTRODUCTION**

The data I used for the project was collected by Carnegie Learning, Inc. between October 12, 2006, and December 20, 2006, and uploaded to DataShop on May 5, 2008. The data is the result of an investigation into student performance solving algebraic equations when presented with worked examples of similar equations, compared to the performance of students not provided with such examples. In this study, students were randomly assigned to two groups. Different from the traditional definition of control and treatment group, the students in the control group were presented with worked examples to copy on the first opportunity to practice a problem type. The treatment group is identical respective to all other items that the researchers are examining (e.g., Teacher, Units, Pre/Post Test, A/B Form, Item) with the exception that it does not receive the experimental manipulation that the control group received. There are two test forms (A, B) for both Unit 7 and Unit 9 which are equation-solving units in 2006 Cognitive Tutor Algebra. Questions on each test form are different, but the level of difficulty and the question types are the same.

The purpose of my project is to: 1) point out which questions students struggle with the most in both Unit 7 and Unit 9 by using simple statistical and graphical methods; 2) test whether or not the students with the benefit of the worked examples will experience benefits to their learning or improved performance over students without examples (pure problem-solving); and 3) build a regression model to investigate the relationship between the dependent variable, the Score, and the independent variables (e.g., Teacher, Condition, Units, Pre/Post Test, A/B Form, Item).

After conducting analysis on the accuracy rate of each question, I found that in Unit 7, questions 4, 6, 7, 9, and 10 on test form A and questions 3, 4, 6, 7, 9, and 10 on test form B were questions that the students struggled with most. More than half of the students failed to answer these questions correctly or at all. In Unit 9, the accuracy rate of questions 4 and 9 on the test form A and questions 3, 4, 7, 9, and 10 on test form B was below fifty percent. Apart from that, by performing the chi-square test, I conclude that the variables “Condition” and “Score” are statistically associated. However, a change in condition does not necessarily directly cause a change in score. In addition, a logistic regression model Score= -0.5317 + 1.0562 \* Units was built to model the outcome variable (score) conditional on the predictor (units) with the accuracy rate of 62.9%.

**2. DATA DESCRIPTION**

Variables included in the dataset are: “School ID”, “Teacher (name of the teacher)”, “ID” (which demonstrates the same information as “Teacher”), “Condition (e.g., control group or treatment group)”, “Units (e.g., Unit 7 or Unit 9)”, “Pre/Post Test”, “A/B Form (two version of the unit test)”, “Item (e.g., question number in the exam)”, “Problem (e.g., the specific equation-solving problem)”, “Score (e.g., 1 represents earning one point, 0 represents earning no point)”, “Answer (e.g., the student’s answer to the question)”, “Explanation (e.g., explanation student wrote near the answer)”, and “Comment (e.g., teacher’s comment).” However, explanations of the following are missing: 1) the definition of review, normal, and transfer under the column named “Type”; 2) the definition of the labels C/I/W/B under the column named “Code”; and 3) the unit of measure of the variable “Period.”

Due to the purpose of my research and the lack of data dictionaries to make some of the variables more understandable (e.g., “Type”, “Code” and “Period”), only variables "Teacher", "Condition", "Units", "Pre/Post Test", "A/B Form", "Item" and "Score" are selected for further analysis.

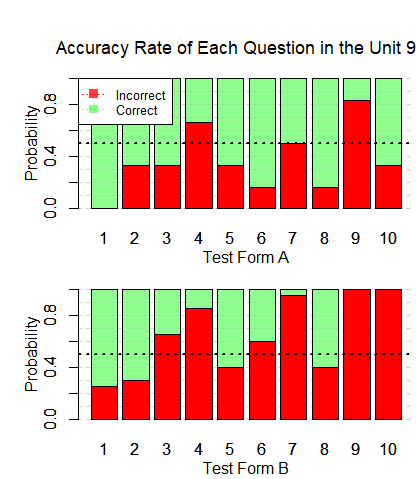
**3. METHODS AND RESULTS**



*Figure 1: Missing Data Patterns of the dataset*

To reveal the structure of missing observations in the 550 observations, Md.pattern() function in the Mice package was used. The missing data pattern is monotone as all of the 10 missing values are under the “Condition” variable (Fig. 1). The missing data is limited to a small number of observations. Thus, instead of imputing data, I removed the 10 observations that have missing values and performed an analysis of the remaining 540 observations. Then, the function factor() is used to encode all variables from vectors to factors.

To answer which questions students struggled with the most in both Unit 7 and Unit 9, the accuracy rate of each question on the two test forms (A, B) for both Unit 7 and Unit 9 are calculated and presented on the parallel histograms. Table() function performed categorical tabulation of data with the variable “Item” and its frequency of “Score.” Then, the function prop.table() computed the marginal proportions based on the score, which reveals what proportion of items are scored 1 and what proportion of items are scored 0. Looking at test form A in Unit 7 as an example, a two-way frequency table of items by score is generated. Then, the proportion of items receiving score 1 and score 0 are calculated respectively (e.g., 26% did not earn a score on item 1 and 74% of students did earn a score on item 1). The parallel histograms with the x-axis representing “Item”, y-axis representing the probability and horizontal dark line marking the 50% probability are drawn to help the visualization. In Unit 7, questions 4, 6, 7, 9, and 10 on the test form A and questions 3, 4, 6, 7, 9, and 10 on test form B were questions students struggled with. More than half of the students failed to answer these questions correctly or at all (Fig.2). Also, in Unit 9, the accuracy rate of questions 4 and 9 on the test form A and questions 3, 4, 7, 9, and 10 on test form B fell below fifty percent (Fig.3). These questions need special attention. Teachers should investigate why students fail to answer those questions correctly, especially for issues that course developers might be able to improve (e.g., clarification of the question asked and the lack of practice for a certain question type).

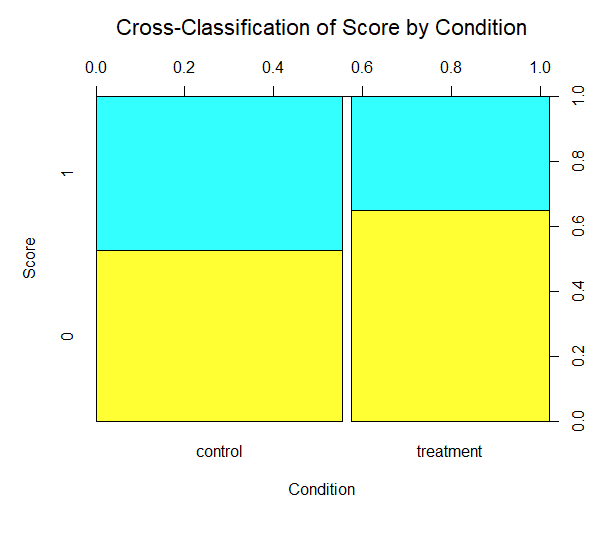


*Figure 3: Accuracy Rate of Each Question in the Unit 9 (Test Form A & B)*



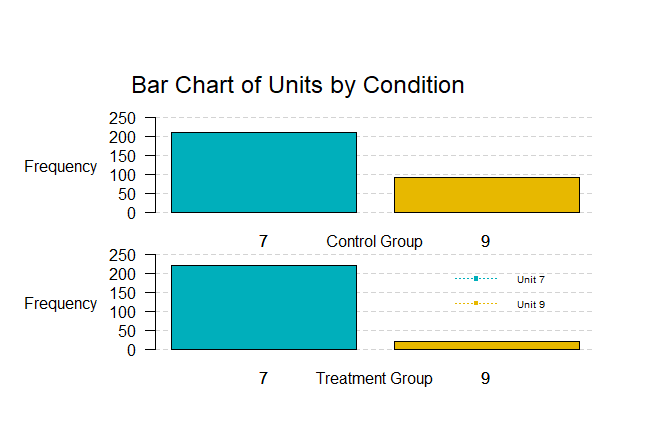
*Figure 2: Accuracy Rate of Each Question in the Unit 7 (Test Form A & B)*

To test whether the students who were given the worked examples will experience benefits to their learning or performance over the students without such examples (pure problem-solving), Spineplots function was first used to plot the relationship between “Condition” and “Score” and then a chi-square test for association was run. Spineplots are useful in that they can visually represent a lack of independence. If the variables are independent, I expect the proportions within each category to be roughly the same. However, the proportion of score 0 accounted for 65% of the population in the treatment group, compared with 50% of score 0 in the control group. This differing proportional distribution by “Score” might reveal departures from independence (Fig. 4). Thus, a chi-square test for association is run to determine if there is a significant relationship between “Score” and “Condition.” The null hypothesis is the independence of Score and Condition. The alternative hypothesis is that they are dependent. Since the p-value is 0.005125, which is less than 0.05, we reject the null hypothesis of independence as there is evidence of a dependency between “Score” and “Condition” meaning the proportion of score 1 differs amongst the control and treatment group. This, combined with the proportion table, leads to the conclusion that the students with the worked examples will experience benefits to their learning and/or performance over students without the same examples.



*Figure 4: Cross- Classification of variable “Score” (e.g.: 1 or 0) by “Condition” (e.g.: control or treatment)*

However, chi-square test for variable “Pre/Post Test” and “Condition”, variable “A/B Form” and “Condition” and variable “Units” and “Condition” were also conducted. The p-values were 0.01565, 0.9155, and 1.032e-09 respectively, which reveal that there is evidence of a dependency between “Pre/Post Test” and “Condition” and a dependency between “Units” and “Condition” (Fig. 5). This finding is contradictory to the claim of the data provider that the sample is representative and are randomly assigned to the treatment or control group, which makes it impossible to conclude whether the intervention had any effect. Random assignment is essential to internal validity or the extent to which the researcher can make causal claims about the effect of the intervention. Non-random assignment leads to non-equivalent groups, meaning that the effect of the intervention might be caused by the groups being different at the outset rather than different at the end as a result of the intervention (“Conduct and Interpret a Logistic Regression”). As a result, we cannot conclude that the change in Condition directly caused a change in Score, but we can say there is an association between the variables “Condition” and “Score.”



*Figure 5: Bar Chart of variable “Units” (e.g.: unit 7 or unit 9) by “Condition” (e.g.: control or treatment)*

To investigate the causal effect relationship, a logistic regression of the form log(odds)=β0+β1∗x1+...+βn∗xn was fit to model the outcome variable (Score) conditional on the predictors (Teacher, Condition, Units, Pre/ Post Test, A/B Form). Logistic regression is used to predict the probability of an outcome that has two values (“Difference between Random Selection and Random Assignment”). In this case, the binary dependent variable is “Score” and the potential outcomes are 1 and 0. Variable “Item” is removed from the independent variable list since items on the two test forms (A, B) for both Unit 7 and Unit 9 are not the same and teachers are not going to use exactly the same questions next year. Data was first split into train and test subsets in 0.8 ratio while preserving relative ratios of different labels in the dataset. Then, the general linear model function was used and the family argument is set to be binomial to make the glm() function do logistic regression, as opposed to some other type of generalized linear model. The summary of the model provides the z-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is o. Comparing each p-value to the preselected value of alpha I chose (alpha= 0.05), I deleted the coefficient having the greatest p-value among others one at a time. Then, I ran the logistic regression without the deleted coefficient again until all the coefficients had a p-value less than 0.05. For example, model 1 shows that variable “Teacher” has p-value at 0.786, which is less than 0.05. The data is not sufficient to prove the variable “Teacher” is significantly different from 0. As a result, the variable “Teacher” was deleted and a logistic regression was fit to model the outcome variable (Score) conditional on the rest of the predictors (Units, Pre/Post Test, and A/B Form). After removing variable “Teacher”, “A/B Form”, “Condition” and “Pre/Post Test”, model 5 was built with the p-value of the intercept and variable “Unit” are both significantly less than 0.05, which are statistically significant. I reject the null hypothesis and conclude that the coefficient is significantly different from 0. The prediction equation is log(p/1-p) = b0 + b1\* Units where p is the probability of getting score 1. Expressed in terms of the variables, the logistic regression equation is log(p/1-p) = -0.5317 + 1.0562 \* Units. The variable “Units” is equal to 1 when it is Unit 9, and 0 when it is Unit 1. It helps to understand that when the item is from Unit 9, the odds of earning score 1 increases by a factor of 1.06. The deviance residuals are close to being centered on 0 and are roughly symmetrical, which is normal.

To calculate the accuracy of the model, I ran the test data through the model and validated the model using confusion Matrix to look at the predictions versus the actual values on the training data. If it is predicted FALSE and the actual value was 0 or it is predicted TRUE and the actual value was 1, the model predicted correctly. Others were predicted incorrectly. The percentage accuracy on the training data is 62.9%. Then, I tested the model by making predictions against the test set and the percentage accuracy rate is 61.7%. From the logistic regression, it is shown that the variable “Condition” is not statistically significant, which means students with the worked examples will not experience benefits to their learning or performance over students without examples (pure problem-solving).

**4. CONCLUSION**

The project uses statistical and graphical methods to answer research questions: 1) which questions students are struggling with the most in both Unit 7 and Unit 9; 2) whether the students with the worked examples will experience benefits to their learning or performance over students without examples (pure problem-solving); and 3) the causal relationship between the dependent variable, Score and the independent variables (e.g., Teacher, Condition, Units, Pre/Post Test, A/B Form, Item).

The results show that in Unit 7, questions 4, 6, 7, 9, and 10 on test form A and questions 3, 4, 6, 7, 9, and 10 on test form B were questions that students struggled with the most. With regard to Unit 9, questions 4 and 9 on test form A and questions 3, 4, 7, 9, and 10 on test form B need special attention. Apart from that, by performing a chi-square test, I found a correlation between “Condition” and “Score” but giving students worked examples does not necessarily directly cause benefits to their learning or performance. In addition, a logistic regression model Score= -0.5317 + 1.0562 \* Units was built to model the outcome variable (score) conditional on the predictor (units) with an accuracy rate of 62.9%.

Although the process has been made in understanding the factors affecting scoring, the accuracy rate of the logistic regression model is not satisfactory. Future research needs to enlarge feature space that introduces more useful covariates, perform feature engineering if needed, bring alternative dataset that has additional information and design the model more widely based on educational insights.

**References:**

Adrian A. Dragulescu and Cole Arendt (2018). xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.1. https://CRAN.R-project.org/package=xlsx

A Multimodal Interface for Solving Equations. *DataShop.* Retrieved from https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=145,

Conduct and Interpret a Logistic Regression. *Statistics Solutions*. Retrieved from https://www.statisticssolutions.com/conduct-interpret-logistic-regression/

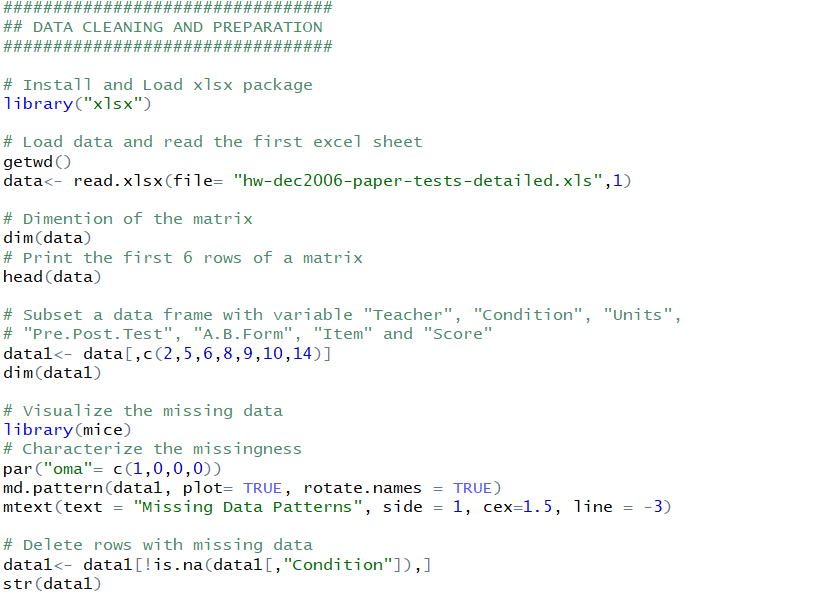
Difference between Random Selection and Random Assignment. *Statistics Solutions*. Retrieved from https://www.statisticssolutions.com/difference-between-random-selection-and-random-assignment/

Jarek Tuszynski (2019). caTools: Tools: moving window statistics, GIF, Base64, ROC AUCetc.. R package version 1.17.1.2. https://CRAN.R-project.org/package=caTools

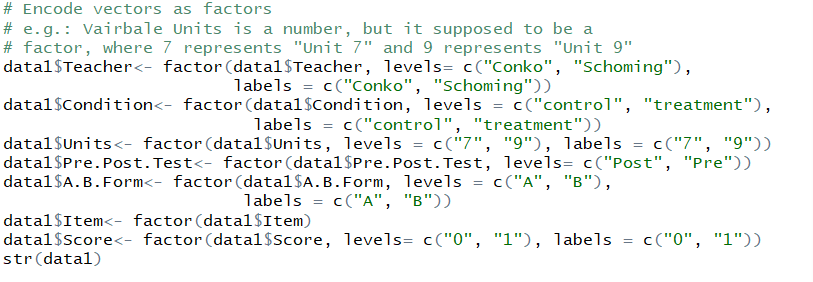
R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

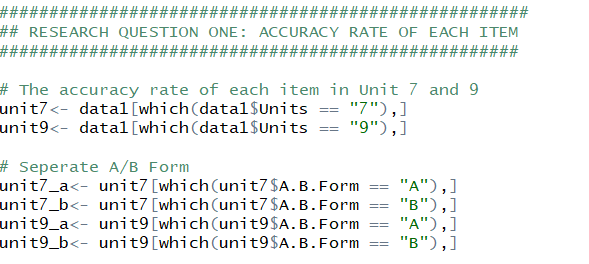
**Appendix A: Data Cleaning and Preparation**



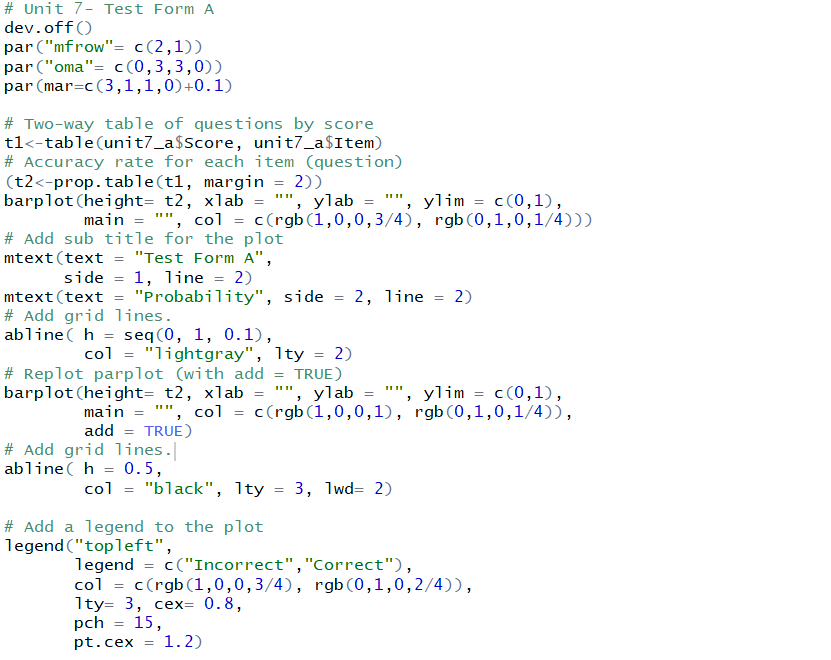
**Appendix B: Encode Vectors as Factors**



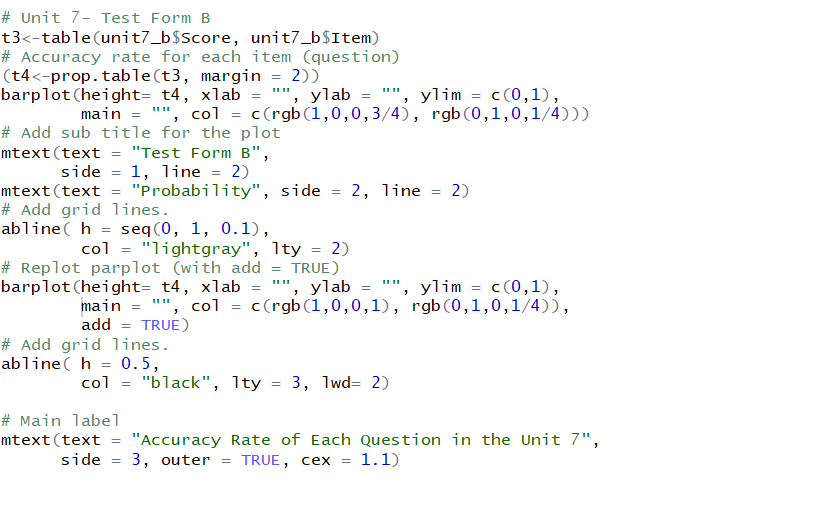
**Appendix C: Data Preparation for Research Question One**



**Appendix D: Accuracy Rate of Each Question in Test Form A for Unit 7**



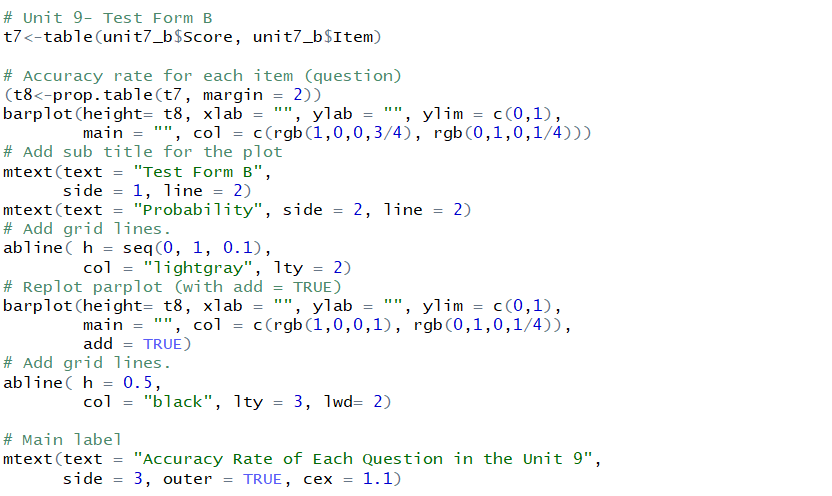
**Appendix E: Accuracy Rate of Each Question in Test Form B for Unit 7**



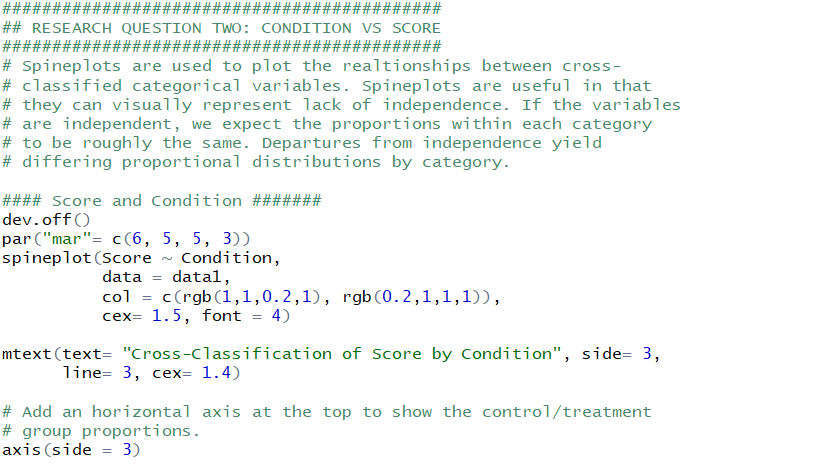
**Appendix F: Accuracy Rate of Each Question in Test Form A for Unit 9**



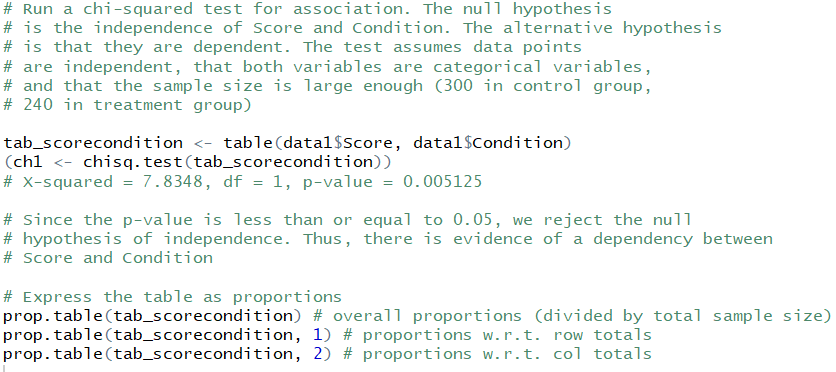
**Appendix G: Accuracy Rate of Each Question in Test Form B for Unit 9**



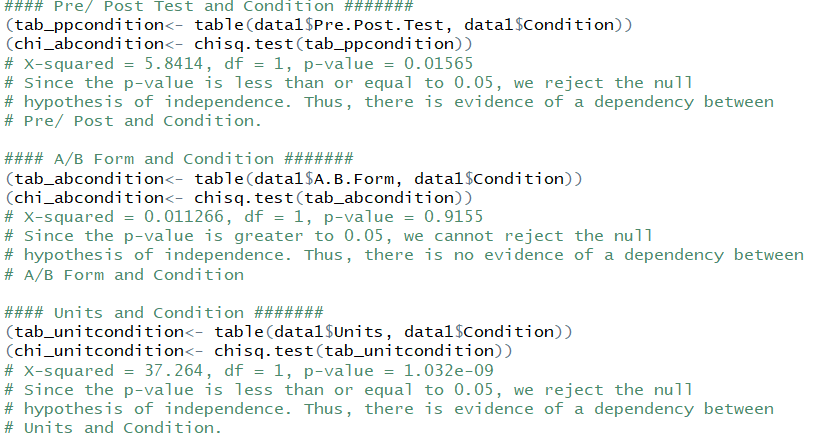
**Appendix H: Cross- Classification of Score by Condition**



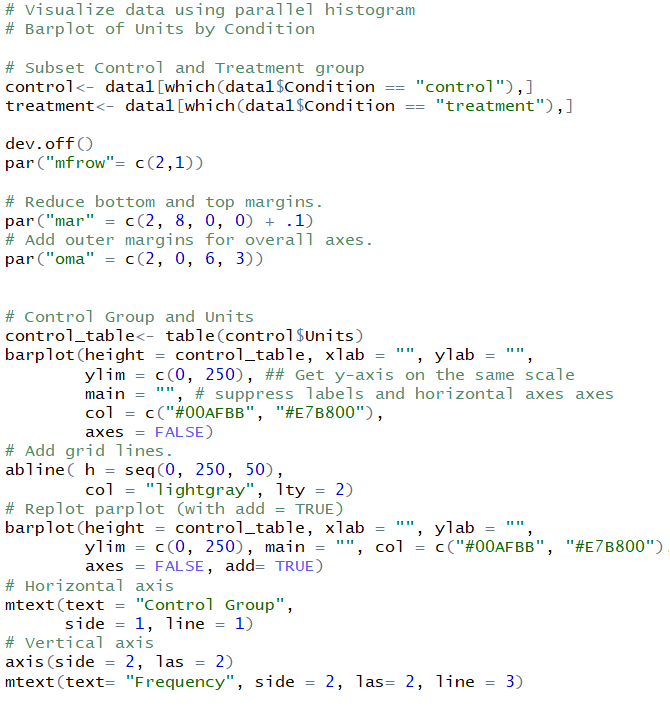
**Appendix I: Chi-Square Test for Score and Condition**



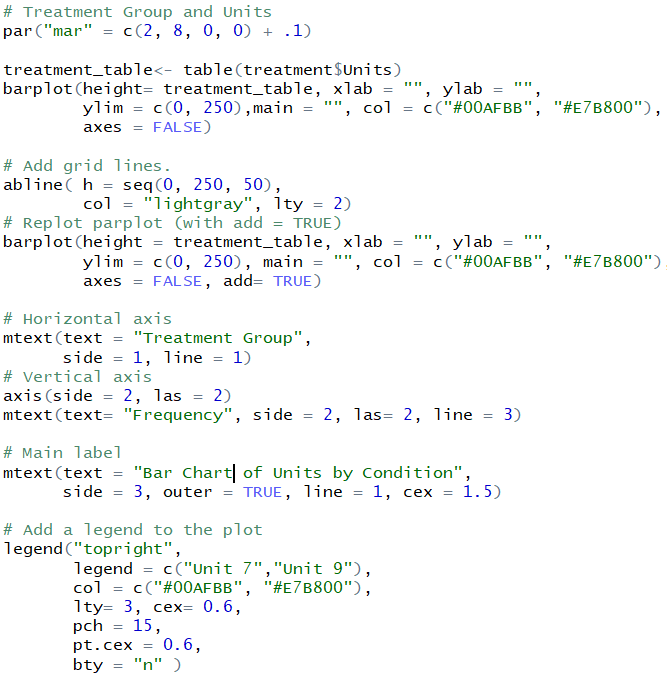
**Appendix J: Chi-Square Test for Association**



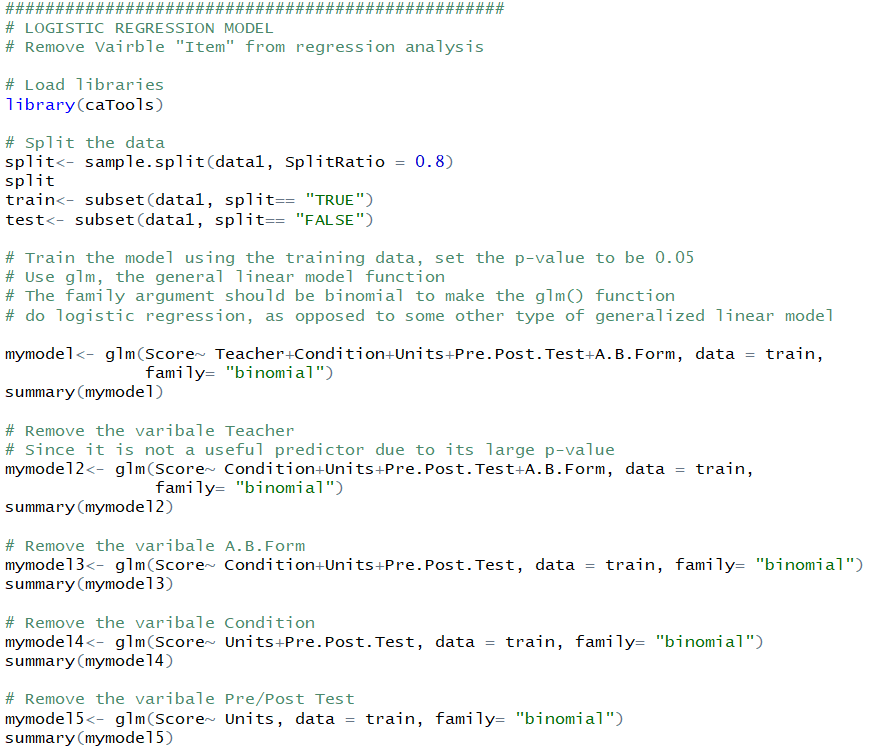
**Appendix K: Bar Chart of Units by Condition (Control Group)**



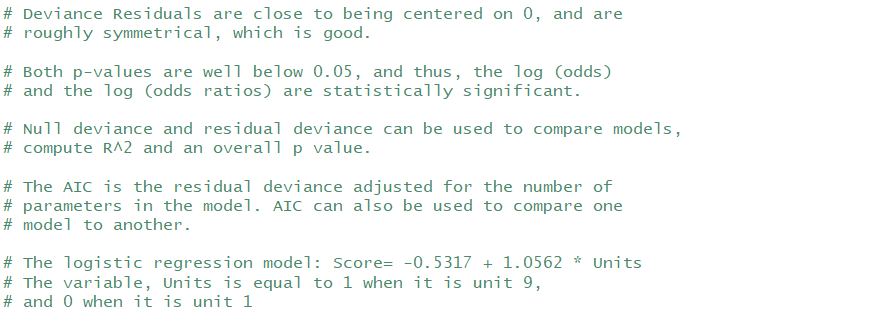
**Appendix L: Bar Chart of Units by Condition (Treatment Group)**



**Appendix M: Logistic Regression Model**



**Appendix N: Logistic Regression Analysis**



**Appendix O: Confusion Matrix and Accuracy Rate of the Logistic Regression Model**

