Milestone 3 Topic: Evaluating Effect of Study Strategies on Learning

Group members: Anantha Rao, Jenna Ji, Desmond Fung, Zhiyi Chen

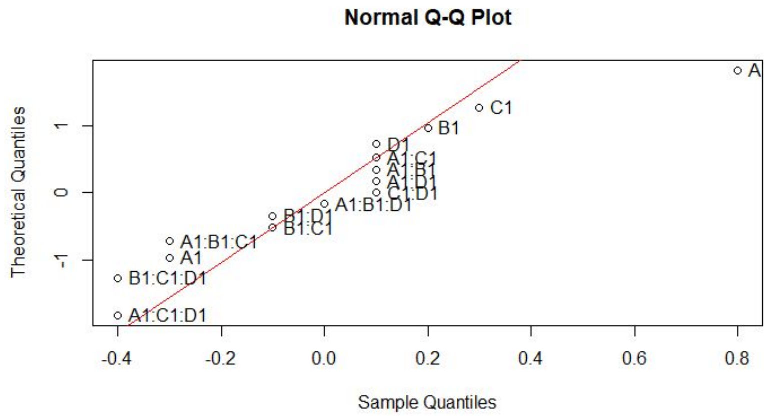
In order to learn the effect of different learning strategies on driver's test scores, we have designed a 2^4 factorial model to evaluate the hypotheses. A total of 16 participants were asked to review learning materials from a driver's education course. We then sent out a google survey to evaluate learning. All the data are collected and stored into a table. Each participant was assigned with either reading or video materials for study (A), a test time of either 5 minutes or 10 minutes (B), a pre-test or not (C), and either possessed a driver's license or didn't (D). We decided to settle on these four factors as they proved to be the most feasible variables to study. In order to ensure the fairness and representativeness of the results, we used the method of blind testing in which participants have no knowledge of which factor combination they are receiving.

Firstly, after factoring the factors in the dataset and computing the final scores to percentage scales, we ran regressions using lm() function in programming language R. In order to determine the effects of factors, we implemented Daniel Plots, as shown below, to visualize which factors are having significant effects on the test scores.

We compacted all our data in a facet grid. It seems like people without a pre-test generally perform worse than those with a pre-test. Another trend we noticed is that people who were given reading material tend to have a higher test score than those given a video.

Further analysis was made to statistically evaluate these observations below.

In the Daniel Plot, we again see that there may be some interesting main effects with effect A and C (video and pre-test). We also see that there appears to be some notable interactions, however we will forego all of these because some of the main effects that comprise them are themselves close to this red normal line. Thus, we invoke the sparsity of effects principle to comfortably dismiss high order effects, and we invoke heritability to dismiss those interactions whose main effects aren’t important. We continued our investigation on factors A and C.



After projecting out the insignificant factors, B and D, we’re left with a 2^2 factorial design with each combination of 4 replicates. We settled on the regression model of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where x_1 represents the study material (reading or video), x_2 represents whether having pre-tests or not, and the β's are the regression coefficients. After summary() and anova(), we got $y = 0.82 - 0.11x_1 + 0.09x_2$. The p values for factors reading/video and pretest/no-pretest are 0.005 and 0.0213. Therefore, at the level of α=0.05, there is significant evidence to conclude that factors reading/video and pretest/no-pretest are significant.

People who were given reading materials tended to have higher test scores than those given a video. We tentatively theorize that this is because people are more concentrated when reading than watching a video, and perhaps paid close attention to the reading. Further explanation should be analyzed via scientific methods. Pre-test seems to have less effect, but still, people who are taking the pre-tests have higher scores than people who don't. This may be because, the pretest primed test takers about the content, or that people who take the tests with pre-tests are more familiar with the form of the tests and thus more prepared.

Stat 424 Final Project

12/3/2020

Names	Video	Test_Time	Pre_Test	Drivers_License		Total score
Jack	-1	1	1	1		16.00 / 20
Lisa	1	-1	1	-1		16.00 / 20
Joel	-1	-1	1	-1		18.00 / 20
Daniel	1	-1	1	1		15.00 / 20
Jose	1	1	1	1		17.00 / 20
Parker	-1	-1	1	1		20.00 / 20
Andy	1	1	1	-1		15.00 / 20
Jessica	-1	1	1	-1		19.00 / 20
Josiah	-1	-1	-1	1		16.00 / 20
Lynda	1	-1	-1	1		14.00 / 20
Summit	1	1	-1	1		16.00 / 20
Zackery	-1	1	-1	1		17.00 / 20
Quinn	1	-1	-1	-1		12.00 /20
Alex	-1	-1	-1	-1		15.00 / 20
Will	-1	1	-1	-1		17.00 / 20
Lucas	1	1	-1	-1		15.00 / 20

Figure 1: An overview of our collected data

```r
# helper functions
code <- function(x) ifelse(x == '-', -1, 1)

daniel_plot <- function(effects) {
  qq <- qqnorm(effects, datax = TRUE)
  qqline(effects, col = "red", probs = c(0.3, 0.7), datax = TRUE)
  text(qq$x, qq$y, names(effects), pos=4)
}
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.0     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## Warning: package 'dplyr' was built under R version 4.0.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(BsMD)
```

```
## Warning: package 'BsMD' was built under R version 4.0.3
```

```r
df <- read.csv("All_Data_FINAL.csv")
df$Total.score <- as.numeric(str_sub(df$Total.score,1,2)) / 20
df<- df%>% rename(A=Video., B=Test_Time, C=Pre_Test., D=Drivers_License.)
df$A <- as.factor(df$A)
df$B <- as.factor(df$B)
df$C <- as.factor(df$C)
df$D <- as.factor(df$D)
# estimate effects

df2 <- df
df2$A <- ifelse(df2$A == -1,"Reading","Video")
df2$B <- ifelse(df2$B == -1,"5min","10min")
df2$C <- ifelse(df2$C == -1,"No Pretest","Pretest")
df2$D <- ifelse(df2$D == -1,"No License","License")

#creating facet grid
ggplot(df2) +
  geom_point(aes(x = A, y = Total.score, col = C)) +
  facet_grid(B ~ D) +
  labs(title = "Visual Overview of Data")
```


```r
#intial lm
fit_coded <- lm(Total.score~A*B*C*D, data=df)
summary(fit_coded)
```

```
##
## Call:
## lm(formula = Total.score ~ A * B * C * D, data = df)
##
## Residuals:
## ALL 16 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.500e-01         NA      NA       NA
## A1          -1.500e-01         NA      NA       NA
## B1           1.000e-01         NA      NA       NA
## C1           1.500e-01         NA      NA       NA
## D1           5.000e-02         NA      NA       NA
## A1:B1        5.000e-02         NA      NA       NA
## A1:C1        5.000e-02         NA      NA       NA
## B1:C1       -5.000e-02         NA      NA       NA
## A1:D1        5.000e-02         NA      NA       NA
## B1:D1       -5.000e-02         NA      NA       NA
## C1:D1        5.000e-02         NA      NA       NA
## A1:B1:C1    -1.500e-01         NA      NA       NA
## A1:B1:D1     4.663e-16         NA      NA       NA
```

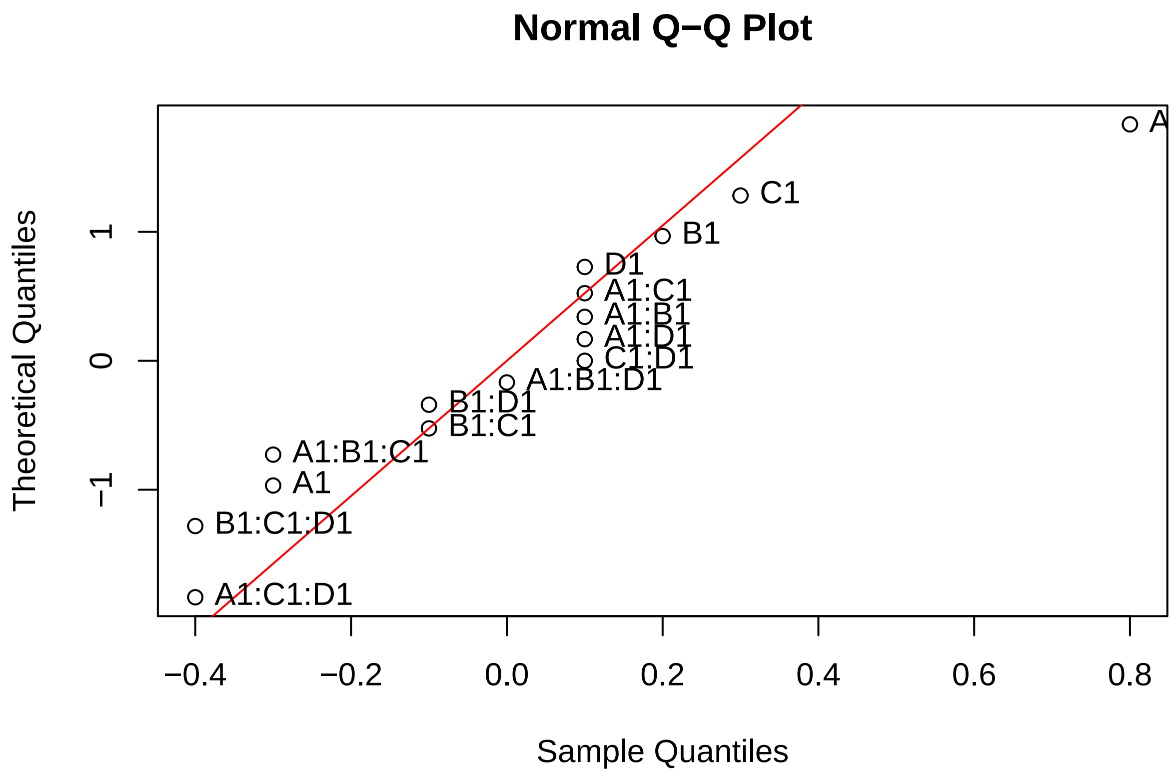
```
## A1:C1:D1    -2.000e-01         NA      NA       NA
## B1:C1:D1    -2.000e-01         NA      NA       NA
## A1:B1:C1:D1  4.000e-01         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 15 and 0 DF,  p-value: NA
```

```r
anova(fit_coded)
```

```
## Warning in anova.lm(fit_coded): ANOVA F-tests on an essentially perfect fit are
## unreliable
```

```
## Analysis of Variance Table
##
## Response: Total.score
##           Df   Sum Sq  Mean Sq F value Pr(>F)
## A          1 0.050625 0.050625
## B          1 0.005625 0.005625
## C          1 0.030625 0.030625
## D          1 0.002500 0.002500
## A:B        1 0.005625 0.005625
## A:C        1 0.000625 0.000625
## B:C        1 0.015625 0.015625
## A:D        1 0.002500 0.002500
## B:D        1 0.002500 0.002500
## C:D        1 0.002500 0.002500
## A:B:C      1 0.000625 0.000625
## A:B:D      1 0.010000 0.010000
## A:C:D      1 0.000000 0.000000
## B:C:D      1 0.000000 0.000000
## A:B:C:D    1 0.010000 0.010000
## Residuals  0 0.000000
```

```r
effects <- 2 * coef(fit_coded)[-1] # exclude intercept
daniel_plot(effects)
```




```r
#after projection
projected_fit <- lm(Total.score~A+C, data=df)
summary(projected_fit)
```

```
##
## Call:
## lm(formula = Total.score ~ A + C, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10625 -0.04375  0.00000  0.04375  0.09375
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.81875    0.02895  28.277 4.63e-13 ***
## A1          -0.11250    0.03343  -3.365  0.00507 **
## C1           0.08750    0.03343   2.617  0.02130 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06687 on 13 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5188
## F-statistic: 9.086 on 2 and 13 DF,  p-value: 0.003397
```

```r
anova(projected_fit)
```

```
## Analysis of Variance Table
```

```
##
## Response: Total.score
##           Df   Sum Sq  Mean Sq F value   Pr(>F)
## A          1 0.050625 0.050625 11.3226 0.005073 **
## C          1 0.030625 0.030625  6.8495 0.021301 *
## Residuals 13 0.058125 0.004471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
effects2 <- 2 * coef(projected_fit)[-1] # exclude intercept
daniel_plot(effects2)
```