Math 338 Final Exam Lab Portion

December 12-19, 2019

You have until 7:00 PM on Thursday, December 19 to complete the exam, Knit this file (click the button that says Knit) and upload the resulting pdf file to Titanium.

You may refer to your notes, your textbook, and any pre-existing online reference (eBook, R/Rguroo help, anything on Titanium). You may search for help online, but you must cite any source found through the search. You may ask Dr. Wynne to clarify what a question is asking for or for technical (R code) help. You may not ask other people for help or use any other resources.

For full credit, show all work except for final numerical calculations (which can be done using a scientific/graphing calculator or R).

ALL BLUE TEXT ARE RESPONSES TO QUESTIONS ASKED ON THIS EXAM

Problem 1

To qualify for the December 2019 debate, a Democratic candidate for President must attract 4% support in four qualifying polls taken between October 16 and December 12.

For the purposes of part (A), you may assume:

- Each participant within a poll is independent
- A candidate's true support is 3%; that is, 3% (0.03) of Democratic voters would vote for this candidate for President in a Democratic primary

If we take a poll of 1000 randomly selected Democratic voters, what is the probability that the candidate has at least 4% support in that poll? [2 pts]

```
n <- 1000
x <- n*0.04
p <- 0.03
pbinom(x, n, p, lower.tail = F) + dbinom(x, n, p)
```

```
## [1] 0.04373438
```

A democratic candidate will have a 4.3734% chance of having at least 4% of the total poll's support.

For the purposes of part (B), you may assume:

- There are a total of 26 qualifying polls between October 16 and December 12
- The probability of earning at least 4% support in a poll is given by your answer to part (A); you may make up an answer to part (A) if you cannot solve it
- Each poll is independent; that is, the result of one poll does not affect any other polls

What is the probability that that this candidate qualifies for the debate by earning at least 4% support in 4 (or more) qualifying polls? [1.5 pts]

```
n <- 1000
x <- n*0.04
p <- 0.03
init_prob <- pbinom(x, n, p, lower.tail = F) + dbinom(x, n, p)

total_polls <- 26
winning_polls <- 4
a <- pbinom(winning_polls, size = total_polls, prob = init_prob, lower.tail = F)
b <- dbinom(winning_polls, total_polls, init_prob)
print(a+b)
```

```
## [1] 0.02531838
```

The probability of this candidate qualifying for the debate given them winning 4 qualifying polls ($P(x \geq 4)$) is 2.531838%.

Problem 2

In ecology and evolutionary biology, it is often important to test whether sex ratios differ from "sex parity" (50% female, 50% male). In a population of riflemen birds, researchers claim that, due to sexual dimorphism, the sex ratio should be 47% female, 53% male.

If the researchers are able to record the sex of 768 independent birds, is this sample size large enough to detect the expected difference from "sex parity"? Why or why not? You can assume a standard Type I Error rate of 5%. [4 pts]

```
# source: Dr. Wynne, Power Analysis for Binomial with R

alpha <- 0.05
n <- 768
p_naught <- 0.5
p_alternate <- 0.47
critical <- qbinom(alpha, n, p_naught)
# pbinom(critical, n, p_naught)
# ^ above is slightly higher than alpha
# go one lower
critical_value <- critical - 1
pbinom(critical_value, n, p_alternate)
```

```
## [1] 0.4870242
```

Given a sample size of 768 riflemen birds, it is not large enough to detect the 3% difference in sex parity. This is because we need a power (β) of at least 80% or better to detect a difference and were given power of 48.70242%.

Problem 3

In a recent study, 91 subjects (all female; 45 adolescents and 46 young adults) participated in a task in which they spent 10 minutes of "study time" either doing boring arithmetic problems or viewing pictures. Each subject participated under two conditions: one in which the pictures were of people ("Social" condition) and one in which the pictures were of landscapes ("Nonsocial" condition).

The pics_vs_math.csv data file on Titanium contains the following variables:

- Age.Group: Adolescent (age 11-17) or Adult (age 23-33)
- S.Time: the approximate number of seconds spent looking at pictures, instead of doing math, in the Social condition (pictures)
- NS.Time: the approximate number of seconds spent looking at pictures, instead of doing math, in the Nonsocial condition (landscapes)
- Time.Diff: the difference in the time spent looking at pictures in the Social vs. Nonsocial condition (this variable is computed for each subject as S.Time – NS.Time)
- S.Rating: how much enjoyment (on a scale from 0-100) the participant got from viewing the pictures in the Social condition (people)
- NS.Rating: how much enjoyment (on a scale from 0-100) the participant got from viewing the pictures in the Nonsocial condition (landscapes)
- Rating.Diff: the difference in enjoyment ratings in the Social vs. Nonsocial conditions (this variable is computed for each subject as S.Rating – NS.Rating)

```
# read in CSV data
data <- read.csv("~/Downloads/pics_vs_math.csv")
```

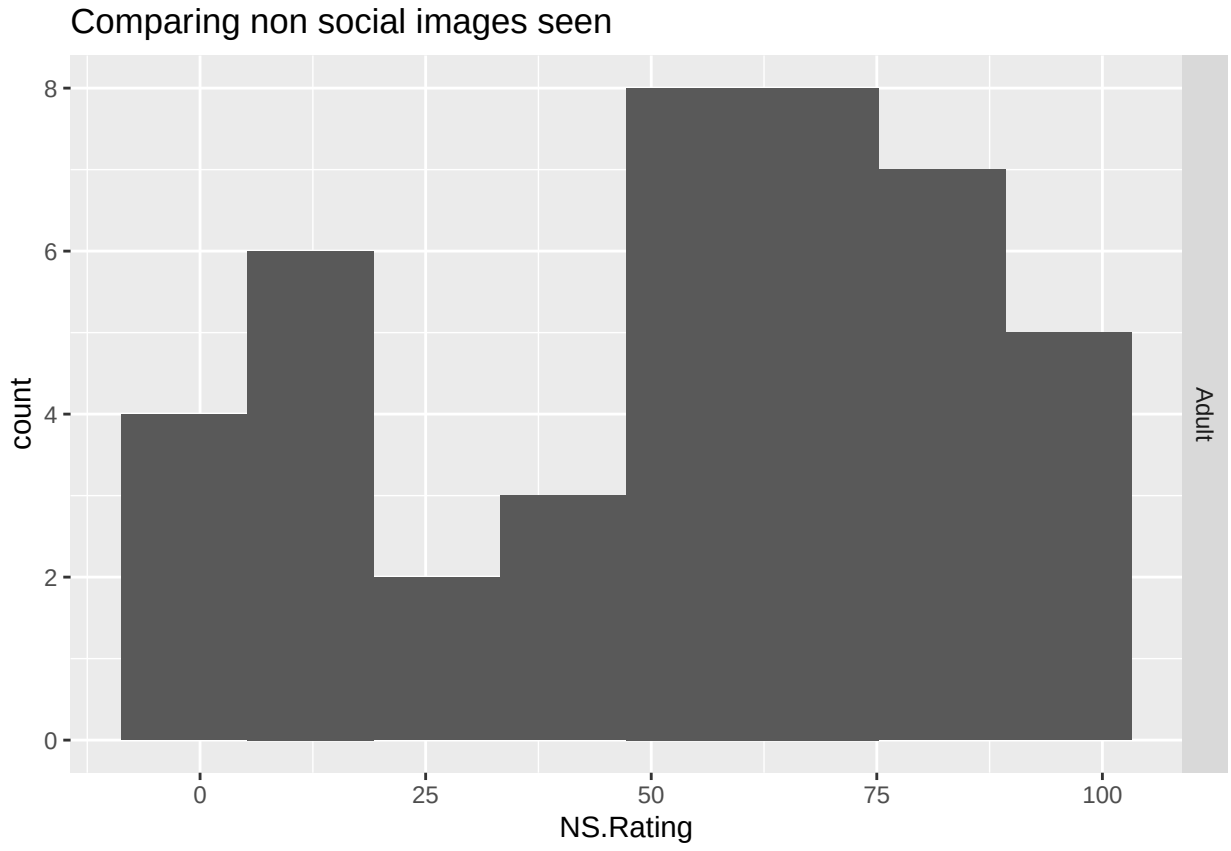
Note that there is some missing data in this file, but we will ignore the missing data.

For this problem, answer one of the two questions. That is, answer EITHER PART A OR PART B BUT NOT BOTH PARTS!!!!! Full credit will be given for the following:

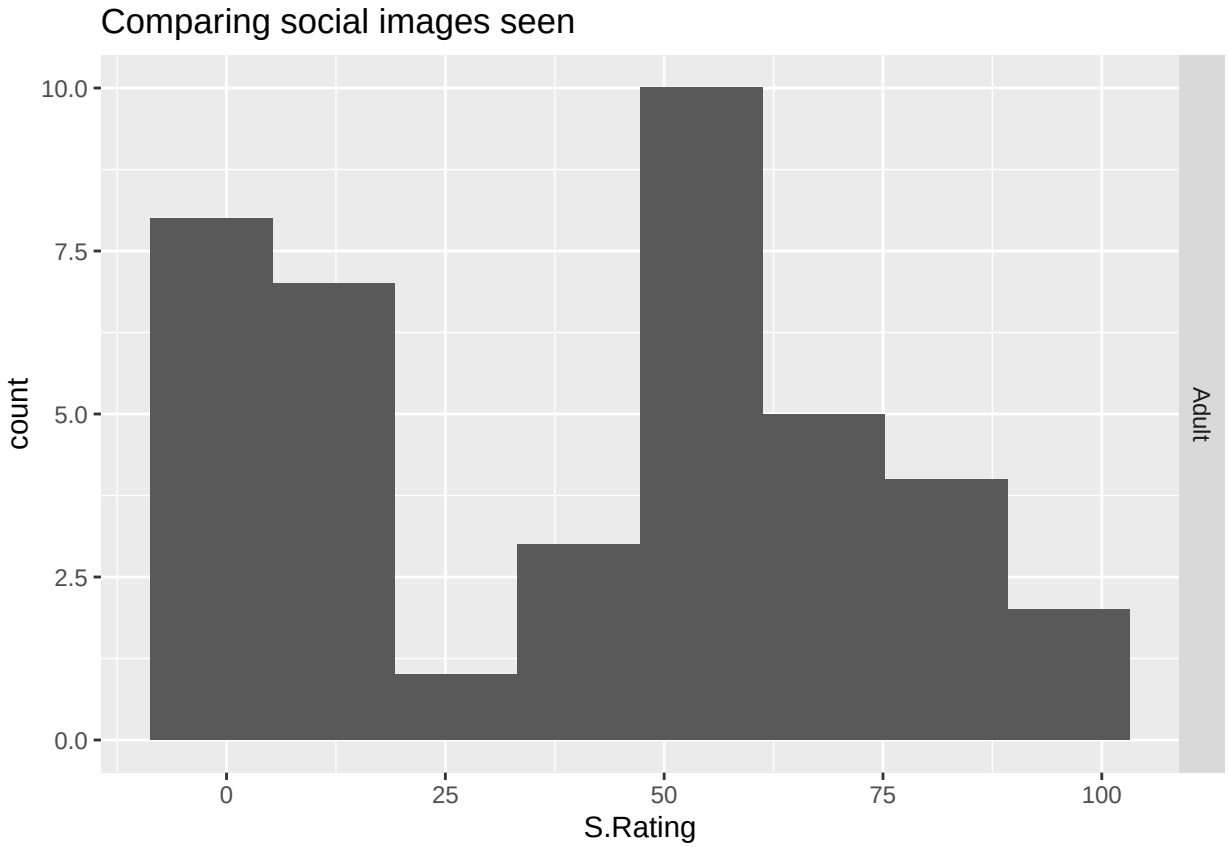
- Correctly identifying the inferential framework (one-sample, matched pairs, two-sample, or one-way ANOVA) suggested by the question. [1 pt]
- Producing one or more graphs that visualize the sample data and checking any necessary assumptions. At least one graph must have a clear connection to both the variable(s) involved in the question and the framework identified. [1.5 pts]
- Correctly performing statistical inference and showing all relevant, necessary (R code and) software output. If a confidence interval is indicated, use a 95% confidence level. If a hypothesis test is indicated, use either a Fisher-type test or NHST (as appropriate) with a 5% significance level. If either type of inference is valid, you can choose. [1 pt]
- Providing a real-world interpretation/conclusion that answers the question in context. Interpretations/conclusions should be supported by the results of your inferential procedure. [2 pts]

Do young women in their 20's and 30's prefer looking at pictures of people to looking at pictures of landscapes? (That is, do they give higher ratings to the pictures of people compared to the pictures of landscapes?) HINT: you may need to subset your data to include only the adult group.

```
# shut up about namespace
library(ggplot2, warn.conflicts = F)
library(dplyr, warn.conflicts = F)
data <- read.csv("~/Downloads/pics_vs_math.csv")
# filter out the people who are older than 20
subset <- data %>% filter(Age.Group == "Adult")

# non social plot

a <- ggplot(subset, aes(x = NS.Rating))
b <- a + geom_histogram(binwidth = 14.0, center = 12.25)
ns_plot <- b + facet_grid(Age.Group ~ .) + labs(title = "Comparing non social images seen")
suppressWarnings(print(ns_plot))
```

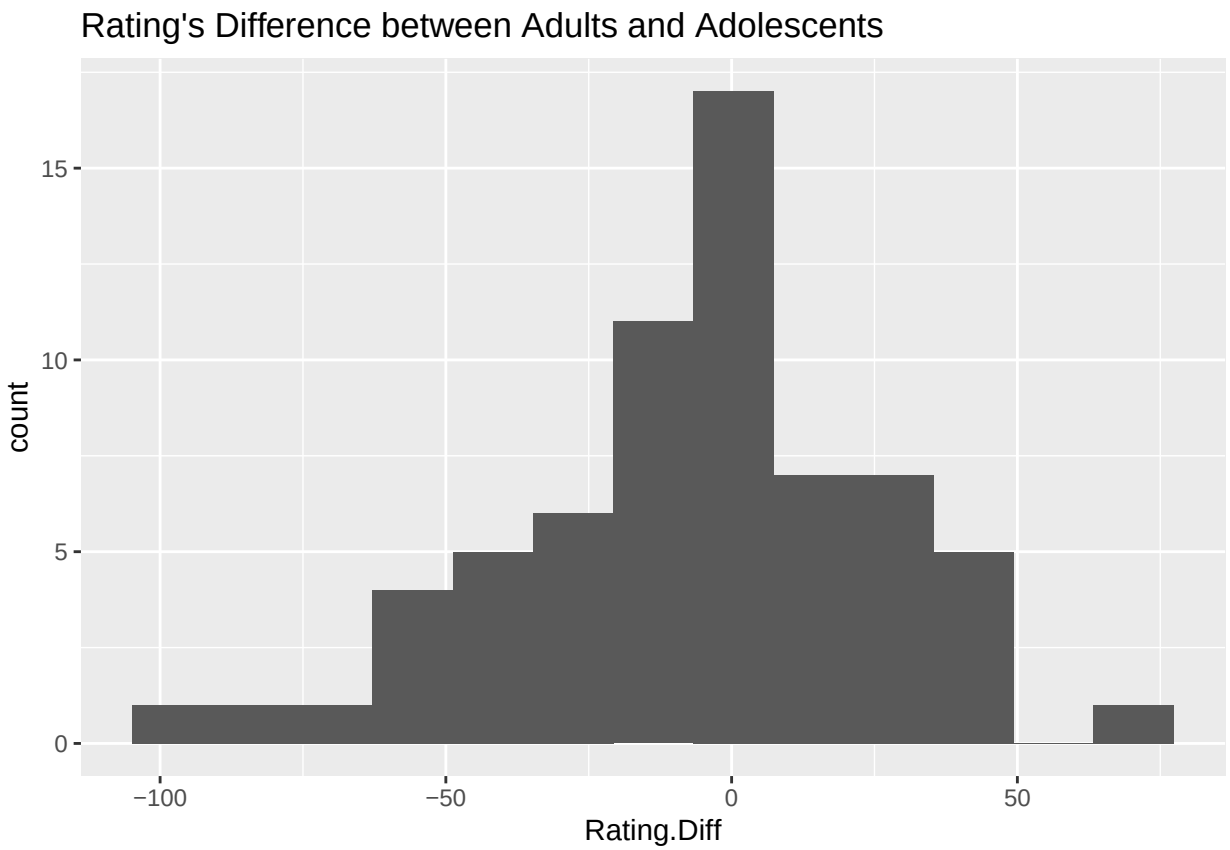



```
# social plot

e <- ggplot(subset, aes(x = S.Rating))
f <- e + geom_histogram(binwidth = 14.0, center = 12.25)
social_plot <- f + facet_grid(Age.Group ~ .) + labs(title = "Comparing social images seen")
suppressWarnings(print(social_plot))
```



The framework I chose was matched pairs because you have two objects coming from a given source. In this case it is adolescents and adults enjoyment level of seeing two sets of images.

```r
library(ggplot2, warn.conflicts = F)
library(dplyr, warn.conflicts = F)
data <- read.csv("~/Downloads/pics_vs_math.csv")
subset <- data %>% filter(Age.Group == "Adult")

a <- ggplot(data, aes(x = Rating.Diff))
b <- a + geom_histogram(binwidth = 14.0, center = 0.25)
diff_plot <- b + labs(title = "Rating's Difference between Adults and Adolescents")

suppressWarnings(print(diff_plot))
```




```r
# seen from Lab 22

t.test(x = subset$NS.Rating, y = subset$S.Rating, paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  subset$NS.Rating and subset$S.Rating
## t = 2.8141, df = 37, p-value = 0.00779
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.494368 27.610895
## sample estimates:
## mean of the differences
##                16.05263
```

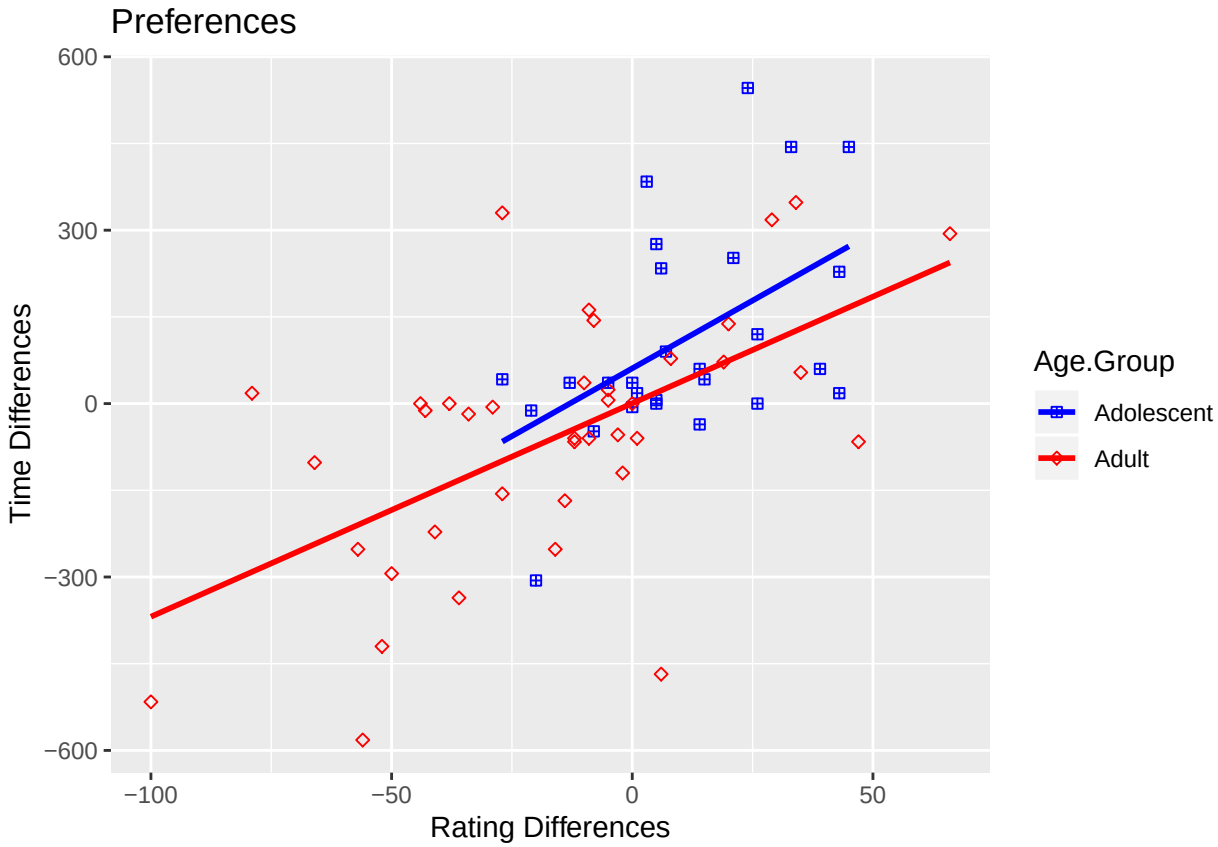
I am 95% confident that the approval rating are between 4.494368 and 27.610895 compared between non social image ratings and social image ratings for adults. Since out confidence interval does not include the value of 0, we can accurately determine the average approval rating difference between social images and non social images in the adult subset of the sample.

Problem 4

66 of the 91 subjects in the study from Problem 3 spent time looking at both pictures of people and pictures of landscapes. Researchers were interested in determining whether how much people preferred the people to the landscapes (Rating.Diff) affected how much more time they spent looking at the people compared to the landscapes (Time.Diff).

Create a scatterplot to visualize this data. Think carefully about which variable is the explanatory (predictor) variable and which is the response variable. Indicate on the scatterplot which subjects were adults and which were adolescents. You do not need to include any regression lines. [1.5 pts]

```
library(ggplot2, warn.conflicts = FALSE)
library(dplyr, warn.conflicts = FALSE)

data <- read.csv("~/Downloads/pics_vs_math.csv")

a <- ggplot(data = data, mapping = aes(
        x = Rating.Diff, y = Time.Diff,
        color = Age.Group, shape = Age.Group
))
b <- a + geom_point()
c <- b + labs(title = "Preferences", x = "Rating Differences", y = "Time Differences")
plot_color <- c + scale_color_manual(
        values = c("blue", "red")) + scale_shape_manual(values = c(12, 23)
)
final_plot <- plot_color + geom_smooth(method = "lm", se = FALSE)
suppressWarnings(print(final_plot))
```



Tell me one interesting thing about this dataset after looking at the scatterplot that wasn't obvious before looking at the plot. Write in context; full points will be awarded for a clear real-world takeaway rather than use of statistical jargon. [1 pt]

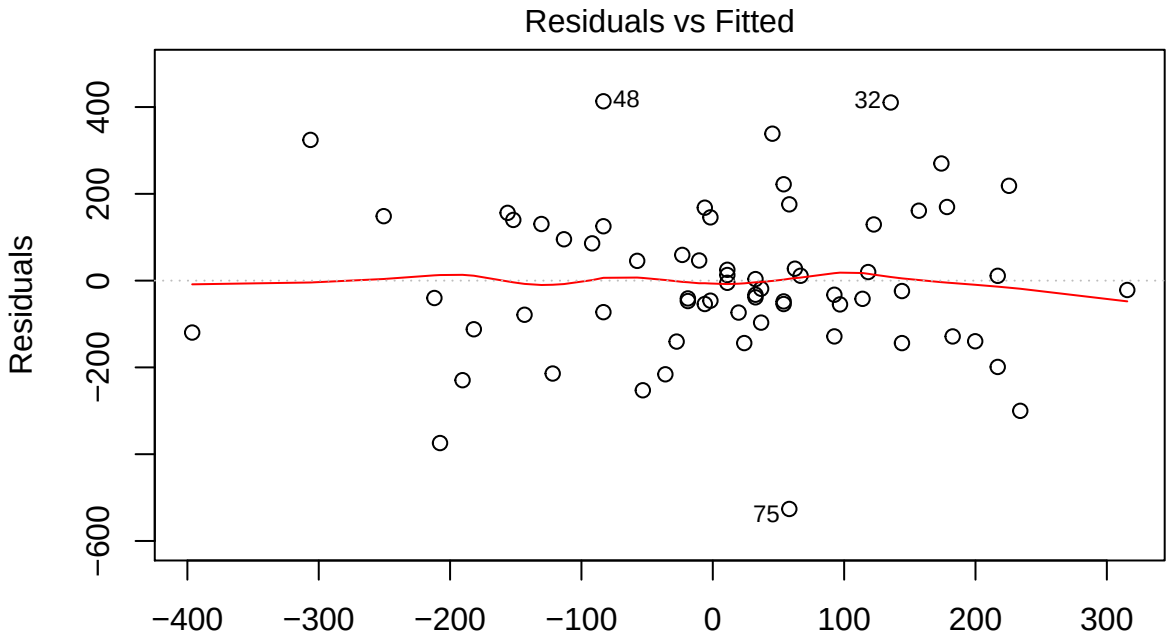
There seems to more plots for the adult group and and this implies that adults are more inclined to answer surveys. There is also a difference in the slope of the blue line and the red line, which seems to be in directly proportional to the amount of plots.

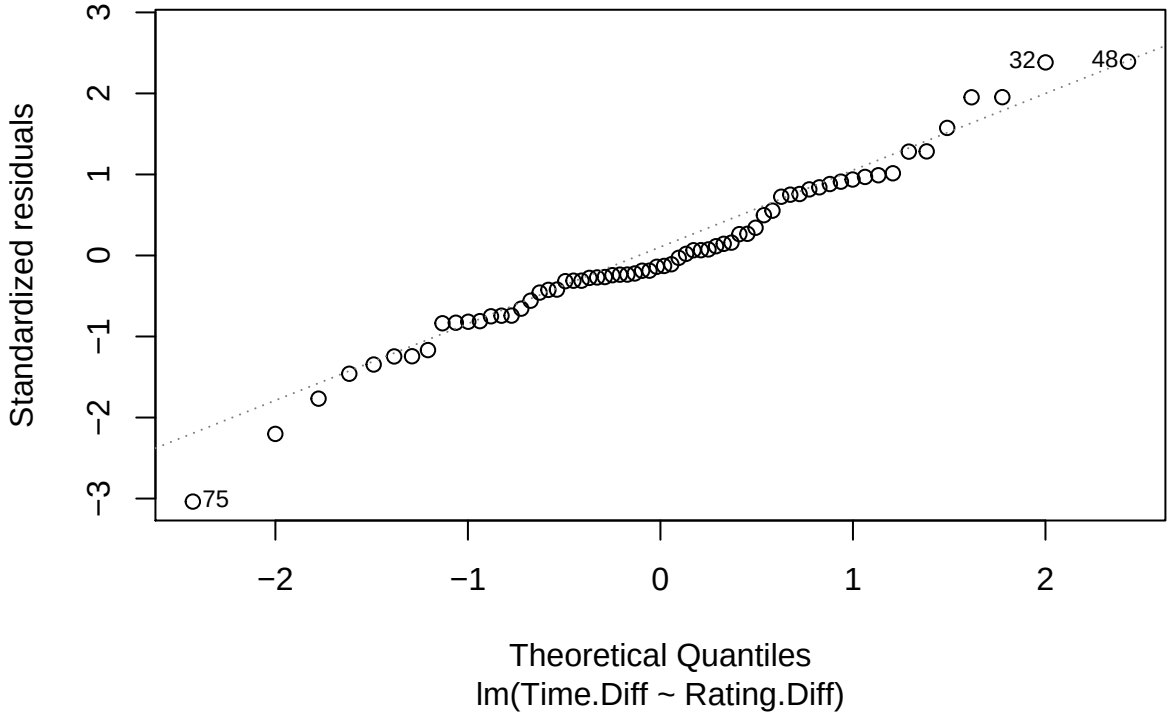
Fit a linear regression model using the entire sample (regardless of group). Include below the coefficient estimates table, the residual plot, and the normal quantile (q-q) plot. [1 pt]

```
# Coefficient Estimates Table
data <- read.csv("~/Downloads/pics_vs_math.csv")
model_one <- lm(Time.Diff ~ Rating.Diff, data = data)
coef(model_one)
```

```
## (Intercept) Rating.Diff
##   32.508219    4.288587
```

```
# Residual plot and Q-Q Plot (in order)
plot(model_one, which = c(1, 2))
```


Residuals vs Fitted
Residuals
Fitted values
lm(Time.Diff ~ Rating.Diff)
48
32
75
400
200
0
−200
−600
−400
−300
−200
−100
0
100
200
300


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Time.Diff ~ Rating.Diff)
32
48
75
3
2
1
0
−1
−2
−3
−2
−1
0
1
2

Obtain a 95% confidence interval for the mean difference in time spent looking at pictures of people compared to landscapes for someone who finds the two types of pictures equally enjoyable (that is, Rating.Diff = 0). Does this suggest that, on average, people with no preference tend to "slack off" more looking at one or the other type of picture? If so, which type? [2 pts]

```
# Confidence interval when Rating.Diff = 0
data <- read.csv("~/Downloads/pics_vs_math.csv")
model_one <- lm(Time.Diff ~ Rating.Diff, data = data)
new_data_frame <- data.frame(Rating.Diff = c(0))

confidence_interval <- predict(model_one, newdata = new_data_frame,
        interval = "confidence", level = 0.95)
data_confidence_level_df <- data.frame(new_data_frame, confidence_interval)

print(data_confidence_level_df)
```

```
##   Rating.Diff      fit       lwr      upr
## 1           0 32.50822 -11.01237 76.02881
```

Since our confidence interval includes the value of 0, we cannot accurately determine if there is a preference to slack off more look at one or the other type of picture.