S&DS 230 Final Project: Sleep Amongst Healthcare Workers During the Pandemic

'NA'

5/15/2021

Part 1: Introduction

The Covid-19 pandemic has undoubtedly had a large toll on people across the world, but one of the groups impacted the most is healthcare workers. The pandemic has brought stress about hospital room shortages, supply shortages, and personnel shortages, and health care workers bear the brunt of all this responsibility. In this project, our group aimed to analyze the best predictors for the amount of sleep workers got in April 2020, right after the pandemic hit America extremely hard and the entire nation went into lock down. With stories of health care workers working 48 hour or longer shifts, our group wanted to see what factors could lead to health care workers receiving differing amounts of sleep during the pandemic. Sleep is connected to the physical well-being of humans, so our group assumed that loss of sleep would mean that health care workers' physical health was impacted (lack of sleep can lead to fatigue, muscle pain, etc.). This model can be useful in predicting which groups of healthcare workers will be hit the hardest (and lose the most sleep) if the pandemic were to rebound, as well as allow health care workers to look back upon their lack of sleep during the early stages of the pandemic and find out what factors led their physical health to deteriorate.

Part 2: Data

All variables were renamed from the original column names for the sake of increased readability and understanding.

Variable Information:

- Age: data from Q1, age in years, continuous
- Gender: from Q2, categorical
- state: data from Q3, categorical
- SleepJan: data from Q12, Average hours of sleep in January 2020, continuous
- SleepApril: data from Q13, Average hours of sleep in last week (April 2020), continuous
- socialMedia: data from Q24, Hours consuming COVID-19 related social media and/or news on average per day, categorical
- ChildrenHome: data from Q11a, Number of children in home, continuous

Reading in the Dataset

```
sleep <-
read.csv("https://raw.githubusercontent.com/peanutsalad/SDS230Final/main/COVIDSleep.csv", as.is = TRUE)
dim(sleep)

## [1] 915  66

#delete first row which is just a label
sleep <- sleep[-1,]
```

Data Cleaning

Cleaning the Age

When cleaning the age variable, after doing some cleaning with gsub, all data values were converted to numeric form. Cleaning for this variable was relatively simple. We simply looked at the data in this variable and cleaned up the few strange entries.

```
sleep$Q1 <- gsub("50-60", "55", sleep$Q1)
sleep$Q1 <- gsub(" years", "", sleep$Q1)
sleep$Q1 <- gsub("seventy five", "75", sleep$Q1)
sleep$Q1 <- gsub(" +", "", sleep$Q1)
sleep$Q1[sleep$Q1 == ""] <- NA
sleep$Q1 <- as.numeric(sleep$Q1)
sleep$Age <- sleep$Q1
```

Cleaning Function for Sleep Time

When cleaning the average amount of sleep received, we made assumptions for any data values that are 7-7.5, 8-8.5, 8.5-9 etc. could be rounded down to the lower value of the range. A function was created to clean the sleep values for both a week in April 2020 and for the month of January 2020. Cleaning sleep times were especially challenging because there were many varying inputs that resulted in different averages.

```
cleandata <- function(question) {
  question <- gsub("3-4.*","3.5", question)
  question <- gsub("4-5.*|4 to 5.*", "4.5", question)
  question <- gsub("5-6.*", "5.5", question)
  question <- gsub("6-7.*|5 to 8.*|6 to 7.*|6 and 1/2|5-8.*", "6.5",
question)
  question <- gsub("7-8.*|7 to 8.*|7:30|7 1/2|5-10.*", "7.5", question)
  question <- gsub("8-9.*|7-10.*|5-12.*", "8.5", question)
  question <- gsub("9-10.*", "9.5", question)
  question <- gsub("3-5.*", "4", question)
  question <- gsub("4 to 6.*|4-6.*", "5", question)
  question <- gsub("3 nights 5.5 hours 2nights 6.5 hours|5-7", "6", question)
  question <- gsub("6-8.*|7-7.5.*|seven|7:15", "7", question)
  question <- gsub("8-8.5.*|7-9.*", "8", question)
```

```
  question <- gsub("8-10.*", "9", question)
  question <- gsub("10-12", "11", question)
  question <- gsub("6.5-7", "6.5", question)
  question <- gsub("7.25-7.5", "7.5", question)
  question <- gsub("2 .*", "2", question)
  question <- gsub("4 .*", "4", question)
  question <- gsub(".* 5.*|5 .*|5, .*", "5", question)
  question <- gsub("6 .*|6-6.5", "6", question)
  question <- gsub("7 .*", "7", question)
  question <- gsub("8 .*", "8", question)
  question <- gsub("9 .*", "9", question)
  question <- gsub("10 .*", "10", question)
  question <- gsub("12 .*|12\\+", "12", question)

}
```

Cleaning Average Hours of Sleep on a Work Night in April 2020

One outlier with 15 hours of sleep was excluded to not skew analysis. We utilized our date cleaning function to clean this variable, and then converted it to numeric form and renamed it.

```
sleep$Q13 <- cleandata(sleep$Q13)
sleep$Q13 <- as.numeric(sleep$Q13)
sleep$Q13[sleep$Q13 > 15] <- NA
sleep$SleepApril <- sleep$Q13
```

Cleaning Average Hours of Sleep on a Work Night in January 2020

We utilized our date cleaning function to clean this variable, and then converted it to numeric form and renamed it.

```
sleep$Q12 <- cleandata(sleep$Q12)
sleep$Q12 <- as.numeric(sleep$Q12)
sleep$SleepJan <- sleep$Q12
```

Cleaning Average Hours Consuming COVID-19 Related Social Media

Blank responses were labeled as NA. Strange characters (â€) appeared in some responses were removed using gsub.

```
sleep$Q24[sleep$Q24 == ""] <- NA
sleep$Q24 <- gsub(" â€“ ","-", sleep$Q24)
sleep$socialMedia <- sleep$Q24
table(sleep$socialMedia)

##
##   0-.5 hour 0.51-1 hour   1-2 hours   2-3 hours    3+ hours
##         134         251         278         119          56
```

Cleaning State Data

The state data did not require much cleaning, as it likely came from a multiple choice question. The only adjustment made (besides labeling blank responses as NA) was to clarify the "I do not reside in the United States" responses as being outside of both the US and Puerto Rico, as Puerto Rico was one of the included "states".

```
sleep$state <- sleep$Q3
sleep$state <- gsub("I do not reside in the United States", "outside of
U.S./Puerto Rico", sleep$state)
sleep$state[sleep$state == ""] <- NA
table(sleep$state)
```

```
##
##                     Alabama                     Arizona
##                           3                           2
##                    Arkansas                  California
##                           1                          15
##                    Colorado                 Connecticut
##                           5                           2
##                     Florida                     Georgia
##                          10                           7
##                       Idaho                    Illinois
##                           1                          12
##                     Indiana                        Iowa
##                           3                           2
##                      Kansas                    Kentucky
##                           2                           1
##                   Louisiana                    Maryland
##                           1                           4
##               Massachusetts                    Michigan
##                           7                         677
##                   Minnesota                 Mississippi
##                           6                           4
##                    Missouri                    Nebraska
##                           4                           2
##               New Hampshire                  New Jersey
##                           1                           2
##                  New Mexico                    New York
##                           1                          13
##              North Carolina                        Ohio
##                           6                          15
##                    Oklahoma                      Oregon
##                           1                           3
## outside of U.S./Puerto Rico                Pennsylvania
##                          33                          11
##                 Puerto Rico              South Carolina
##                           2                           3
##                   Tennessee                       Texas
##                           5                          11
##                        Utah                     Vermont
##                           6                           2
```

```
##                    Virginia                  Washington
##                           4                           4
##               West Virginia                   Wisconsin
##                           3                           6

#It is important to note that there are many more responses from Michigan
compared to other states. Some states only have one respondent.
```

Cleaning Data on Number of Children at Home

It seems that question 11a was only added after the first 10 or so respondents, as there is no data for question 11a for the first 10-15 respondents. Additionally, these first respondents answered question 11 with numbers instead of yes or no, leading us to believe that the original wording of question 11 was how many children there were in the home. Thus, we are manually inputting the data from these first respondents into the data of Q11a to ensure that all the relevant data show up in the same variable.

```
sleep$ChildrenHome <- sleep$Q11a
sleep$ChildrenHome[3] <- 1
sleep$ChildrenHome[4] <- 1
sleep$ChildrenHome[6] <- 2
sleep$ChildrenHome[7] <- 1
sleep$ChildrenHome[8] <- 1
sleep$ChildrenHome[9] <- 2
sleep$ChildrenHome[10] <- 2
#Assuming that people who answered no or not applicable to Question 11
(whether there were children at home) have zero children at home.
sleep$ChildrenHome[sleep$Q11 == "No"] <- 0
sleep$ChildrenHome[sleep$Q11 == "Not applicable"] <- 0
sleep$ChildrenHome[sleep$ChildrenHome == ""] <- 0
sleep$ChildrenHome <- as.numeric(sleep$ChildrenHome)
```

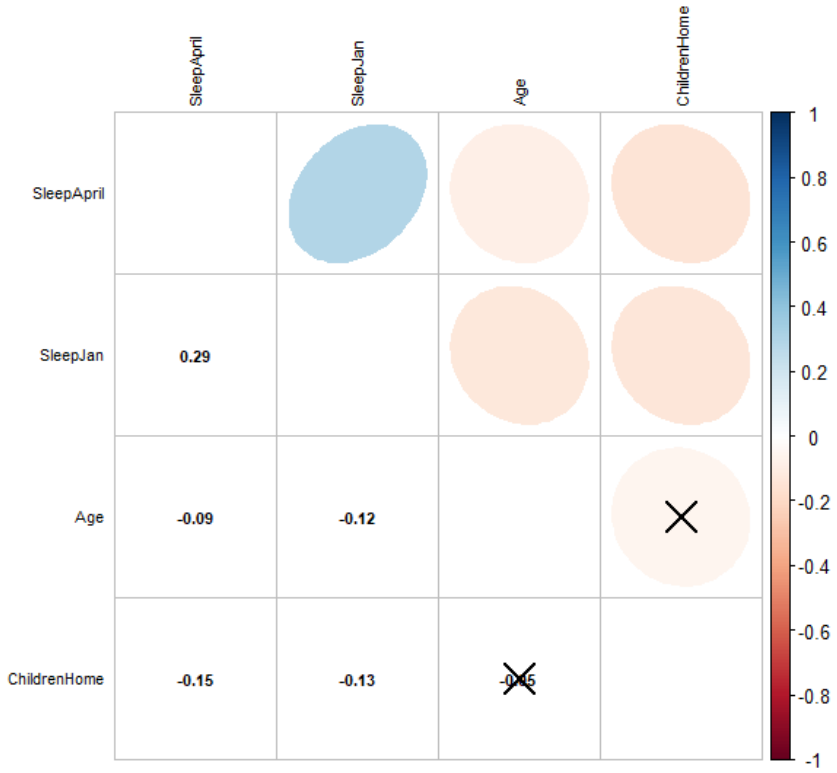
Cleaning Gender Data

This variable did not require much cleaning. We renamed it for clarity.

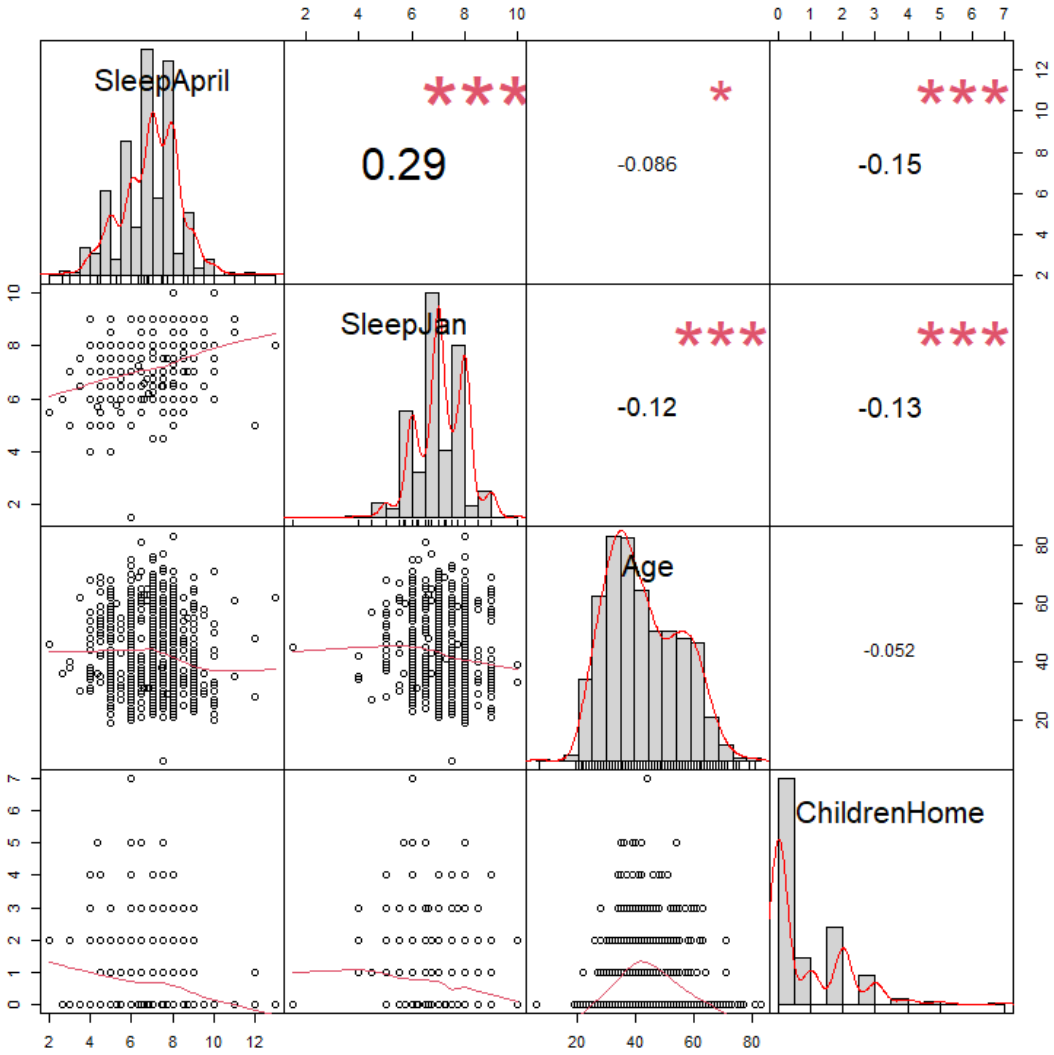
```
sleep$Gender <- sleep$Q2
```

Correlation Plots

```
sleep2 <- sleep[, c("SleepApril", "SleepJan", "Age", "ChildrenHome")]

sleep2 <- sleep2[complete.cases(sleep2), ]

sigcorr <- cor.mtest(sleep2[,], conf.level = .95)

par(xpd = TRUE)
corrplot.mixed(cor(sleep2[,]), lower.col = "black", upper = "ellipse", tl.col
= "black", number.cex = .7,
               order = "hclust", tl.pos = "lt", tl.cex=.7, p.mat = sigcorr$p,
sig.level = .05, mar = c(1, 1, 5, 1))
```




```
chart.Correlation(sleep2[,], histogram = TRUE, pch = 19)
```



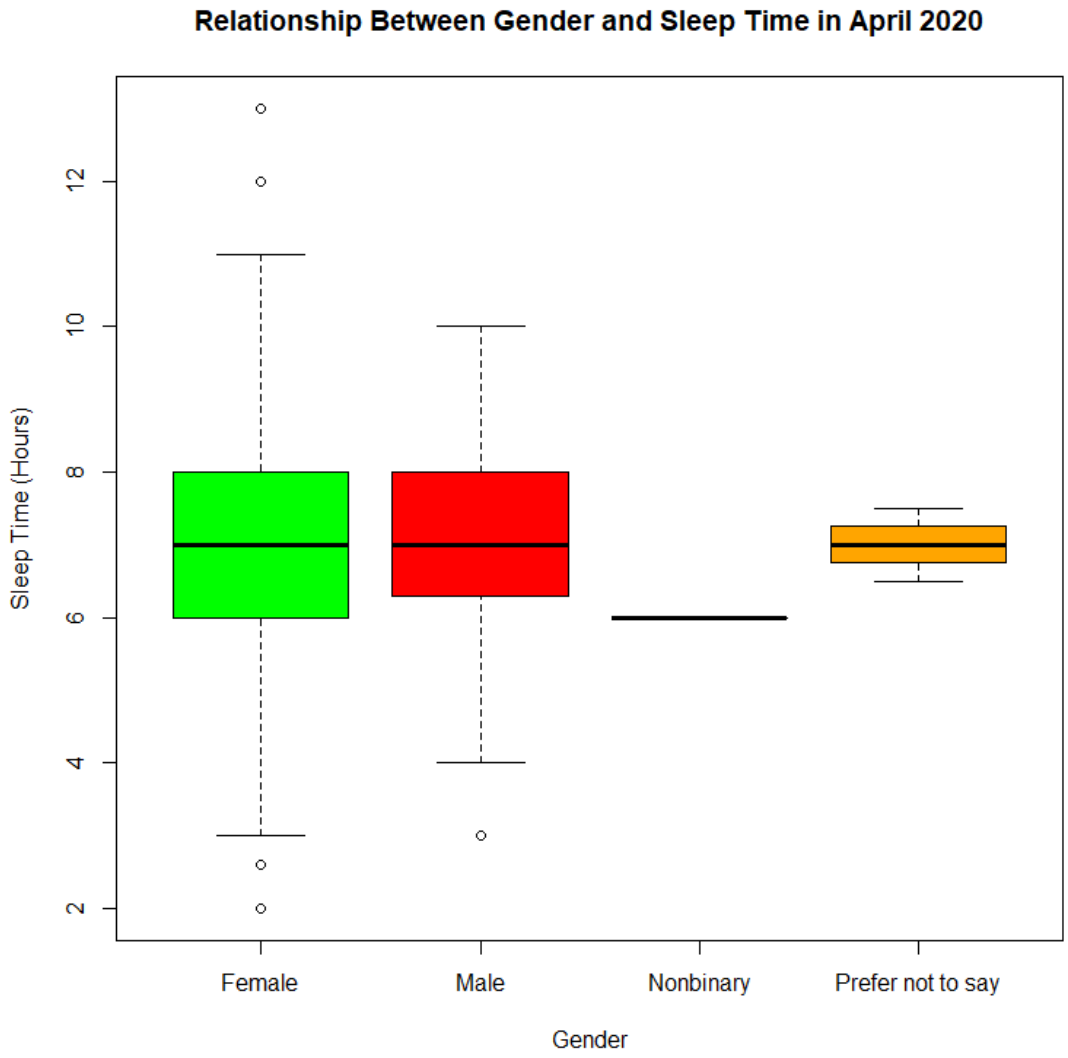
The corrplot shows that all correlations appear to be significant except for the correlation between the number of children at home and the age of the survey participant. The histograms appear to be slightly right skewed for children at home and age. THe histogram for average hours of sleep for the week in April 2020 appears normally distributed. The histogram is slightly left skewed for average hours of sleep in January. None of the correlations seem to be extremely linear, as the scatterplots seem to be more like blobs than lines.

Case Study 1: Gender's Impact on Sleep in a Week in April 2020

Boxplot

```
sleep$Q2[sleep$Q2 == ""] <- NA

boxplot(sleep$Q13 ~ sleep$Q2, data = sleep, col = c("green", "red", "blue",
"orange"),
```

```
    main = "Relationship Between Gender and Sleep Time in April 2020",
    xlab = "Gender", ylab = "Sleep Time (Hours)")
```



The boxplot shows that values do not differ greatly for sleep time in April 2020 based on gender.

T-test

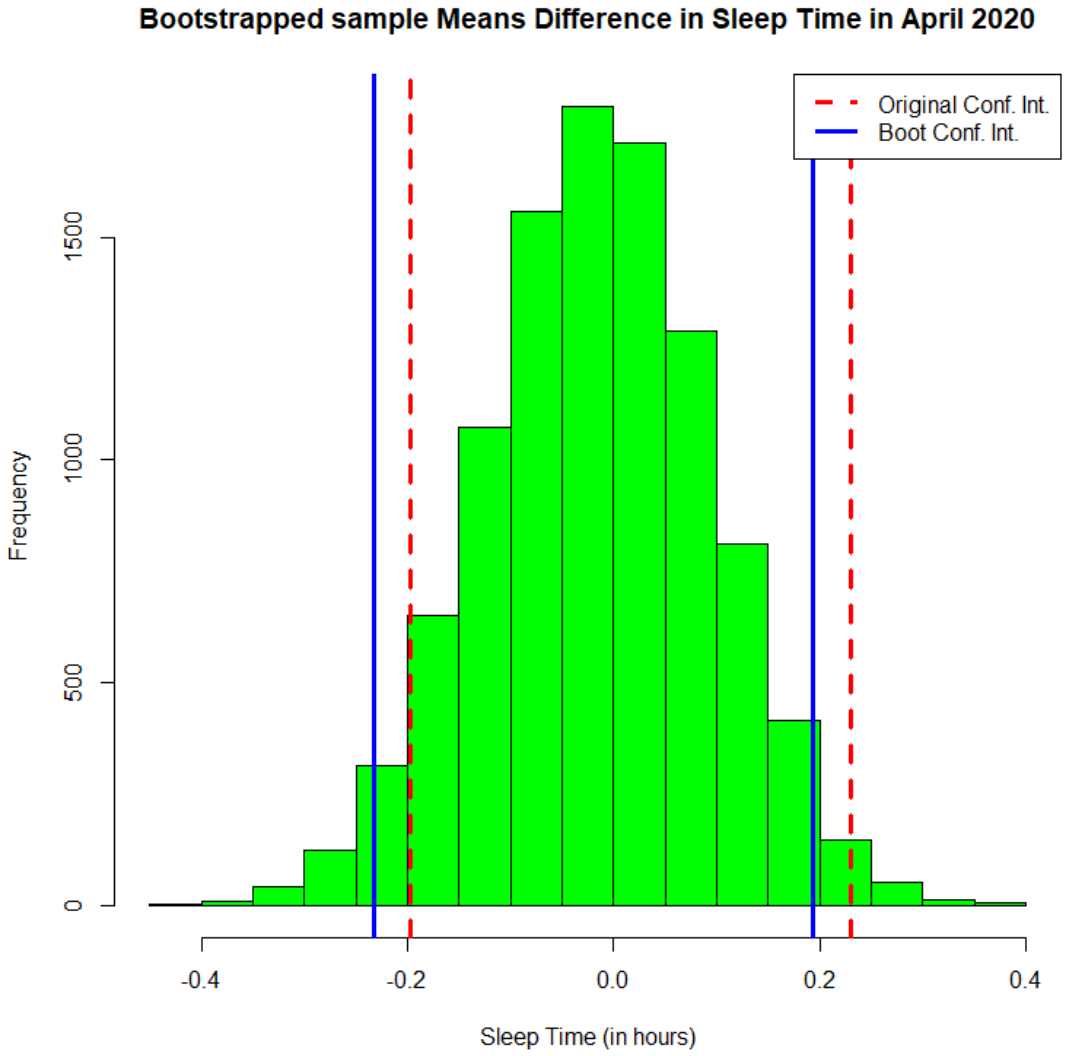
```
sleep3 <- na.omit(sleep[, c("SleepApril", "SleepJan", "socialMedia", "state",
"ChildrenHome", "Gender", "Age")])

table(sleep3$Gender)

##
##            Female              Male         Nonbinary Prefer not to say
##               648               171                 2                 3
```

```r
sleep3$Gender <- as.factor(sleep3$Gender)
sleep3 <- sleep3[sleep3$Gender %in% c("Female", "Male"), ]
sleep3$Gender <- droplevels(sleep3$Gender)
levels(sleep3$Gender)

## [1] "Female" "Male"

t.test(SleepApril ~ Gender, data = sleep3)

##
##  Welch Two Sample t-test
##
## data:  SleepApril by Gender
## t = 0.14769, df = 308.31, p-value = 0.8827
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1975818  0.2296481
## sample estimates:
## mean in group Female   mean in group Male
##             6.987963             6.971930

Sleept_test <- t.test(SleepApril ~ Gender, data = sleep3)$conf.int
```

The t-test comparing recent sleep in male and female healthcare workers produces a value of well above 0.05 (approximately 0.70), suggesting that there is no significant difference in sleep time in April 2020 between male healthcare workers and female healthcare workers.

Bootstrap

```r
N <- 10000

diffSleep <- rep(NA, N)

for (i in 1:N) {
  sM <- sample(sleep3$SleepApril[sleep3$Gender == "Male"], sum(sleep3$Gender
== "Male"), replace = TRUE)
  sF <- sample(sleep3$SleepApril[sleep3$Gender == "Female"],
sum(sleep3$Gender == "Female"), replace = TRUE)
  diffSleep[i] <- mean(sM) - mean(sF)
}

ci <- quantile(diffSleep, c(0.025, 0.975))

hist(diffSleep, col = "green", main = "Bootstrapped sample Means Difference
in Sleep Time in April 2020", xlab = "Sleep Time (in hours)")
abline(v = ci, lwd = 3, col = "blue")
abline(v = Sleept_test, lwd = 3, col = "red", lty = 2)
legend("topright", c("Original Conf. Int.", "Boot Conf. Int."), lwd = 3, col
= c("red", "blue"), lty = c(2, 1))
```

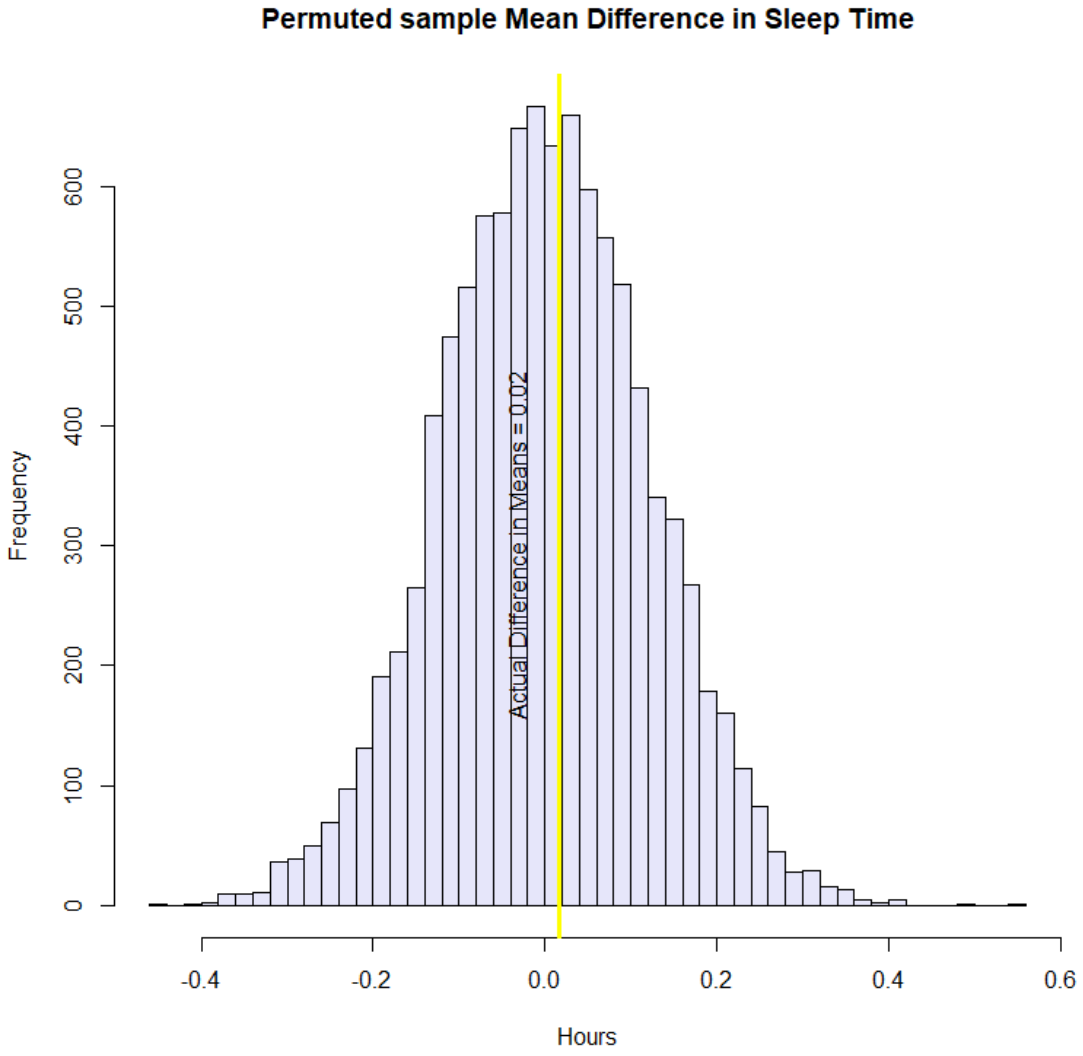


While both the lower and higher bounds of the theoretical (original) confidence interval are considerably lower than those of the bootstrapped confidence interval, both confidence intervals clearly include zero; this fortifies the idea that there is no statistically significant difference in a recent week's sleep time between men and women working in healthcare settings.

Permutation test

```
(meanSleep <- by(sleep3$SleepApril, sleep3$Gender, mean))

## sleep3$Gender: Female
## [1] 6.987963
## ------------------------------------------------------------
## sleep3$Gender: Male
## [1] 6.97193

(actualdifference <- meanSleep[1] - meanSleep[2])
```

```
##     Female
## 0.01603314

N <- 10000
permdiffSleep <- rep(NA, N)
for (i in 1:N) {
  fakeGender <- sample(sleep3$Gender)
  permdiffSleep[i] <- mean(sleep3$SleepApril[fakeGender == "Female"]) -
mean(sleep3$SleepApril[fakeGender == "Male"])
}


hist(permdiffSleep, col = "lavender", main = "Permuted sample Mean Difference
in Sleep Time", xlab = "Hours", breaks = 50)
abline(v = actualdifference, col = "yellow", lwd = 3)
text(actualdifference - 0.05, 300, paste("Actual Difference in Means =",
round(actualdifference, 2)), srt = 90)
```




```
mean(abs(permdiffSleep) >= abs(actualdifference))
```

```
## [1] 0.8966
```

From the histogram, it seems that the actual difference is close to the middle of the permuted sample distribution (which is centered around zero), suggesting that there is not a noticeable difference in sleep time between males and females. This is supported by the simple test above, which shows that the absolute value of the permuted difference in sleep values has approximately a 89% chance of being higher than the absolute value of the actual difference in sleep time between genders (about 0.05).

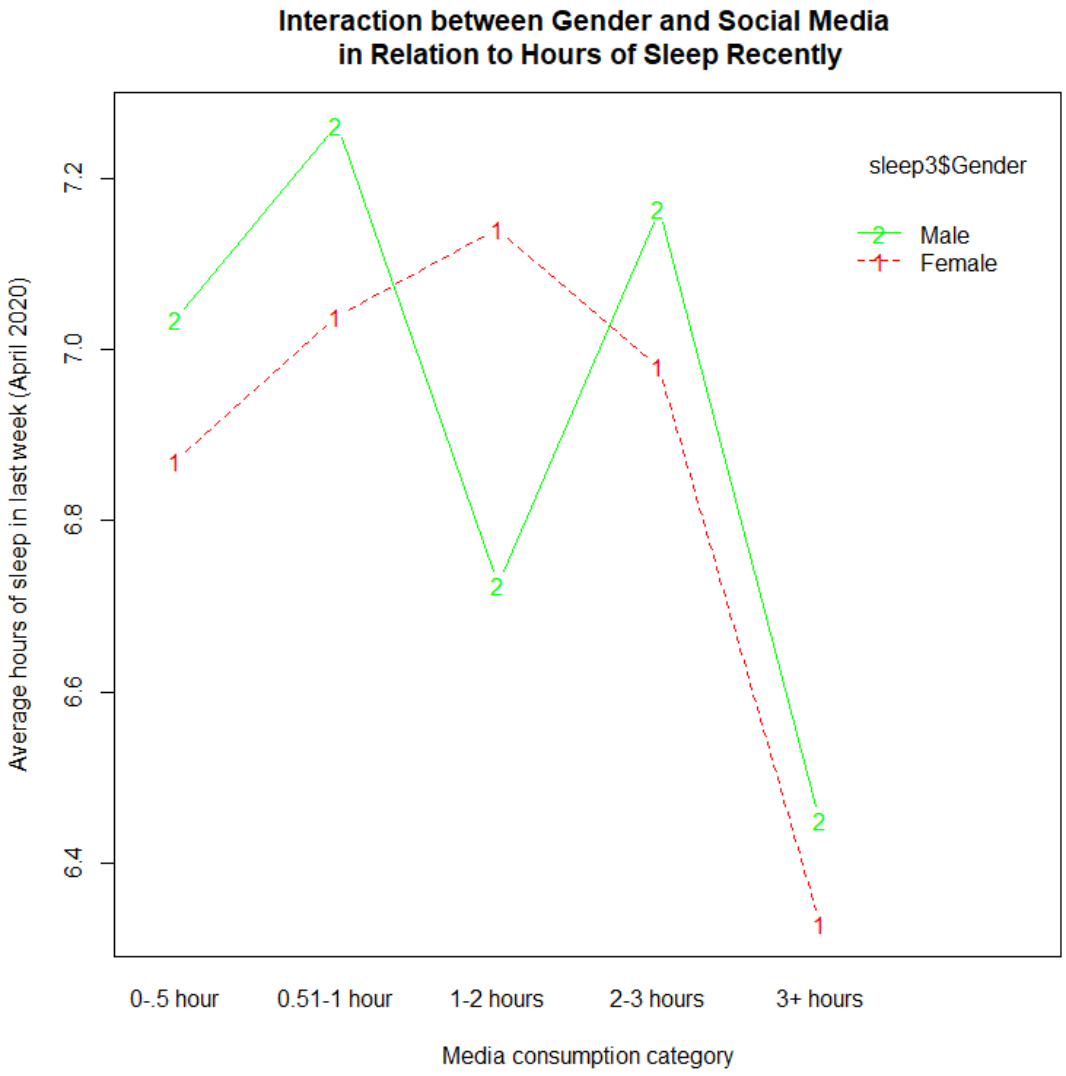
Case Study 2: Gender and COVID-19 Related Social Media Time Impact on Sleep in a Week in April 2020

ANOVA

For ANOVA, we are doing a two-way analysis on how gender and the amount of time spent on social media affects the amount of sleep people have gotten in a week in April 2020.

Interaction Plot

```
interaction.plot(sleep3$socialMedia, sleep3$Gender, sleep3$SleepApril, col = c("red","green"), type = 'b', main = "Interaction between Gender and Social Media \n in Relation to Hours of Sleep Recently", ylab = "Average hours of sleep in last week (April 2020)", xlab = "Media consumption category")
```



Based on this interaction plot, it appears unlikely that gender and hours spent on social media have an interaction because portions of the slope are parallel.

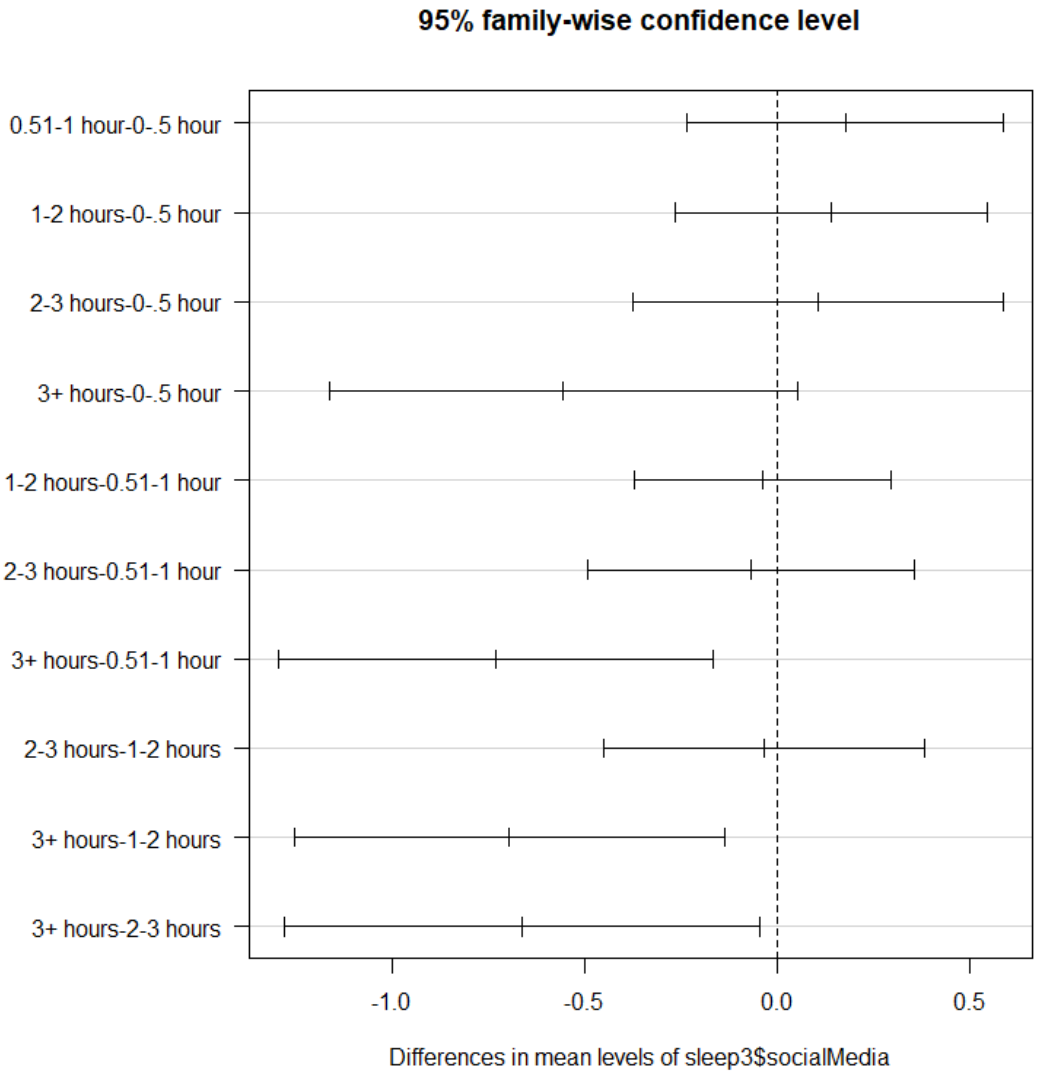
```
aov1 <- aov(sleep3$SleepApril ~ sleep3$socialMedia + sleep3$Gender + sleep3$Gender*sleep3$socialMedia)
Anova(aov1, type = 'III')
```

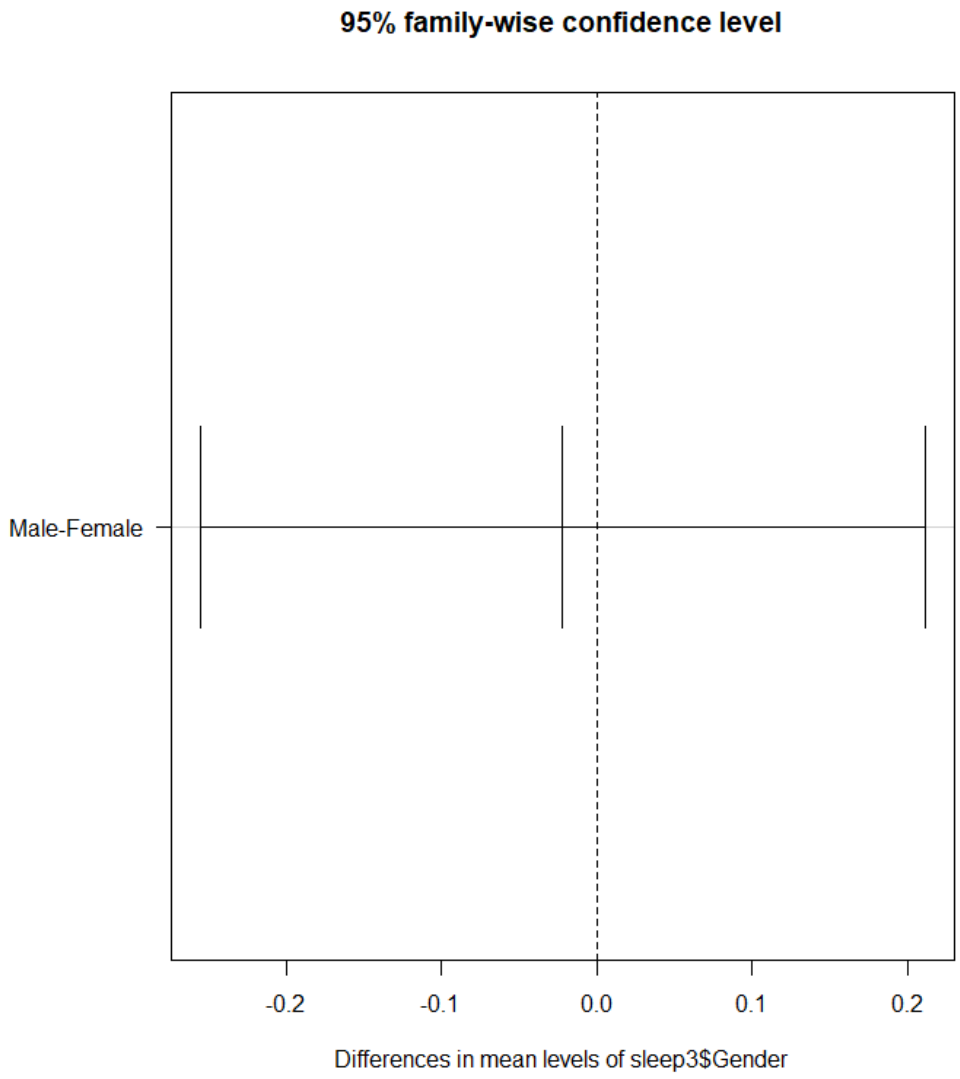
```
## Anova Table (Type III tests)
##
## Response: sleep3$SleepApril
##                                 Sum Sq  Df   F value    Pr(>F)
## (Intercept)                     4719.7   1 2465.4783 < 2.2e-16 ***
## sleep3$socialMedia                26.4   4    3.4446  0.008392 **
## sleep3$Gender                      0.6   1    0.3177  0.573141
## sleep3$socialMedia:sleep3$Gender  11.4   4    1.4864  0.204346
## Residuals                       1548.7 809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov2 <- aov(sleep3$SleepApril ~ sleep3$socialMedia + sleep3$Gender)
TukeyHSD(aov2)  # <-- Tukey's "Honest Significant Difference"
```

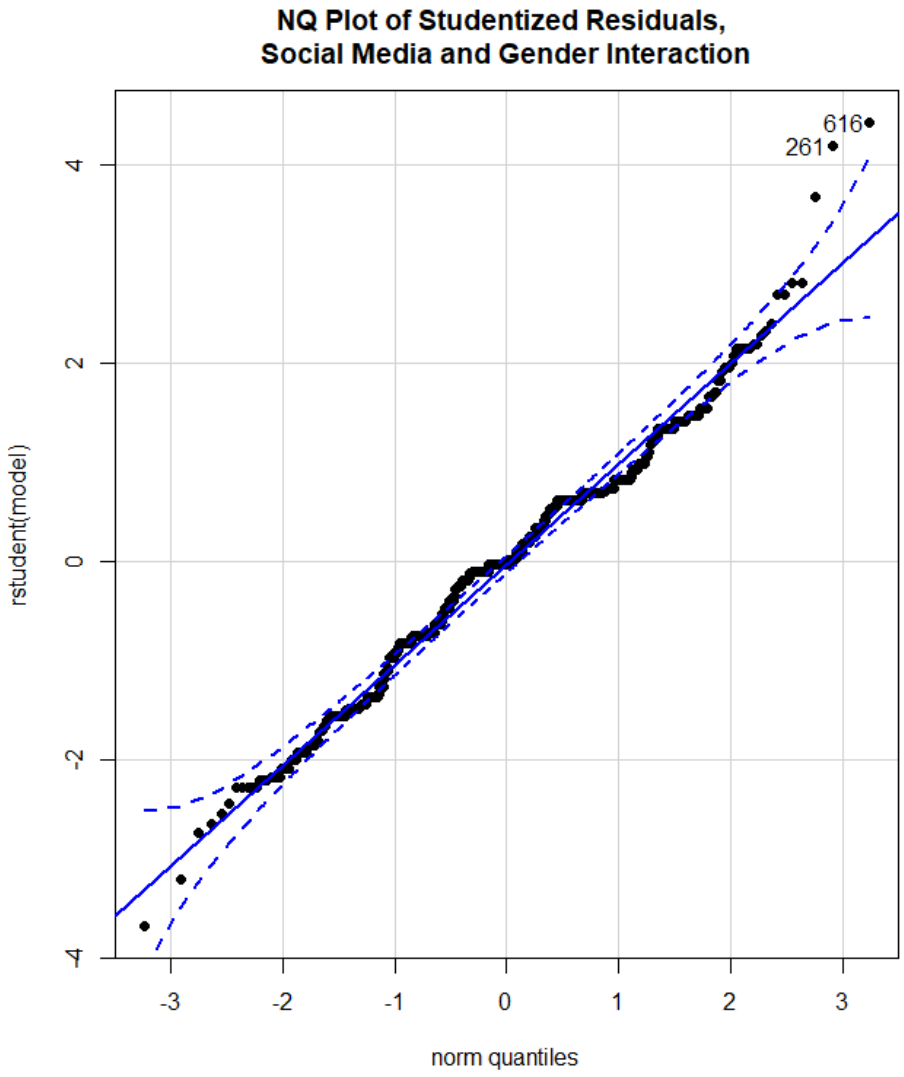
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sleep3$SleepApril ~ sleep3$socialMedia + sleep3$Gender)
##
## $`sleep3$socialMedia`
##                               diff        lwr         upr     p adj
## 0.51-1 hour-0-.5 hour  0.17676811 -0.2357877  0.58932394 0.7677610
## 1-2 hours-0-.5 hour    0.14083347 -0.2635484  0.54521531 0.8762133
## 2-3 hours-0-.5 hour    0.10700631 -0.3754099  0.58942249 0.9740896
## 3+ hours-0-.5 hour    -0.55606765 -1.1659442  0.05380894 0.0932448
## 1-2 hours-0.51-1 hour -0.03593464 -0.3696499  0.29778065 0.9983619
## 2-3 hours-0.51-1 hour -0.06976181 -0.4946931  0.35516949 0.9916080
## 3+ hours-0.51-1 hour  -0.73283577 -1.2983361 -0.16733538 0.0038221
## 2-3 hours-1-2 hours   -0.03382717 -0.4508271  0.38317278 0.9994631
## 3+ hours-1-2 hours    -0.69690113 -1.2564662 -0.13733609 0.0062307
## 3+ hours-2-3 hours    -0.66307396 -1.2813892 -0.04475871 0.0285022
##
## $`sleep3$Gender`
##                     diff        lwr       upr    p adj
## Male-Female -0.02172237 -0.2554859 0.2120411 0.855314
```

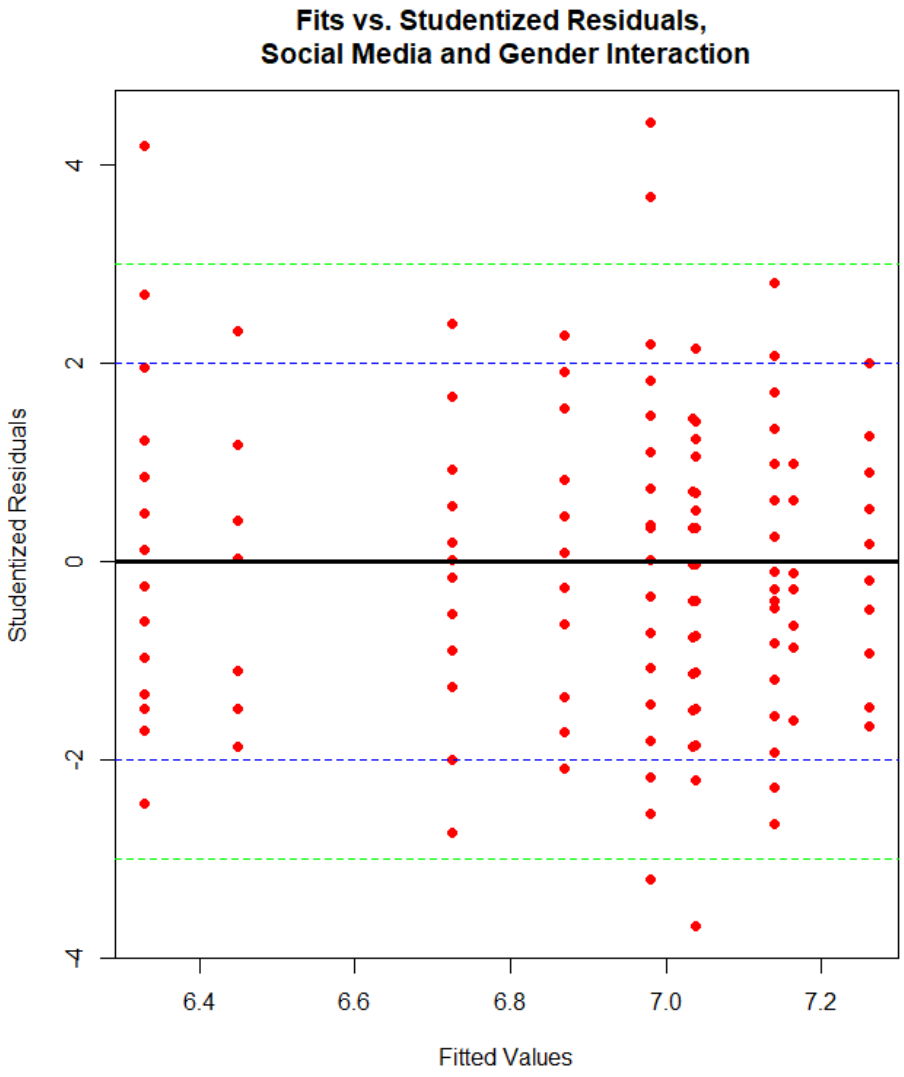
```
#Fix margins
par(mar=c(5, 11, 4, 1))
plot(TukeyHSD(aov2), las = 1)
```


95% family-wise confidence level
0.51-1 hour-0-.5 hour
1-2 hours-0-.5 hour
2-3 hours-0-.5 hour
3+ hours-0-.5 hour
1-2 hours-0.51-1 hour
2-3 hours-0.51-1 hour
3+ hours-0.51-1 hour
2-3 hours-1-2 hours
3+ hours-1-2 hours
3+ hours-2-3 hours
-1.0
-0.5
0.0
0.5
Differences in mean levels of sleep3$socialMedia




```
myResPlots2(aov1, label = "\nSocial Media and Gender Interaction")
```


NQ Plot of Studentized Residuals,
Social Media and Gender Interaction
616
261
rstudent(model)
norm quantiles
4
2
0
-2
-4
-3
-2
-1
0
1
2
3



The conclusion reached from the Anova test matches what was discovered in the interaction plot. Social Media and Gender do not have a significant interaction. Because the Tukey comparison is limited to one-way Anova tests, and our Anova test was two-way, the Tukey analysis split out two categorical variables into two Tukey plots. It appears that the only significant pairs that had significantly different average sleep times were 3+ hours-0.51-1 hour pair, the 3+ hours-1-2 hours pair, and the 3+ hours-2-3 hours when the one-way Anova test was performed on these two categorical variables. Male and Female were not a significant pair. From the residual plots, the quantile plot looks fairly linear, meaning the distribution is fairly normal for the residuals, and there does not seem to be evidence of heteroskedasticity.

Case Study 3: Investigating if Healthcare Workers Had Less Sleep in April 2020 vs. January 2020

To investigate if health care workers received less sleep in April 2020 vs. January 2020, we performed a paired t-test because the the two groups we are comparing come from the

same individuals, and are thus paired based on the individual. We have already confirmed the data we are comparing is relatively normally distributed from the chart.Correlation result.

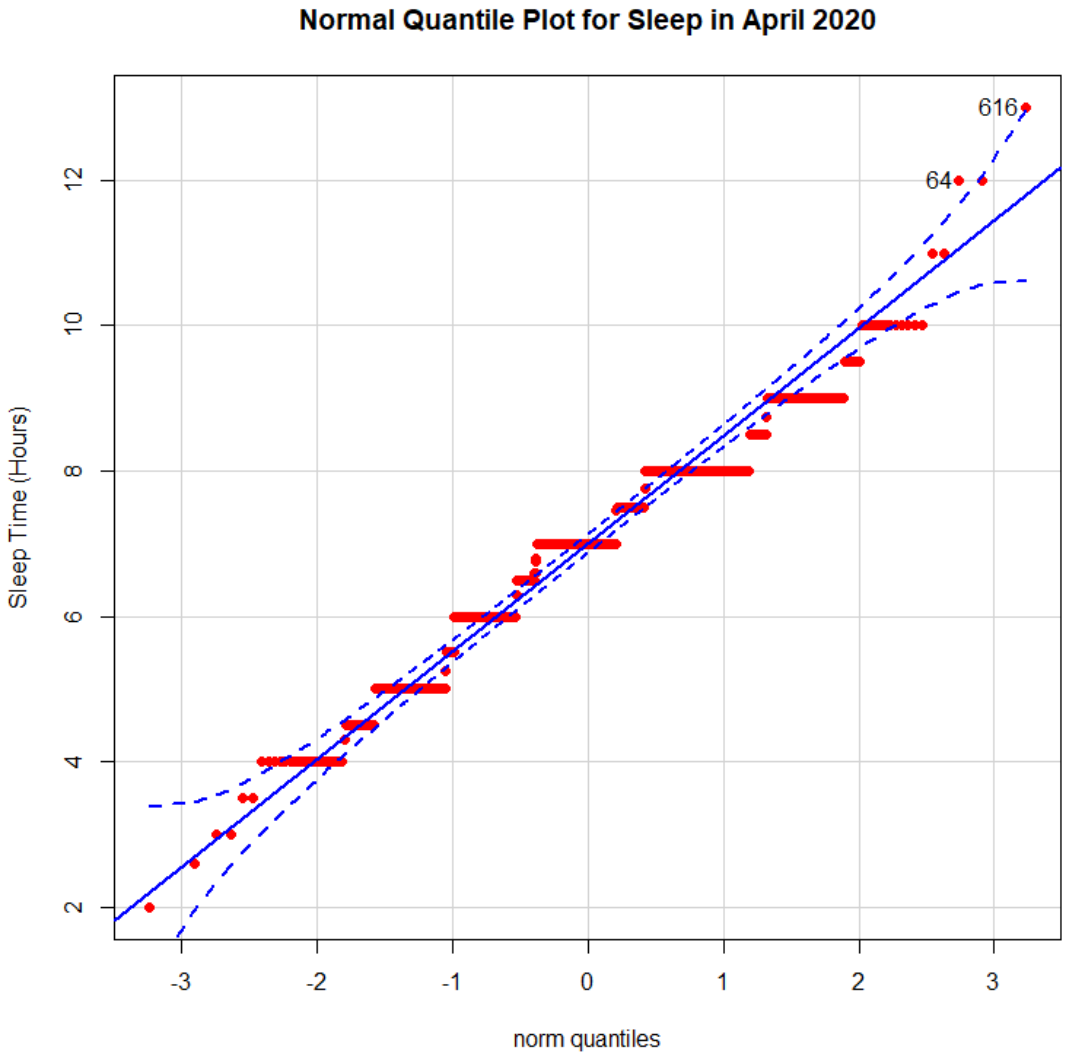
```
sleep4 <- sleep3[complete.cases(sleep3),]
t.test(sleep4$SleepApril, sleep4$SleepJan, paired = TRUE)

##
##  Paired t-test
##
## data:  sleep4$SleepApril and sleep4$SleepJan
## t = -2.873, df = 818, p-value = 0.004171
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.24271856 -0.04568193
## sample estimates:
## mean of the differences
##              -0.1442002
```

Based off of the paired t-test, the difference in healthcare workers sleep between January 2020 and April 2020 was significant (p-value was 0.004 which is less than 0.05). The t-test also said the mean of the differences was -0.144, concluding that healthcare workers received less sleep in April 2020.

Final Model: Using Multiple Regression to Build the Final Model that Includes All Variables to predict recent sleep of Health Care workers

As stated in the introduction, one of the most useful applications of this dataset is to create a linear model that predicts hours of sleep since it is a known predictor of quality of life and physical health. In order to do this, we plan to use backwards stepwise regression to create the final model. To do so, we use all variables we cleaned, then use Anova to find which variables and interactions are non-significant to create the final model.

We want to make sure our response variable is normally distributed. We can look at a quantile plot to test this.




```
## [1] 616  64
```

Because the quantile plot for sleep time is pretty linear, we can say that the distribution of April 2020 sleep time for healthcare workers is approximately normal.

```
# State is removed initially as it had perfect multicollinearity for some
predictors, likely due to unbalanced sample regarding state
m1 <- lm(SleepApril ~ Gender + Age + SleepJan + ChildrenHome + socialMedia +
Gender*socialMedia, data = sleep3)
Anova(m1, type = 3)

## Anova Table (Type III tests)
##
## Response: SleepApril
##                      Sum Sq  Df  F value    Pr(>F)
```

```
## (Intercept)         199.09   1 114.0820 < 2.2e-16 ***
## Gender                1.31   1   0.7531 0.3857443
## Age                   5.56   1   3.1877 0.0745711 .
## SleepJan             91.92   1  52.6710 9.283e-13 ***
## ChildrenHome         23.15   1  13.2637 0.0002879 ***
## socialMedia          18.12   4   2.5964 0.0351922 *
## Gender:socialMedia   10.07   4   1.4421 0.2181951
## Residuals          1406.62 806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Backwards stepwise regression into m2: Removing Gender*socialMedia, gender, Age
m2 <- lm(SleepApril ~ SleepJan + ChildrenHome + socialMedia, data = sleep3)
Anova(m2, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: SleepApril
##               Sum Sq  Df  F value    Pr(>F)
## (Intercept)   222.59   1 127.0255 < 2.2e-16 ***
## SleepJan      103.04   1  58.8032 4.982e-14 ***
## ChildrenHome   19.88   1  11.3472 0.0007912 ***
## socialMedia    15.60   4   2.2253 0.0645983 .
## Residuals    1422.89 812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

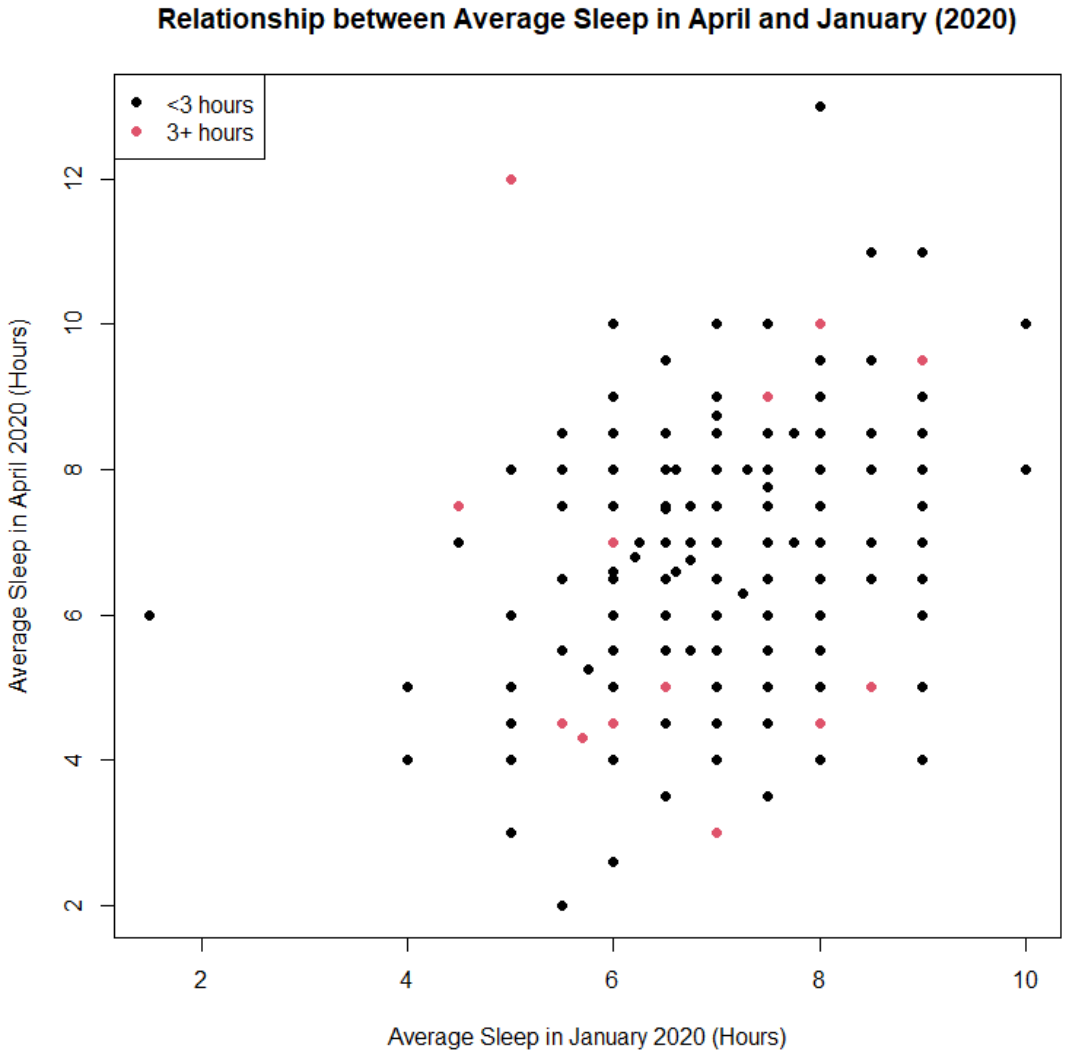
```
summary(m2)
```

```
##
## Call:
## lm(formula = SleepApril ~ SleepJan + ChildrenHome + socialMedia,
##     data = sleep3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2801 -0.8478  0.1356  0.8649  6.2014
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.28667    0.38034  11.271  < 2e-16 ***
## SleepJan               0.38955    0.05080   7.668 4.98e-14 ***
## ChildrenHome          -0.13537    0.04019  -3.369 0.000791 ***
## socialMedia0.51-1 hour  0.12160    0.14452   0.841 0.400381
## socialMedia1-2 hours    0.09893    0.14160   0.699 0.484979
## socialMedia2-3 hours   -0.02981    0.16935  -0.176 0.860316
## socialMedia3+ hours    -0.43581    0.21363  -2.040 0.041670 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.324 on 812 degrees of freedom
## Multiple R-squared:  0.1032, Adjusted R-squared:  0.09654
## F-statistic: 15.57 on 6 and 812 DF,  p-value: < 2.2e-16
```

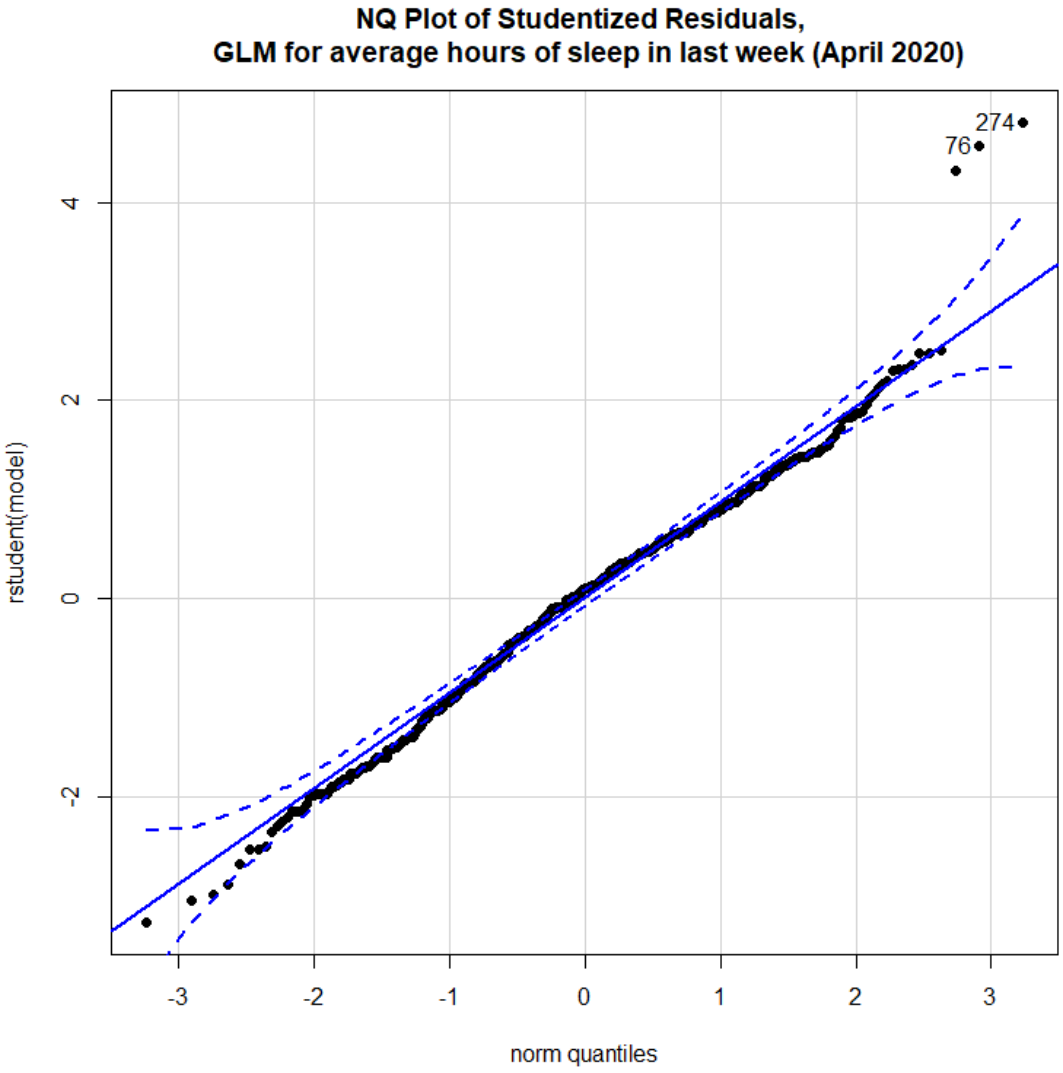
The final model predicting the recent sleep time scale included Sleep time in January 2020, Children at home, and Social media (we felt this p-value is worth considering as significant because 0.06 is very close to 0.05). We chose to do backwards stepwise regression to accommodate both categorical and continuous variables, and the final model was created by only keeping the significant predictors. Overall sleep time was lowest in those who were in the highest category of social media use (see coefficients). Overall, more children at home was associated with lower sleep times. Healthcare workers who had more sleep in January 2020 also generally had more sleep in April. The overall R-squared was relatively low (only about 10%).

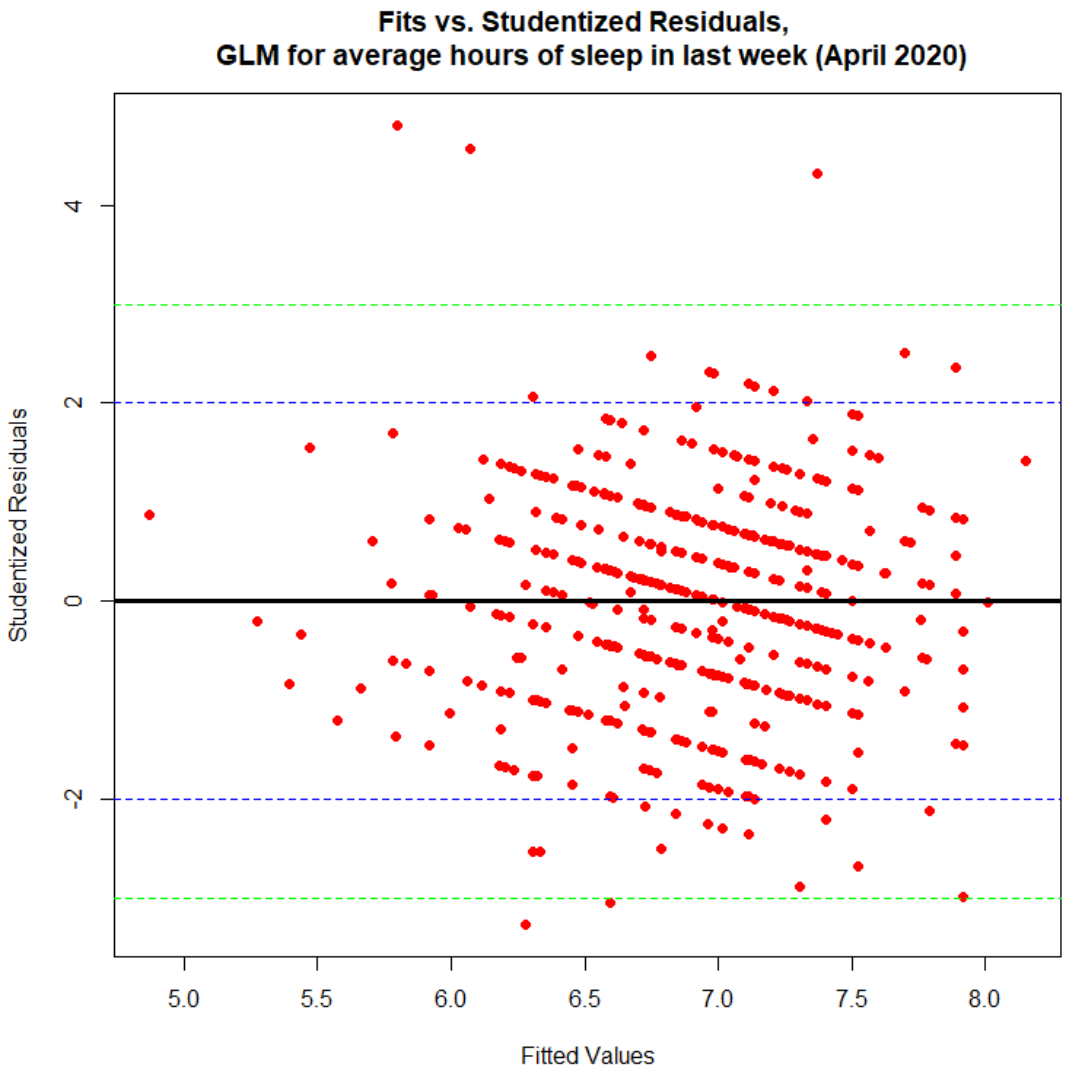
```
sleep3$mediatemp <- sleep3$socialMedia
sleep3$mediatemp[sleep3$mediatemp != "3+ hours"] <- "<3 hours"
plot(SleepApril ~ SleepJan, data = sleep3, main = "Relationship between
Average Sleep in April and January (2020)", xlab = "Average Sleep in January
2020 (Hours)", ylab = "Average Sleep in April 2020 (Hours)", pch = 16, col =
factor(sleep3$mediatemp))
legend("topleft", col = 1:5, legend = levels(factor(sleep3$mediatemp)), pch =
16)
```



As can be seen with the coefficients found in the final linear model, the relationship between hours of sleep in April vs. January 2020 is more negative for people who consume COVID related social media for more than 3 hours.

The fit of the model was good can be visualized in residual plots:





The first plot shows residuals were approximately normally distributed with no evidence of non-linear trends and the second plot shows there was no evidence of heteroskedasticity, extreme outliers or influential points. These are signs of a well fit model as the linearity, homoscedasticity, independence and normality assumptions of linear regressions are met.

Part 4: Summary of Findings and Conclusion

Throughout this project we have found that the best predictors for the average amount of sleep healthcare workers received in a week in April 2020 were the number of children they had at home, their social media usage, and the amount of sleep they received in January. Due to the data being (unfortunately) quite concentrated in Michigan, the state variable could not be used in the final model because several states only had one participant in the survey, resulting in issues with perfect multicollinearity. We began by testing if sleep time in an April 2020 week differed by gender using t-test, bootstrapped mean differences, and a permutation test. As discussed above, we determined that sleep

time did not differ significantly by gender. A t-test was also conducted to evaluate whether each individual's sleep time differed significantly between January 2020 and April 2020. The results of the t-test, paired with the difference in means, showed us that healthcare workers unfortunately received statistically significantly less sleep in January 2020 compared to April 2020. This illustrates the detrimental effect of the COVID-19 pandemic on the lifestyle of healthcare workers. We proceeded to carry out a two-way ANOVA test that showed us that the interaction between gender and social media time was not significant. Next, backwards stepwise multiple regression was carried out to remove all non-significant predictors and build a model based exclusively on significant predictors. As discussed above, we found the significant predictors of April 2020 sleep to be January 2020 sleep (positive), the amount of children at home (negative), and social media usage (negative). Ultimately, this illustrates that good habits such as adequate sleep can persist even in the face of extenuating circumstances such as a global pandemic. It also suggests that a healthcare worker needs to think carefully about their lifestyle choices– namely, how much time they are spending consuming health related news and how many children they decide to have– to ensure that they are not jeopardizing their ability to get a good night's sleep. Based on current knowledge that healthcare workers perform better when more well rested, understanding what factors affect healthcare workers' sleep may not only help improve their quality of life but also may improve patient outcomes.