Homework 09

Due: 2024-11-17

1.

We see a strong relationship between income and amount spent on gambling among males in the study, though not so much for the females. Looking at our model, we see that for each unit change of income we expect to see an increase of 6.5181 in males, while we would only expect to see an increase of 0.1749 in gambling amount for females. This is illustrated by the regression line on our plot.

Looking at similar graphs that measure against amount spent gambling yields similar results. In all comparisons, females have very little amount spent gambling. An interesting note is that males with higher verbal scores and families from a higher socioeconomic status will gamble less than males with lower scores in both variables.

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2
```

```
## Warning: package 'tidyr' was built under R version 4.4.2
```

```
## Warning: package 'readr' was built under R version 4.4.2
```

```
## Warning: package 'purrr' was built under R version 4.4.2
```

```
## Warning: package 'forcats' was built under R version 4.4.2
```

```
## Warning: package 'lubridate' was built under R version 4.4.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
data(teengamb, package = "faraway")

teengamb$sex <- as.factor(teengamb$sex)

by(teengamb, teengamb$sex, summary)
```

```
## teengamb$sex: 0
##  sex        status          income           verbal           gamble
##  0:28   Min.   :18.00   Min.   : 0.600   Min.   : 1.000   Min.   :  0.000
##  1: 0   1st Qu.:38.00   1st Qu.: 2.000   1st Qu.: 6.000   1st Qu.:  2.775
##         Median :51.00   Median : 3.375   Median : 7.000   Median : 14.250
##         Mean   :52.00   Mean   : 4.976   Mean   : 6.821   Mean   : 29.775
##         3rd Qu.:65.25   3rd Qu.: 6.625   3rd Qu.: 8.250   3rd Qu.: 42.175
##         Max.   :75.00   Max.   :15.000   Max.   :10.000   Max.   :156.000
## ------------------------------------------------------------
## teengamb$sex: 1
##  sex        status          income           verbal          gamble
##  0: 0   Min.   :18.00   Min.   : 1.500   Min.   :4.000   Min.   : 0.000
##  1:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.:6.000   1st Qu.: 0.100
##         Median :30.00   Median : 3.000   Median :6.000   Median : 1.700
##         Mean   :35.26   Mean   : 4.149   Mean   :6.421   Mean   : 3.866
##         3rd Qu.:43.00   3rd Qu.: 5.750   3rd Qu.:8.000   3rd Qu.: 6.000
##         Max.   :65.00   Max.   :10.000   Max.   :8.000   Max.   :19.600
```

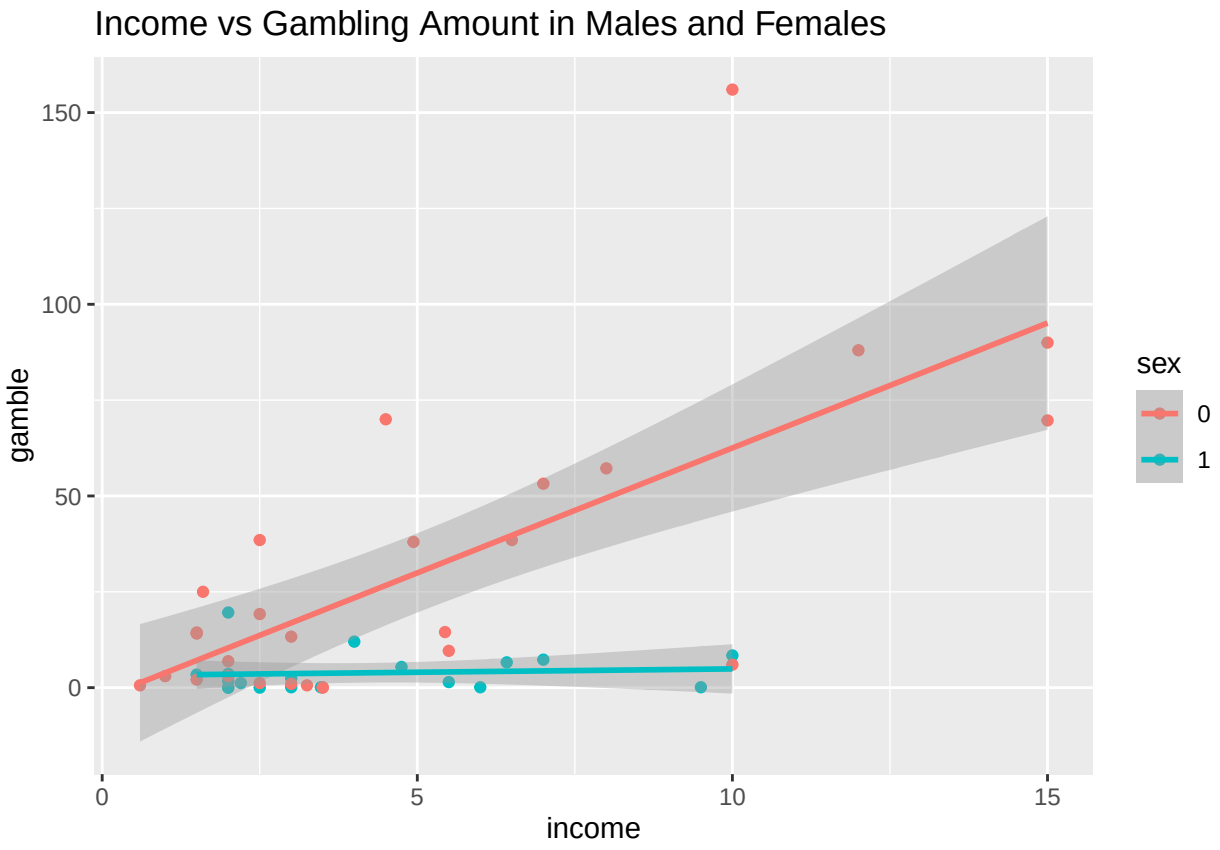
```r
lmod <- lm(gamble ~ income + sex + sex*income, data = teengamb)

summary(lmod)
```

```
##
## Call:
## lm(formula = gamble ~ income + sex + sex * income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.522  -4.860  -1.790   6.273  93.478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.6596     6.3164  -0.421  0.67580
## income        6.5181     0.9881   6.597 4.95e-08 ***
## sex1          5.7996    11.2003   0.518  0.60724
## income:sex1  -6.3432     2.1446  -2.958  0.00502 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.98 on 43 degrees of freedom
## Multiple R-squared:  0.5857, Adjusted R-squared:  0.5568
## F-statistic: 20.26 on 3 and 43 DF,  p-value: 2.451e-08
```

```r
teengamb %>%
  ggplot(aes(x = income, y = gamble, color = sex)) +
  geom_point() +
```

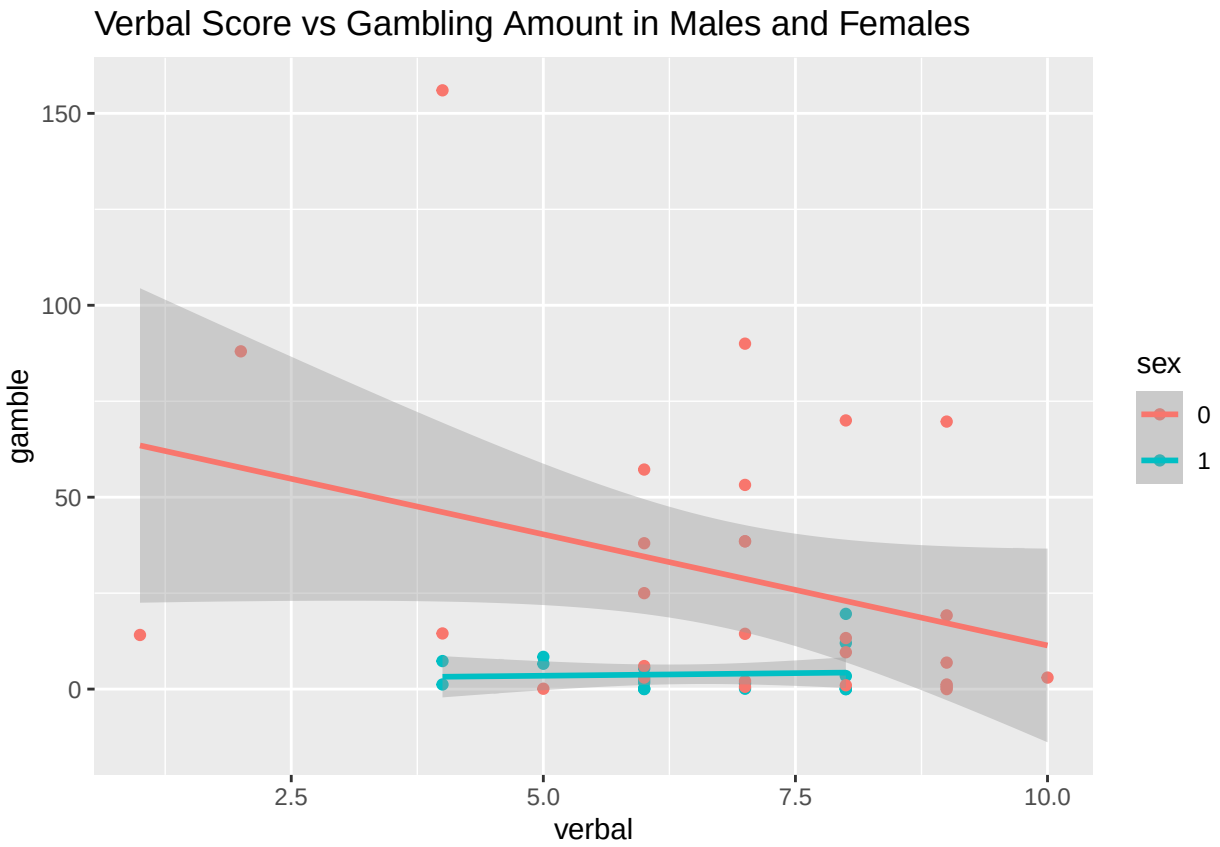
```
  geom_smooth(method = "lm") +
  labs(title = "Income vs Gambling Amount in Males and Females")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



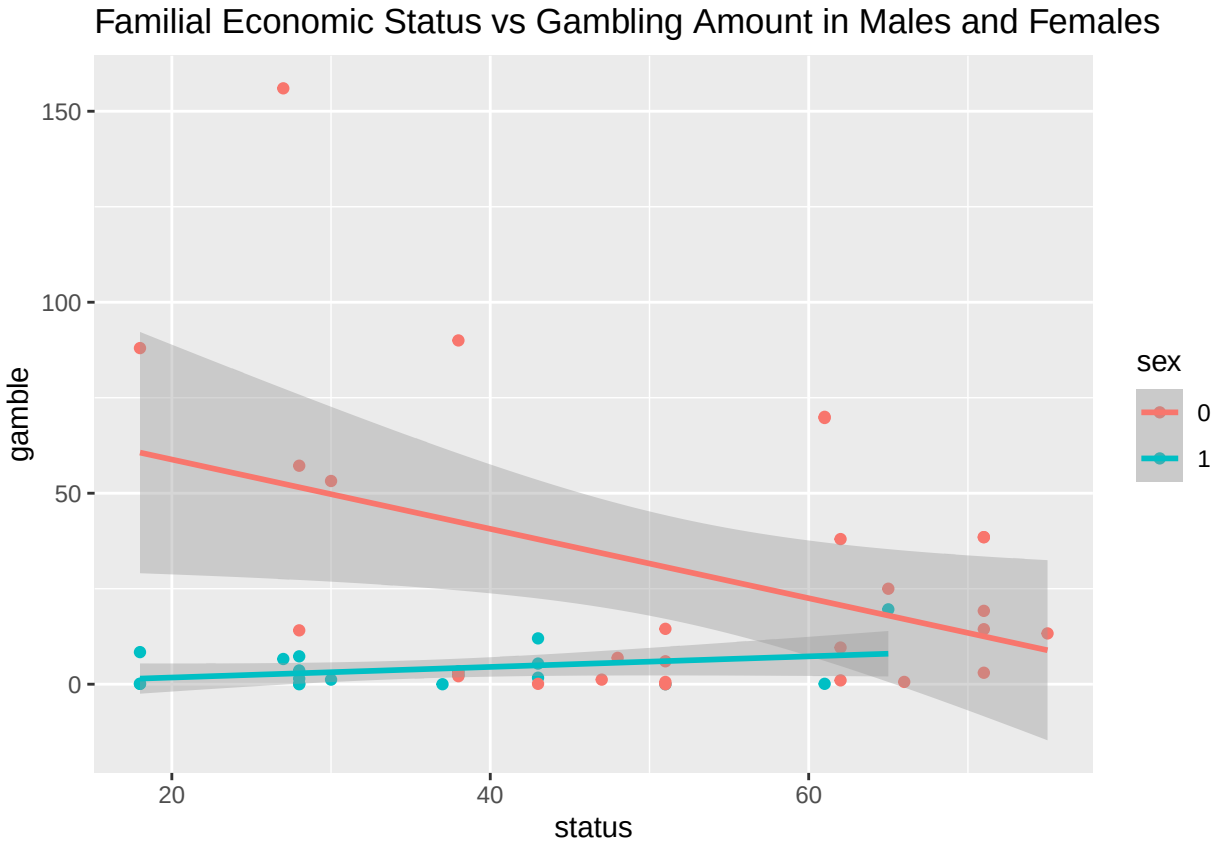

```
teengamb %>%
  ggplot(aes(x = verbal, y = gamble, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Verbal Score vs Gambling Amount in Males and Females")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
teengamb %>%
  ggplot(aes(x = status, y = gamble, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Familial Economic Status vs Gambling Amount in Males and Females")
```

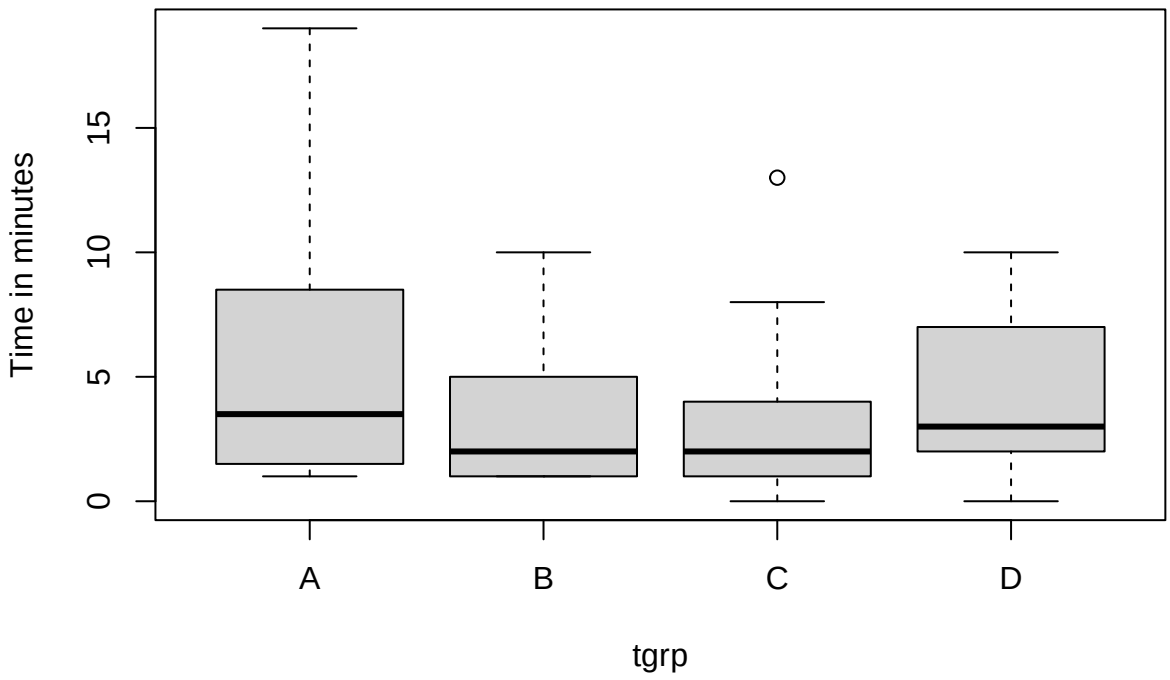
```
## `geom_smooth()` using formula = 'y ~ x'
```



2.

a. The most notable feature is that group B has a lower quartile nearly the same as its minimum. Also, the "A" tgrp group has a much larger range in values, going nearly as high as 20. Group B and D have a median close to the lower quartile.

```
data(anaesthetic, package="faraway")
plot(breath ~ tgrp, data=anaesthetic,ylab="Time in minutes")
```



b. We do not see a very large difference in the levels of the treatments.

```
require(faraway)
```

```
## Loading required package: faraway
```

```
##
## Attaching package: 'faraway'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     teengamb
```

```
lmod2 <- lm(breath ~ tgrp, data=anaesthetic)
summary(lmod2)
```

```
##
## Call:
## lm(formula = breath ~ tgrp, data = anaesthetic)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -2.250 -1.250  1.613 13.600
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.400      0.816   6.618 4.57e-09 ***
## tgrpB         -2.150      1.154  -1.863   0.0663 .
## tgrpC         -2.350      1.154  -2.036   0.0452 *
## tgrpD         -1.050      1.154  -0.910   0.3658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.649 on 76 degrees of freedom
## Multiple R-squared:  0.0655, Adjusted R-squared:  0.02861
## F-statistic: 1.776 on 3 and 76 DF,  p-value: 0.1589
```

```
round(coef(lmod2),1)
```

```
## (Intercept)       tgrpB       tgrpC       tgrpD
##         5.4        -2.1        -2.3        -1.1
```

```
model.matrix(lmod2)
```

```
##    (Intercept) tgrpB tgrpC tgrpD
## 1            1     0     0     0
## 2            1     1     0     0
## 3            1     0     1     0
## 4            1     0     0     1
## 5            1     0     0     0
## 6            1     1     0     0
## 7            1     0     1     0
## 8            1     0     0     1
## 9            1     0     0     0
## 10           1     1     0     0
## 11           1     0     1     0
## 12           1     0     0     1
## 13           1     0     0     0
## 14           1     1     0     0
## 15           1     0     1     0
## 16           1     0     0     1
## 17           1     0     0     0
## 18           1     1     0     0
## 19           1     0     1     0
## 20           1     0     0     1
## 21           1     0     0     0
## 22           1     1     0     0
## 23           1     0     1     0
## 24           1     0     0     1
## 25           1     0     0     0
## 26           1     1     0     0
## 27           1     0     1     0
## 28           1     0     0     1
## 29           1     0     0     0
## 30           1     1     0     0
## 31           1     0     1     0
```

```
## 32           1     0     0     1
## 33           1     0     0     0
## 34           1     1     0     0
## 35           1     0     1     0
## 36           1     0     0     1
## 37           1     0     0     0
## 38           1     1     0     0
## 39           1     0     1     0
## 40           1     0     0     1
## 41           1     0     0     0
## 42           1     1     0     0
## 43           1     0     1     0
## 44           1     0     0     1
## 45           1     0     0     0
## 46           1     1     0     0
## 47           1     0     1     0
## 48           1     0     0     1
## 49           1     0     0     0
## 50           1     1     0     0
## 51           1     0     1     0
## 52           1     0     0     1
## 53           1     0     0     0
## 54           1     1     0     0
## 55           1     0     1     0
## 56           1     0     0     1
## 57           1     0     0     0
## 58           1     1     0     0
## 59           1     0     1     0
## 60           1     0     0     1
## 61           1     0     0     0
## 62           1     1     0     0
## 63           1     0     1     0
## 64           1     0     0     1
## 65           1     0     0     0
## 66           1     1     0     0
## 67           1     0     1     0
## 68           1     0     0     1
## 69           1     0     0     0
## 70           1     1     0     0
## 71           1     0     1     0
## 72           1     0     0     1
## 73           1     0     0     0
## 74           1     1     0     0
## 75           1     0     1     0
## 76           1     0     0     1
## 77           1     0     0     0
## 78           1     1     0     0
## 79           1     0     1     0
## 80           1     0     0     1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$tgrp
## [1] "contr.treatment"
```

```r
anova(lmod2)
```

```
## Analysis of Variance Table
##
## Response: breath
##           Df  Sum Sq Mean Sq F value Pr(>F)
## tgrp       3   70.94  23.646  1.7757 0.1589
## Residuals 76 1012.05  13.316
```

```r
## Fitting the model without intercept term
## lmodi <- lm(breath ~ tgrp -1, anaesthetic)

## summary(lmodi)
```

c. A box-cox transformation is not possible because the response variable must be positive for a box-cox transformation. I have allowed the error to push through to illustrate this.

```r
require(MASS)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
boxcox(lmod2, plotit=T)
```

```
## Error in boxcox.default(lmod2, plotit = T): response variable must be positive
```

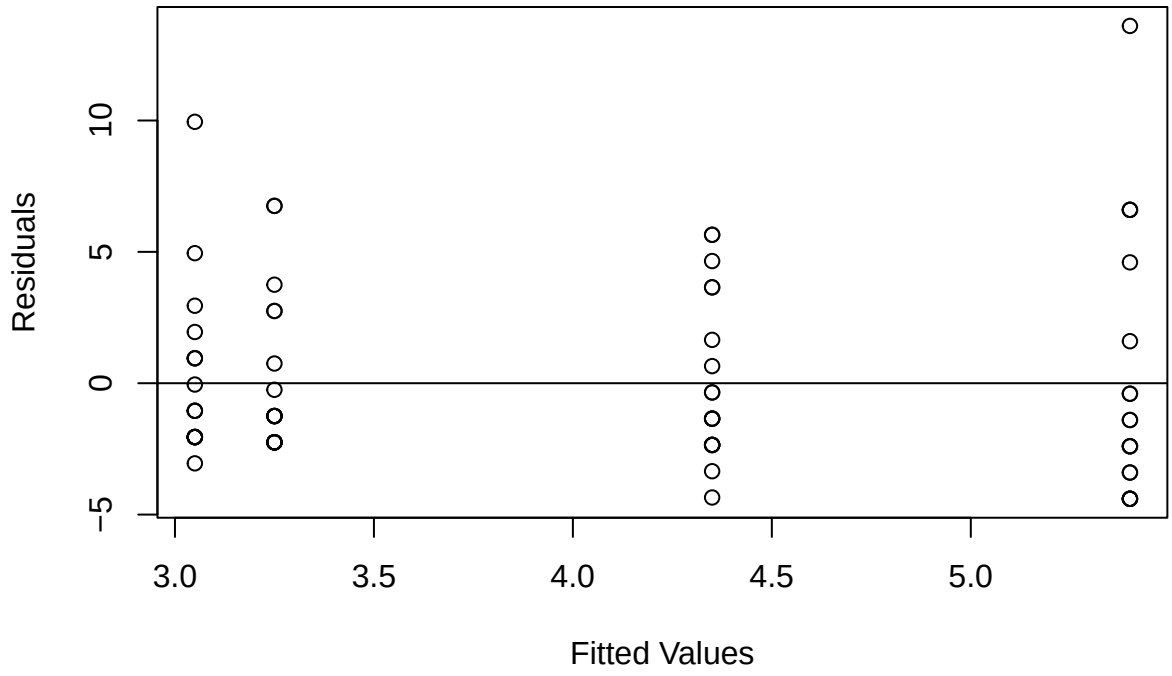
d. The diagnostics appear satisfactory, and there does not appear to be a significant difference among the treatment groups after performing a square root transformation.

```r
lmod3 <- lm(sqrt(abs(residuals(lmod2))) ~ fitted(lmod2))

summary(lmod3)
```

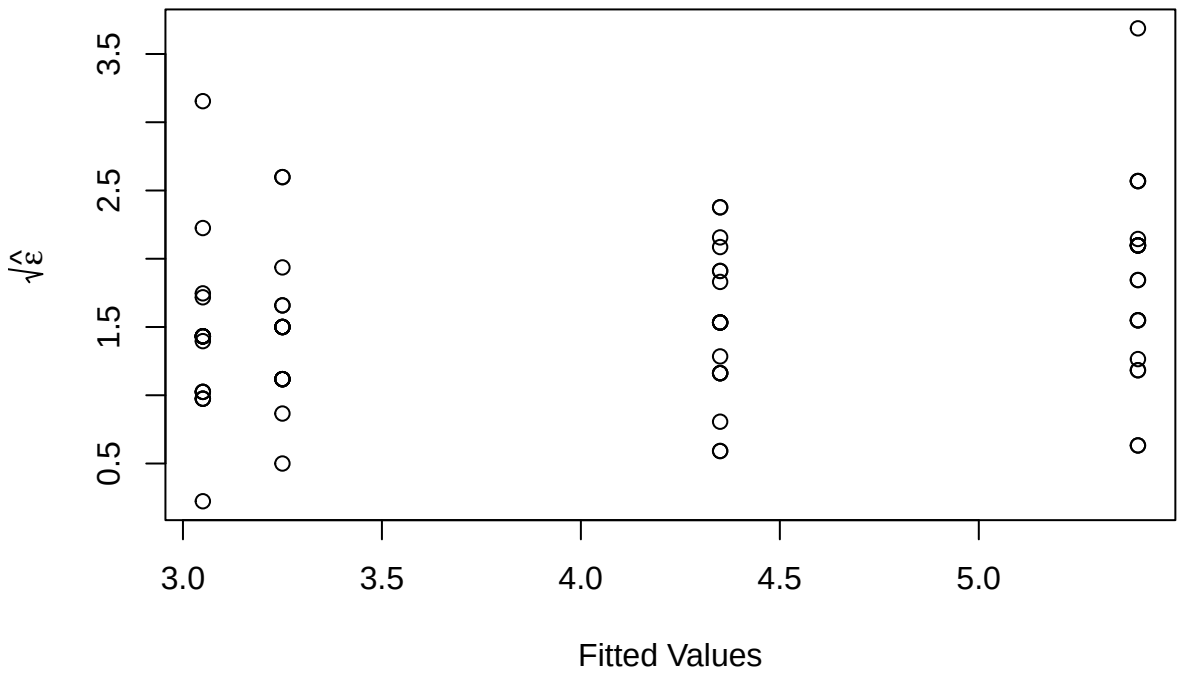
```
##
## Call:
## lm(formula = sqrt(abs(residuals(lmod2))) ~ fitted(lmod2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17531 -0.35565  0.05144  0.28985  1.88005
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    0.82561    0.28636   2.883  0.00509 **
## fitted(lmod2)  0.18188    0.06948   2.618  0.01063 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5852 on 78 degrees of freedom
## Multiple R-squared:  0.08076,    Adjusted R-squared:  0.06897
## F-statistic: 6.853 on 1 and 78 DF,  p-value: 0.01063
```

```r
plot(fitted(lmod2), residuals(lmod2), xlab="Fitted Values", ylab="Residuals")
abline(h=0)
```




```r
plot(fitted(lmod2), sqrt(abs(residuals(lmod2))), xlab="Fitted Values", ylab=expression(sqrt(hat(epsilon)
abline(h=0)
```


$\sqrt{|\hat{\varepsilon}|}$
0.5
1.5
2.5
3.5
3.0
3.5
4.0
4.5
5.0
Fitted Values