ANOVA Functions and Residual Analysis

Ethan Tsao

2024-10-23

Part 1: ANOVA function

(a)

```
aov_function <- function(df) {

  outcome_var <- df[[1]]
  factor_var <- df[[2]]

  groups <- split(outcome_var, factor_var)
  k <- length(groups)
  n <- length(outcome_var)
  y_ddot <- mean(outcome_var)
  df_bw_groups <- k-1
  df_in_groups <- n-k

  SSE <- sum(sapply(groups, function(group) (length(group) - 1) * var(group)))

  MSE <- SSE/df_in_groups

  SST <- sum(sapply(groups, function(group) length(group) * (mean(group) - y_ddot)^2))

  MST <- SST/df_bw_groups

  F0 <- MST/MSE

  aov_table <- matrix(c(df_bw_groups, SST, MST, F0,
                        df_in_groups, SSE, MSE, NA),
                      nrow = 2, byrow = TRUE)

  rownames(aov_table) <- c("method", "Residuals")
  colnames(aov_table) <- c("Df", "Sum Sq", "Mean Sq", "F value")

  return(aov_table)
}
```

(b)

```
google <- c(46, 49, 51, 42)
waze <- c(44, 47, 47, 43)
gut <- c(50, 51, 45, 43)

commute_df <- data.frame(
  time = c(google, waze, gut),
  method = c(rep("google", 4),
             rep("waze", 4),
             rep("gut", 4))
)

aov_function(commute_df)
```

```
##           Df Sum Sq Mean Sq   F value
## method     2    9.5    4.75 0.4130435
## Residuals  9  103.5   11.50        NA
```

```
model1 <- aov(time ~ method, data=commute_df)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## method       2    9.5    4.75   0.413  0.674
## Residuals    9  103.5   11.50
```

(c)

```
temp_100 <- c(21.8, 21.9, 21.7, 21.6, 21.7)
temp_125 <- c(21.7, 21.4, 21.5, 21.4)
temp_150 <- c(21.9, 21.8, 21.8, 21.6, 21.5)
temp_175 <- c(21.9, 21.7, 21.8, 21.4)

temp_df <- data.frame(
  density = c(temp_100, temp_125, temp_150, temp_175),
  method = c(rep("temp_100", 5),
             rep("temp_125", 4),
             rep("temp_150", 5),
             rep("temp_175", 4))
)

aov_function(temp_df)
```

```
##           Df    Sum Sq    Mean Sq  F value
## method     3 0.1561111 0.05203704 2.023663
## Residuals 14 0.3600000 0.02571429       NA
```

```r
model2 <- aov(density ~ method, data=temp_df)
summary(model2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## method       3 0.1561 0.05204   2.024  0.157
## Residuals   14 0.3600 0.02571
```

Part 2: Other Exercises

Exercise 3.8

1. Not a Designed Experiment:

- During a designed experiment, researchers manipulate one or more factors and control other variables to gauge their effects on an outcome. In the case of this experiment, the researchers did not manipulate chocolate consumption. They only observed and recorded the participants' habits and depression levels, which makes it an observational study rather than a controlled experiment.

2. Not a Cause and Effect Link Established:

- Since we concluded that this is an observational study, it cannot establish a cause and effect relationship between chocolate consumption and depression.
- The study does find a correlation, but this does not imply that there is causation. There could be confonding vairables that explain this relationship, such as stress, pre-existing eating habits, and other factors. Without controlling for these variables we cannot say that eating more chocolate causes depression or vice versa.

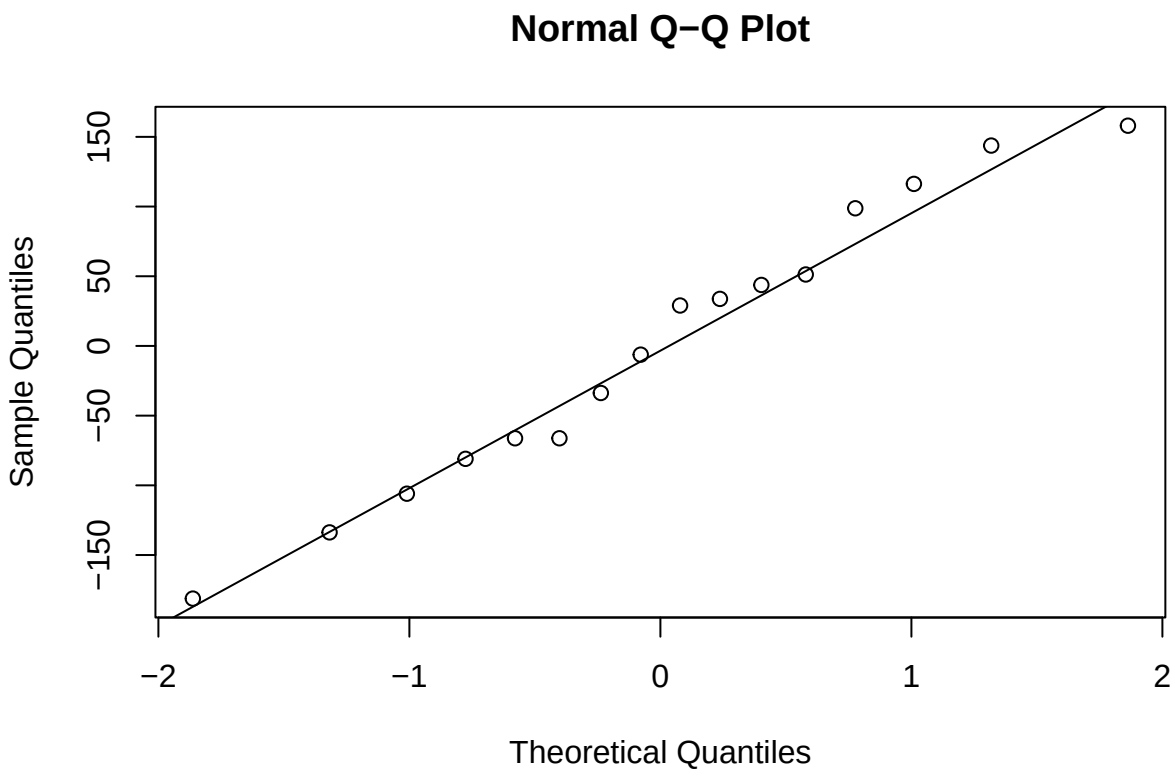
3. How to Establish a Cause and Effect Link:

- In order to establish a cause and effect relationship, the study would need to be a randomized controlled trial.
- This would require participants to be randomly assigned to two or more groups. One group for example, could be assigned to consume a speciffic amount of chocolate (treatment group), and the other would consume no chocolate (control group).
- Control of Variables: All other factors should be controlled and accounted for to isolate the effects of the chocolate consumption on the study participants.
- Blinding: The study could also usea double blind format, in which neither the participants nor researchers know which group is receiving the chocolate. This would prevent bias in the results found.

Exercise 3.9 parts (d) and (e)

(d)

```r
mix_1 <- c(3129, 3000, 2865, 2890)
mix_2 <- c(3200, 3300, 2975, 3150)
mix_3 <- c(2800, 2900, 2985, 3050)
mix_4 <- c(2600, 2700, 2600, 2765)

tensile_strength_df <- data.frame(
  strength = c(mix_1, mix_2, mix_3, mix_4),
  technique = c(rep("mix_1", 4),
                rep("mix_2", 4),
                rep("mix_3", 4),
                rep("mix_4", 4))

)

model <- aov(strength ~ technique, data=tensile_strength_df)

residuals <- residuals(model)

qqnorm(residuals)
qqline(residuals)
```

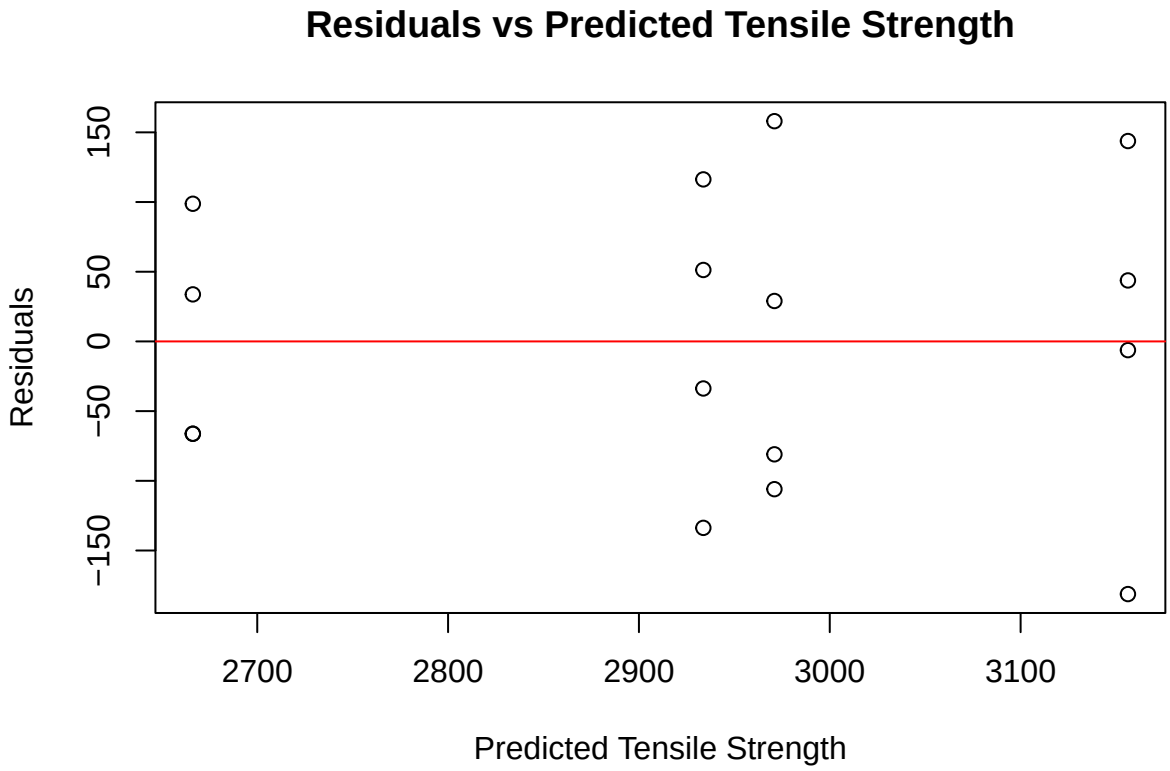


- The Q-Q plot shows that most points lie fairly close to the diagonal line, indicating that the residuals are approximately normally distributed. However, there are some deviations at the tails (specifically, a slight upward curve at the top right and downward curve at the bottom left).

- Conclusion: The normality assumption seems reasonable, but the deviations at the tails might suggest a the data is not perfectly normal.

(e)

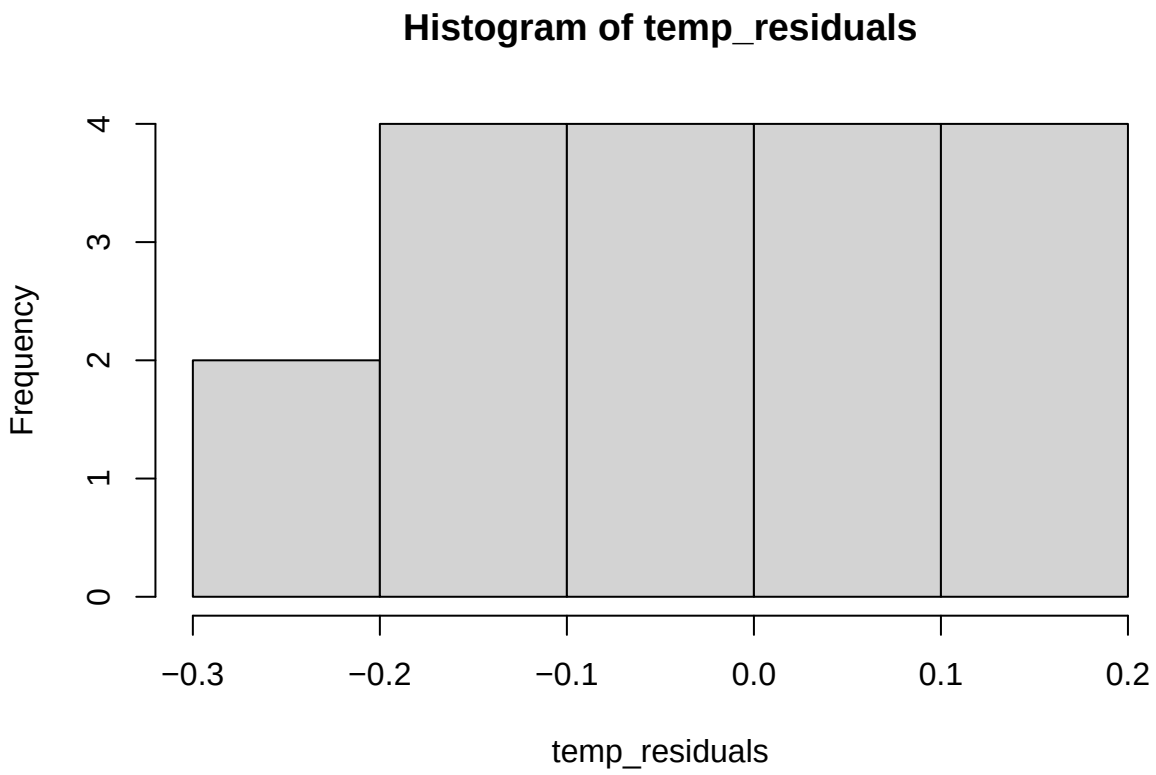
```r
predicted <- fitted(model)

plot(predicted, residuals,
     xlab = "Predicted Tensile Strength",
     ylab = "Residuals",
     main = "Residuals vs Predicted Tensile Strength")

abline(h = 0, col = "red")
```



- The residuals appear to be scattered randomly around the horizontal line at zero, with no clear pattern or trend.
- Conclusion: This indicates that the residuals have constant variance (homoscedasticity), and there is no indication of non-linearity or other problematic patterns. The residuals vs. predicted plot suggests that the model assumptions regarding constant variance are met.

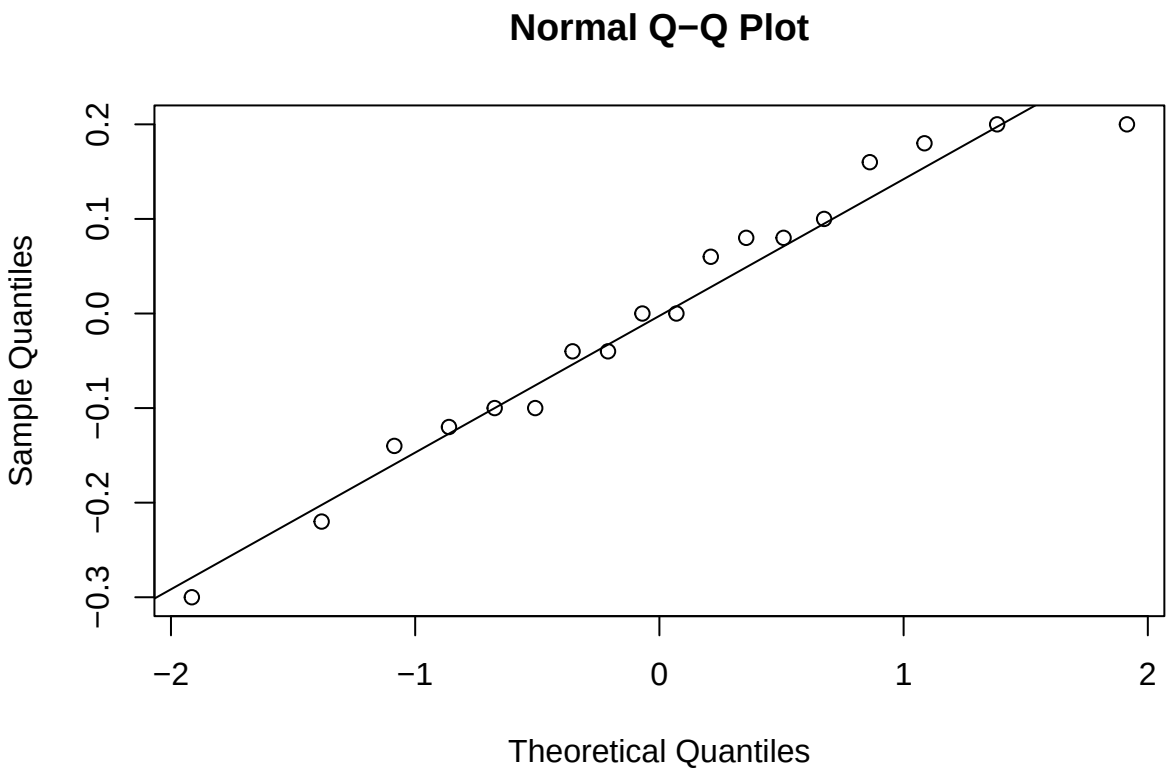
Exercise 3.18 part (c) only

```
model3 <- aov(density ~ method, data=temp_df)

temp_residuals <- residuals(model3)

hist(temp_residuals)
```



- The residuals have equal frequency among values -0.2 to 0.2, with half that from -0.3 to -0.2.

```
qqnorm(temp_residuals)
qqline(temp_residuals)
```



- The Q-Q plot shows that all but one of the points lie very close to the diagonal line, indicating that the residuals are approximately normally distributed. There seems to be one outlier that deviates near the top of the graph, but it doesn't seem to be too extreme.

```
shapiro.test(temp_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  temp_residuals
## W = 0.95925, p-value = 0.5873
```

- Based on our p-value of 0.5873 which is greater than our $\alpha = 0.05$, we fail to reject the null hypothesis that the residuals are normally distributed.

```
names(model3)
```

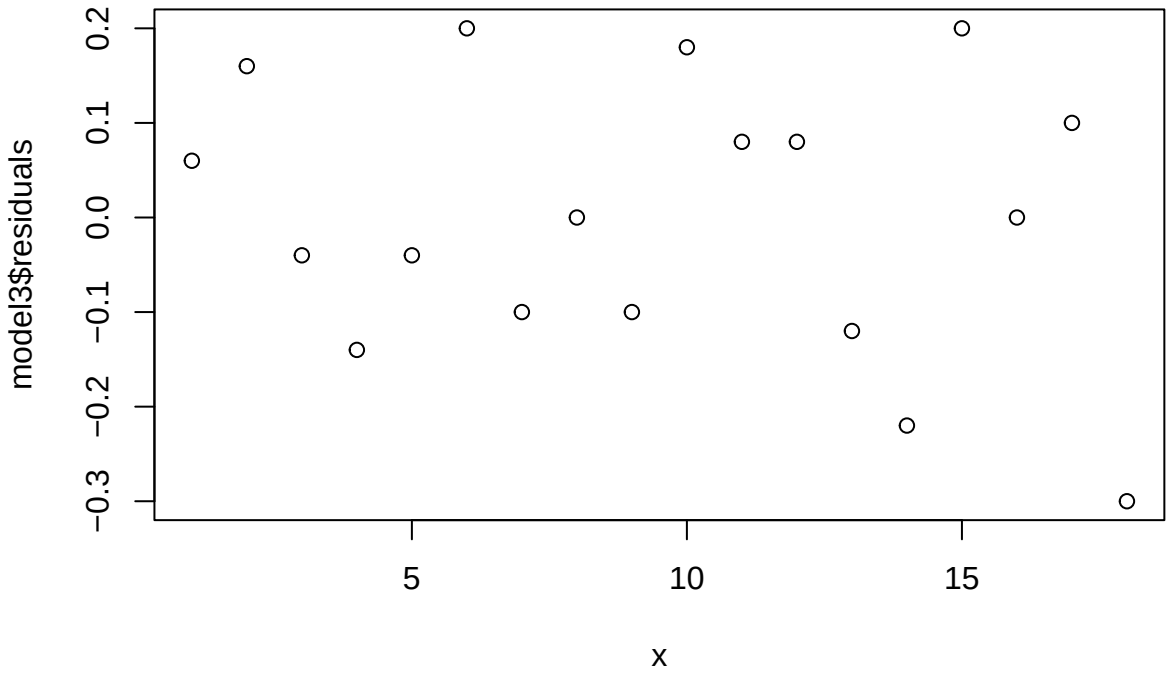
```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "contrasts"     "xlevels"       "call"          "terms"
## [13] "model"
```

```
model3$residuals
```

```
##             1             2             3             4             5
##  6.000000e-02  1.600000e-01 -4.000000e-02 -1.400000e-01 -4.000000e-02
##             6             7             8             9            10
##  2.000000e-01 -1.000000e-01  9.576148e-16 -1.000000e-01  1.800000e-01
##            11            12            13            14            15
##  8.000000e-02  8.000000e-02 -1.200000e-01 -2.200000e-01  2.000000e-01
##            16            17            18
## -1.687297e-18  1.000000e-01 -3.000000e-01
```

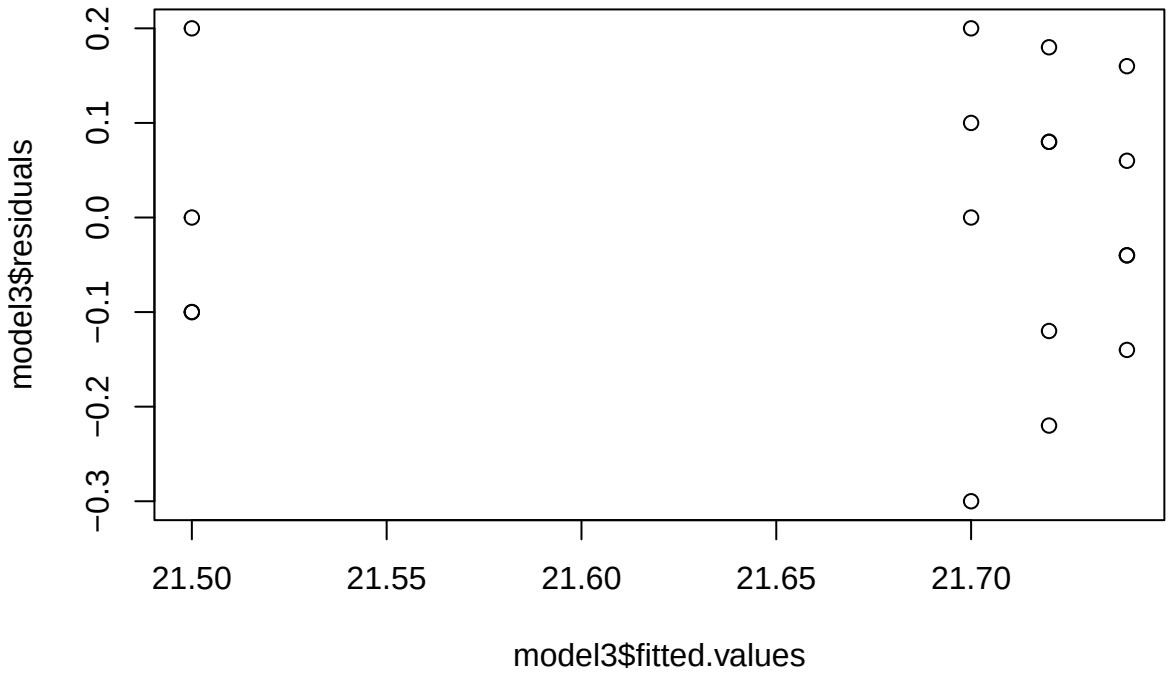
- The residuals in model3$residuals show a range of positive and negative values, with some very close to zero. These near-zero values indicate observations that align almost exactly with the model's predictions, which suggests a good fit for those specific points.

```
x <- 1:length(model3$residuals)
plot(model3$residuals ~ x)
```



- Based on the graph, there doesn't appear to be any significant structure to the data. Values aren't systematically tending to go up or down as the data are collected over time.

```
plot(model3$residuals~model3$fitted.values)
```



- The residuals seem to cluster more on the higher end of fitted values, with a greater spread on the positive side. This uneven spread implies that the residuals might not have constant variance across fitted values, which suggests a potential violation of the homoscedasticity anova assumption.

Conclusion:

1. Normality of Residuals: The Shapiro-Wilk test for normality ($W = 0.95925$, p-value $= 0.5873$) indicates that the residuals do not significantly deviate from normality, as the p-value is well above the standard significance level.
2. No Structure in Residuals: The residuals vs. order plot shows no clear trend in the residuals over time, suggesting independence in data collection. Thus, we can reasonably conclude that the residuals are independent.
3. Equal Variances: The residuals vs. fitted values plot, however, may suggest heteroscedasticity, as there is a greater spread in residuals at certain fitted values, with residuals clustering more on the higher end. This uneven spread could indicate a violation of the equal variance assumption.

While the normality and independence assumptions appear to be satisfied, the assumption of equal variances may be in question. This potential heteroscedasticity could limit the reliability of the ANOVA results.