

4160 project

Jiayi Niu, Justin Chuderewicz, Sampath Ravva, Spencer Edwards

2023-04-24

The specific goal of this study is to investigate how four factors—length of putt (10 or 30 feet), type of putter (mallet or cavity-back), break of putt (breaking or straight), and slope of putt (level or downhill)—influence putting accuracy in golf. The study aims to determine the significance of these factors and their interactions on the distance from the ball to the center of the cup (in inches) after the ball comes to rest. A response of zero indicates that the putt was made. Ultimately, the study seeks to provide insights and recommendations for golfers to improve their putting accuracy by understanding the impact of these factors.

```
y <- c(10, 18, 14, 12.5, 19, 16, 18.5,
      0, 16.5, 4.5, 17.5, 20.5, 17.5, 33,
      4, 6, 1, 14.5, 12, 14, 5,
      0, 10, 34, 11, 25.5, 21.5, 0,
      0, 0, 18.5, 19.5, 16, 15, 11,
      5, 20.5, 18, 20, 29.5, 19, 10,
      6.5, 18.5, 7.5, 6, 0, 10, 0,
      16.5, 4.5, 0, 23.5, 8, 8, 8,
      4.5, 18, 14.5, 10, 0, 17.5, 6,
      19.5, 18, 16, 5.5, 10, 7, 36,
      15, 16, 8.5, 0, 0.5, 9, 3,
      41.5, 39, 6.5, 3.5, 7, 8.5, 36,
      8, 4.5, 6.5, 10, 13, 41, 14,
      21.5, 10.5, 6.5, 0, 15.5, 24, 16,
      0, 0, 0, 4.5, 1, 4, 6.5,
      18, 5, 7, 10, 32.5, 18.5, 8)

lengths <- rep(rep(c(1, 2), each = 7), 8) # 1 for 10ft, 2 for 30ft
putters <- rep(rep(c(1, 2), each = 14), 4) # 1 for Mallet (M), 2 for Cavity-Back (CB)
breaks <- rep(rep(c(1, 2), each = 28), 2) # 1 for Straight, 2 for Breaking
slopes <- rep(c(1, 2), each = 56) # 1 for Level, 2 for Downhill

data <- data.frame(Length = lengths , Putter = putters, Break = breaks, Slope = slopes, Response = y)
```

First we can do the exploratory data analysis to identify general patterns in the data

```
library(gridExtra)
library(ggplot2)

plot1<-ggplot(data, aes(x = factor(lengths), y = y)) +
  geom_boxplot() +
  xlab("Length (1: 10ft, 2: 30ft)") +
  ylab("Distance from the center of the cup (inches)") +
  ggtitle("Boxplot of Length vs Response")
```

```

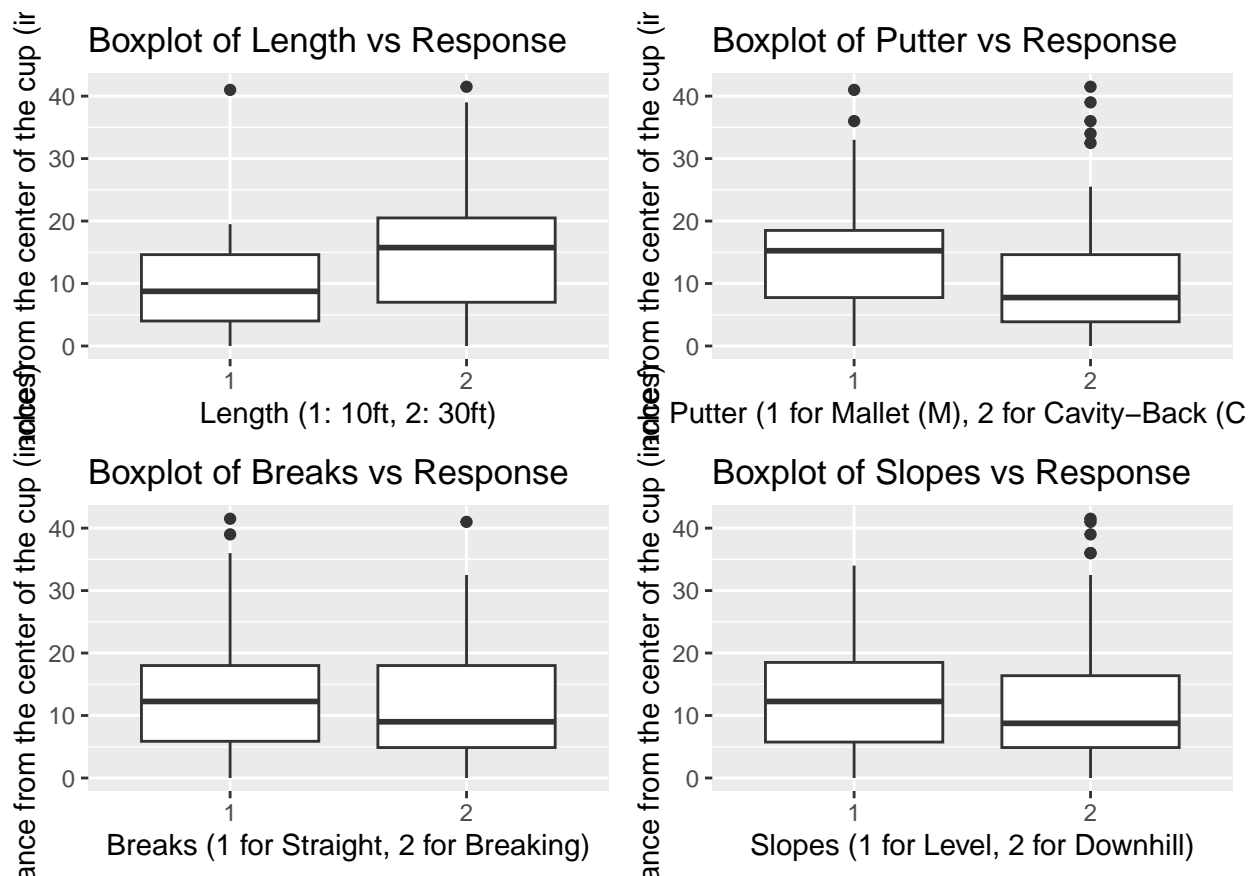
plot2<-ggplot(data, aes(x = factor(putters), y = y)) +
  geom_boxplot() +
  xlab("Putter (1 for Mallet (M), 2 for Cavity-Back (CB))") +
  ylab("Distance from the center of the cup (inches)") +
  ggtitle("Boxplot of Putter vs Response")

plot3<-ggplot(data, aes(x = factor(breaks), y = y)) +
  geom_boxplot() +
  xlab("Breaks (1 for Straight, 2 for Breaking)") +
  ylab("Distance from the center of the cup (inches)") +
  ggtitle("Boxplot of Breaks vs Response")

plot4<-ggplot(data, aes(x = factor(slopes), y = y)) +
  geom_boxplot() +
  xlab("Slopes (1 for Level, 2 for Downhill)") +
  ylab("Distance from the center of the cup (inches)") +
  ggtitle("Boxplot of Slopes vs Response")

grid.arrange(plot1, plot2, plot3, plot4, nrow=2, ncol=2)

```



From the boxplots, we can see the general distributions between the exploratory variables and response variable. Since for length and putter variable, there are significant differences in distances as the length and putter changes. We should pay more attention into these two variables and see if it is true that the effects are significant. We should also keep in mind that EDA is only the first step in the analysis process. We need to perform further statistical tests or use more advanced techniques, such as regression or ANOVA, to make

specific conclusions or predictions based on the data.

This dataset is a 2^4 factorial design because it involves four factors (Length, Putter, Break, and Slope), and each factor has two levels. A factorial design is an experimental design that examines the effects of multiple factors and their interactions on a response variable. The number of levels in each factor is raised to the power of the number of factors to determine the total number of experimental conditions or treatments. In this case, it's 2^4 , which equals 16 experimental conditions.

The 2^4 factorial model is: $Y_{ijklp} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_p + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + (\alpha\beta\gamma\delta)_{ijkl} + \epsilon_{ijklp}$ where $\epsilon \sim N(0, \sigma^2)$

```
model1<-aov(y~as.factor(lengths)*as.factor(putters)*as.factor(breaks)*as.factor(slopes))
summary(model1)
```

```
##                                                                 Df
## as.factor(lengths)                                           1
## as.factor(putters)                                           1
## as.factor(breaks)                                            1
## as.factor(slopes)                                            1
## as.factor(lengths):as.factor(putters)                       1
## as.factor(lengths):as.factor(breaks)                        1
## as.factor(putters):as.factor(breaks)                        1
## as.factor(lengths):as.factor(slopes)                        1
## as.factor(putters):as.factor(slopes)                        1
## as.factor(breaks):as.factor(slopes)                         1
## as.factor(lengths):as.factor(putters):as.factor(breaks)    1
## as.factor(lengths):as.factor(putters):as.factor(slopes)    1
## as.factor(lengths):as.factor(breaks):as.factor(slopes)     1
## as.factor(putters):as.factor(breaks):as.factor(slopes)     1
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 1
## Residuals                                                    96
##                                                                 Sum Sq
## as.factor(lengths)                                           917
## as.factor(putters)                                           388
## as.factor(breaks)                                            145
## as.factor(slopes)                                            1
## as.factor(lengths):as.factor(putters)                       219
## as.factor(lengths):as.factor(breaks)                        12
## as.factor(putters):as.factor(breaks)                        115
## as.factor(lengths):as.factor(slopes)                        94
## as.factor(putters):as.factor(slopes)                        56
## as.factor(breaks):as.factor(slopes)                         2
## as.factor(lengths):as.factor(putters):as.factor(breaks)     7
## as.factor(lengths):as.factor(putters):as.factor(slopes)    113
## as.factor(lengths):as.factor(breaks):as.factor(slopes)     39
## as.factor(putters):as.factor(breaks):as.factor(slopes)     34
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 96
## Residuals                                                    8316
##                                                                 Mean Sq
## as.factor(lengths)                                           917.1
## as.factor(putters)                                           388.1
## as.factor(breaks)                                            145.1
## as.factor(slopes)                                            1.4
## as.factor(lengths):as.factor(putters)                       218.7
## as.factor(lengths):as.factor(breaks)                        11.9
```

```

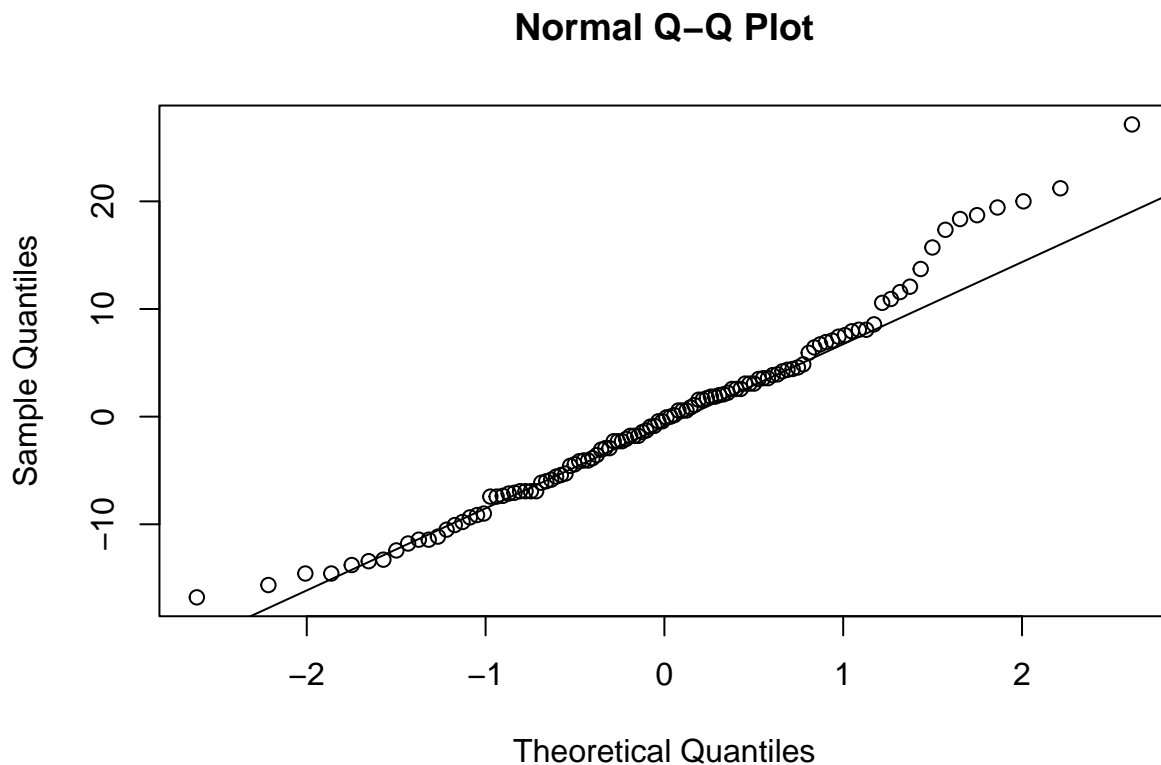
## as.factor(putters):as.factor(breaks) 115.0
## as.factor(lengths):as.factor(slopes) 93.8
## as.factor(putters):as.factor(slopes) 56.4
## as.factor(breaks):as.factor(slopes) 1.6
## as.factor(lengths):as.factor(putters):as.factor(breaks) 7.3
## as.factor(lengths):as.factor(putters):as.factor(slopes) 113.0
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 39.5
## as.factor(putters):as.factor(breaks):as.factor(slopes) 33.8
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 95.6
## Residuals 86.6
## F value
## as.factor(lengths) 10.588
## as.factor(putters) 4.481
## as.factor(breaks) 1.676
## as.factor(slopes) 0.016
## as.factor(lengths):as.factor(putters) 2.525
## as.factor(lengths):as.factor(breaks) 0.137
## as.factor(putters):as.factor(breaks) 1.328
## as.factor(lengths):as.factor(slopes) 1.083
## as.factor(putters):as.factor(slopes) 0.651
## as.factor(breaks):as.factor(slopes) 0.019
## as.factor(lengths):as.factor(putters):as.factor(breaks) 0.084
## as.factor(lengths):as.factor(putters):as.factor(slopes) 1.305
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 0.456
## as.factor(putters):as.factor(breaks):as.factor(slopes) 0.390
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 1.104
## Residuals
## Pr(>F)
## as.factor(lengths) 0.00157
## as.factor(putters) 0.03686
## as.factor(breaks) 0.19862
## as.factor(slopes) 0.89928
## as.factor(lengths):as.factor(putters) 0.11538
## as.factor(lengths):as.factor(breaks) 0.71178
## as.factor(putters):as.factor(breaks) 0.25205
## as.factor(lengths):as.factor(slopes) 0.30066
## as.factor(putters):as.factor(slopes) 0.42159
## as.factor(breaks):as.factor(slopes) 0.89127
## as.factor(lengths):as.factor(putters):as.factor(breaks) 0.77294
## as.factor(lengths):as.factor(putters):as.factor(slopes) 0.25623
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 0.50121
## as.factor(putters):as.factor(breaks):as.factor(slopes) 0.53386
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 0.29599
## Residuals
##
## as.factor(lengths) **
## as.factor(putters) *
## as.factor(breaks)
## as.factor(slopes)
## as.factor(lengths):as.factor(putters)
## as.factor(lengths):as.factor(breaks)
## as.factor(putters):as.factor(breaks)
## as.factor(lengths):as.factor(slopes)
## as.factor(putters):as.factor(slopes)

```

```
## as.factor(breaks):as.factor(slopes)
## as.factor(lengths):as.factor(putters):as.factor(breaks)
## as.factor(lengths):as.factor(putters):as.factor(slopes)
## as.factor(lengths):as.factor(breaks):as.factor(slopes)
## as.factor(putters):as.factor(breaks):as.factor(slopes)
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes)
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

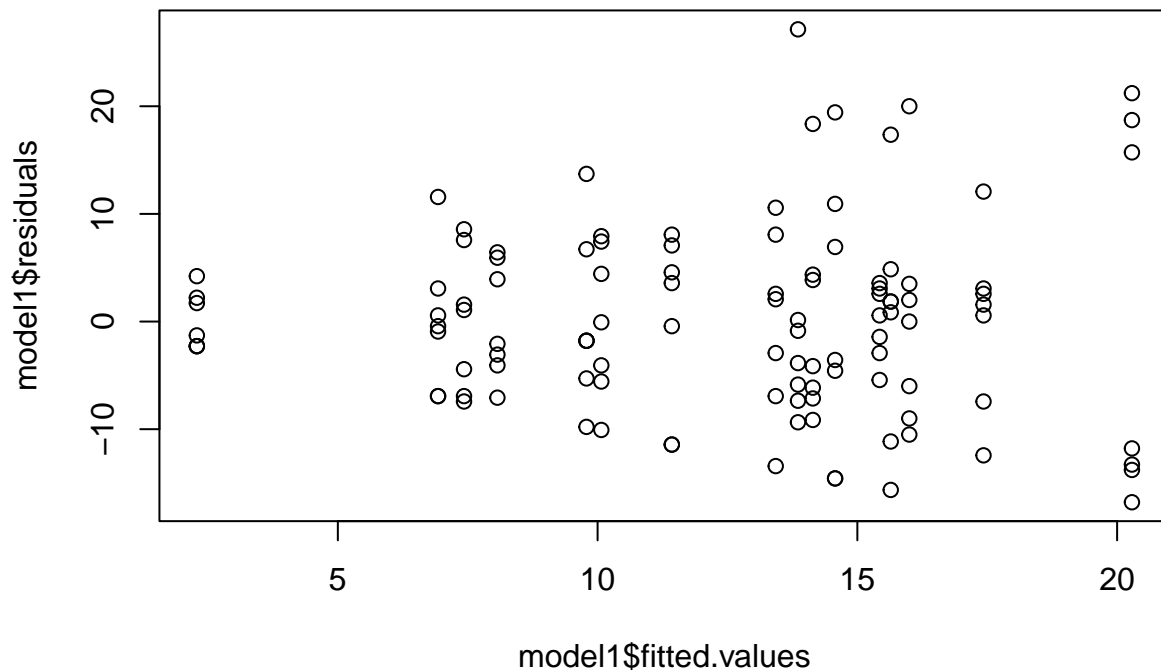
To ensure the validity and accuracy of the ANOVA tests, we should first check the two assumptions of the 2^k factorial experiment, which are assumption of normality and assumption of constant variance.

```
qqnorm(model1$residuals)
qqline(model1$residuals)
```



Since the datapoints generally fall onto the qq-line except some minor deviations, the assumption of normality is met. We will check the assumption of constant variance by the residual plot as follows:

```
plot(model1$fitted.values, model1$residuals)
```



Based on the residual plot, the residuals do not generally show a constant vertical distances across x and predicted values as there is a fanning-out pattern. This pattern violates the assumption of constant variance and can lead to biased and inefficient parameter estimates.

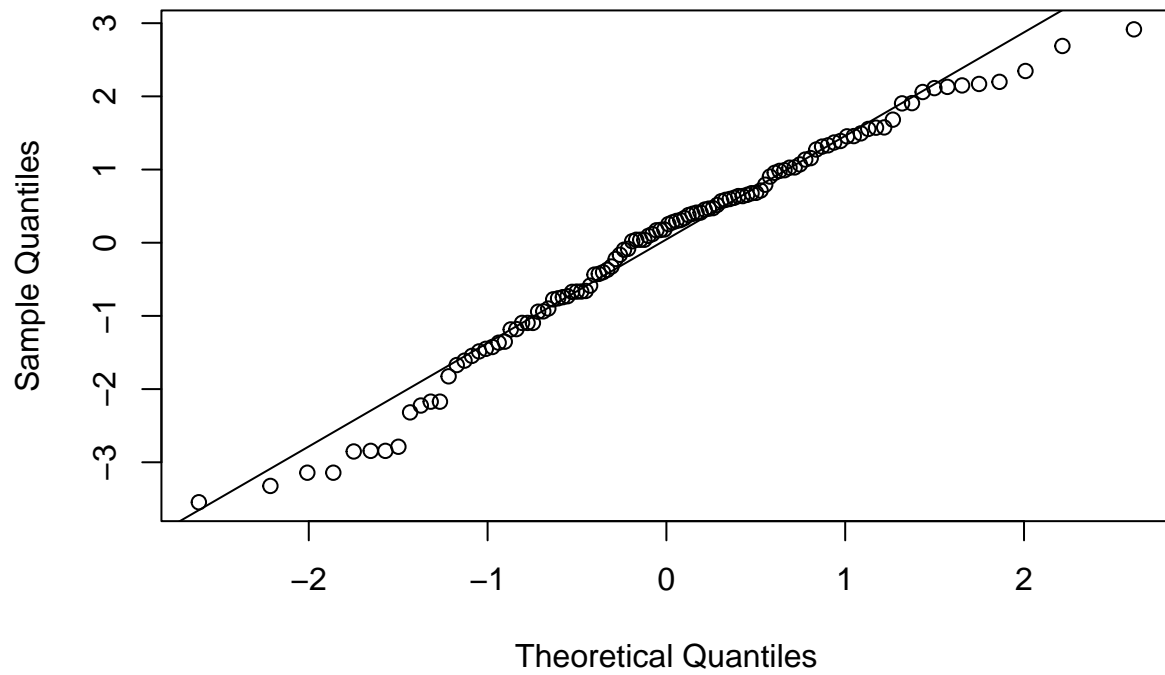
Since the constant variance assumption is not met, we should considering transforming the variables. We would transform the y variable to stabilize the variance by taking the square root of y because there is an upward trend shown in the normal qq-plot. As we go down the ladder of transformation, it is important to note that there are values with 0, so log transformation cannot be considered. Thus, we choose the power of $\frac{1}{2}$. The transformed y is now: $y^* = \sqrt{y}$

```
data$ystar<-sqrt(data$Response)
model1.ystar<-aov(ystar ~ as.factor(lengths)*as.factor(putters)
                  *as.factor(breaks)*as.factor(slopes),data=data)
```

Now we can check the assumptions again to make sure that it is a successful transformation.

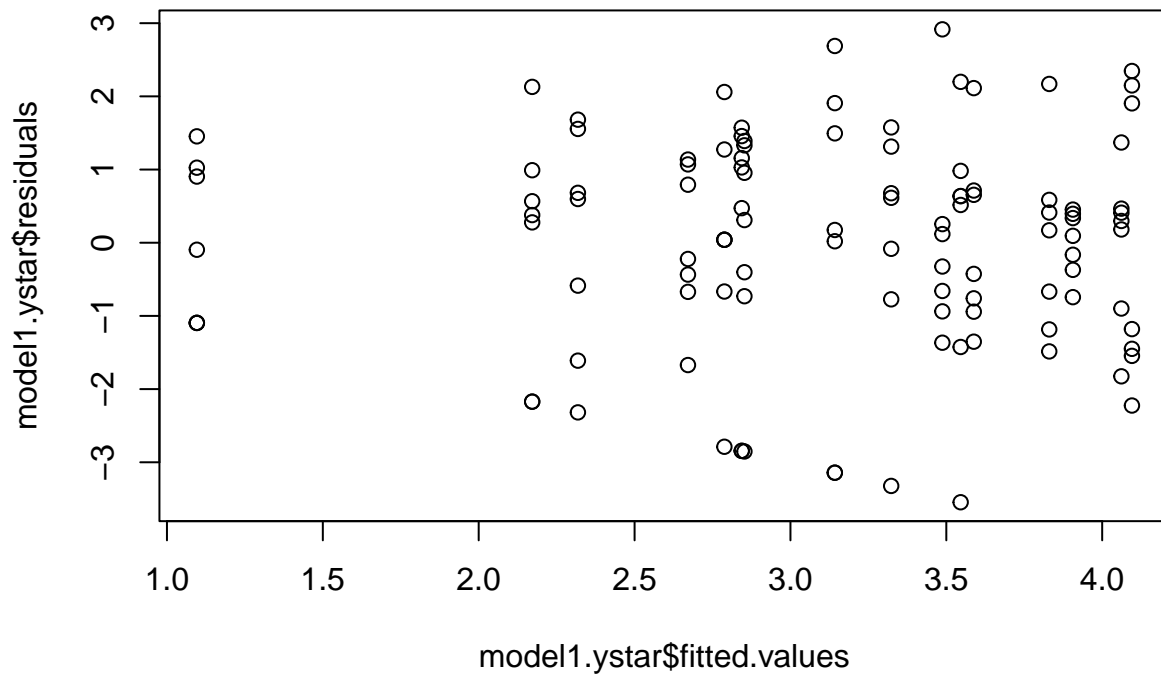
```
qqnorm(model1.ystar$residuals)
qqline(model1.ystar$residuals)
```

Normal Q-Q Plot



The assumption of normality is met because the datapoint generally fall onto the qq-line based on the qq-plot.

```
plot(model1.ystar$fitted.values, model1.ystar$residuals)
```



Assumption of constant variance is met because the residuals show a constant vertical distances across x and predicted values. As the assumption of normality and constant variance are all met. We can continue variable screening.

```
summary(model1.ystar)
```

##	Df
## as.factor(lengths)	1
## as.factor(putters)	1
## as.factor(breaks)	1
## as.factor(slopes)	1
## as.factor(lengths):as.factor(putters)	1
## as.factor(lengths):as.factor(breaks)	1
## as.factor(putters):as.factor(breaks)	1
## as.factor(lengths):as.factor(slopes)	1
## as.factor(putters):as.factor(slopes)	1
## as.factor(breaks):as.factor(slopes)	1
## as.factor(lengths):as.factor(putters):as.factor(breaks)	1
## as.factor(lengths):as.factor(putters):as.factor(slopes)	1
## as.factor(lengths):as.factor(breaks):as.factor(slopes)	1
## as.factor(putters):as.factor(breaks):as.factor(slopes)	1
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes)	1
## Residuals	96
##	Sum Sq
## as.factor(lengths)	21.61
## as.factor(putters)	15.64


```

## as.factor(breaks) 3.94
## as.factor(slopes) 0.13
## as.factor(lengths):as.factor(putters) 5.94
## as.factor(lengths):as.factor(breaks) 0.74
## as.factor(putters):as.factor(breaks) 2.05
## as.factor(lengths):as.factor(slopes) 4.31
## as.factor(putters):as.factor(slopes) 0.62
## as.factor(breaks):as.factor(slopes) 0.02
## as.factor(lengths):as.factor(putters):as.factor(breaks) 0.08
## as.factor(lengths):as.factor(putters):as.factor(slopes) 4.55
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 2.02
## as.factor(putters):as.factor(breaks):as.factor(slopes) 1.05
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 4.73
## Residuals 228.45
## Mean Sq
## as.factor(lengths) 21.613
## as.factor(putters) 15.644
## as.factor(breaks) 3.944
## as.factor(slopes) 0.127
## as.factor(lengths):as.factor(putters) 5.943
## as.factor(lengths):as.factor(breaks) 0.735
## as.factor(putters):as.factor(breaks) 2.051
## as.factor(lengths):as.factor(slopes) 4.308
## as.factor(putters):as.factor(slopes) 0.618
## as.factor(breaks):as.factor(slopes) 0.018
## as.factor(lengths):as.factor(putters):as.factor(breaks) 0.079
## as.factor(lengths):as.factor(putters):as.factor(slopes) 4.554
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 2.023
## as.factor(putters):as.factor(breaks):as.factor(slopes) 1.052
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 4.734
## Residuals 2.380
## F value
## as.factor(lengths) 9.082
## as.factor(putters) 6.574
## as.factor(breaks) 1.657
## as.factor(slopes) 0.054
## as.factor(lengths):as.factor(putters) 2.497
## as.factor(lengths):as.factor(breaks) 0.309
## as.factor(putters):as.factor(breaks) 0.862
## as.factor(lengths):as.factor(slopes) 1.810
## as.factor(putters):as.factor(slopes) 0.260
## as.factor(breaks):as.factor(slopes) 0.007
## as.factor(lengths):as.factor(putters):as.factor(breaks) 0.033
## as.factor(lengths):as.factor(putters):as.factor(slopes) 1.914
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 0.850
## as.factor(putters):as.factor(breaks):as.factor(slopes) 0.442
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 1.989
## Residuals
## Pr(>F)
## as.factor(lengths) 0.0033
## as.factor(putters) 0.0119
## as.factor(breaks) 0.2011
## as.factor(slopes) 0.8175
## as.factor(lengths):as.factor(putters) 0.1173

```

```

## as.factor(lengths):as.factor(breaks) 0.5796
## as.factor(putters):as.factor(breaks) 0.3556
## as.factor(lengths):as.factor(slopes) 0.1816
## as.factor(putters):as.factor(slopes) 0.6115
## as.factor(breaks):as.factor(slopes) 0.9315
## as.factor(lengths):as.factor(putters):as.factor(breaks) 0.8559
## as.factor(lengths):as.factor(putters):as.factor(slopes) 0.1698
## as.factor(lengths):as.factor(breaks):as.factor(slopes) 0.3589
## as.factor(putters):as.factor(breaks):as.factor(slopes) 0.5078
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes) 0.1616
## Residuals
##
## as.factor(lengths) **
## as.factor(putters) *
## as.factor(breaks)
## as.factor(slopes)
## as.factor(lengths):as.factor(putters)
## as.factor(lengths):as.factor(breaks)
## as.factor(putters):as.factor(breaks)
## as.factor(lengths):as.factor(slopes)
## as.factor(putters):as.factor(slopes)
## as.factor(breaks):as.factor(slopes)
## as.factor(lengths):as.factor(putters):as.factor(breaks)
## as.factor(lengths):as.factor(putters):as.factor(slopes)
## as.factor(lengths):as.factor(breaks):as.factor(slopes)
## as.factor(putters):as.factor(breaks):as.factor(slopes)
## as.factor(lengths):as.factor(putters):as.factor(breaks):as.factor(slopes)
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

When we think about simplifying our ANOVA model (i.e. screening unimportant factors and interactions), we want to abide by three principles of interaction: Based on the anova table, 1. Effect Sparsity Principle: The number of relatively important effects in a factorial experiment is small, so that factor and interaction screening is a good idea. 2. Effect Hierarchy Principle: This principle states that lower-order effects are more likely to be significant than higher-order effects. In this case, there are no significant higher-order interactions, so we can safely ignore them for now. 3. Effect Heredity Principle: This principle states that if an interaction is significant, its corresponding main effects should also be included in the model, even if they are not significant by themselves. In this case, there are no significant interactions, so this principle doesn't apply.

Specifically, the order should be for four-way interactions: $A * B * C * D$ insignificant, then we remove it from the model, for $\{A,B,C,D\}=\{\text{lengths}, \text{putters}, \text{breaks}, \text{slopes}\}$

For three-way interactions: $A * B * C$ insignificant, then we remove them from the model

For two-way interactions: $A * B$ insignificant, then we remove them from the model

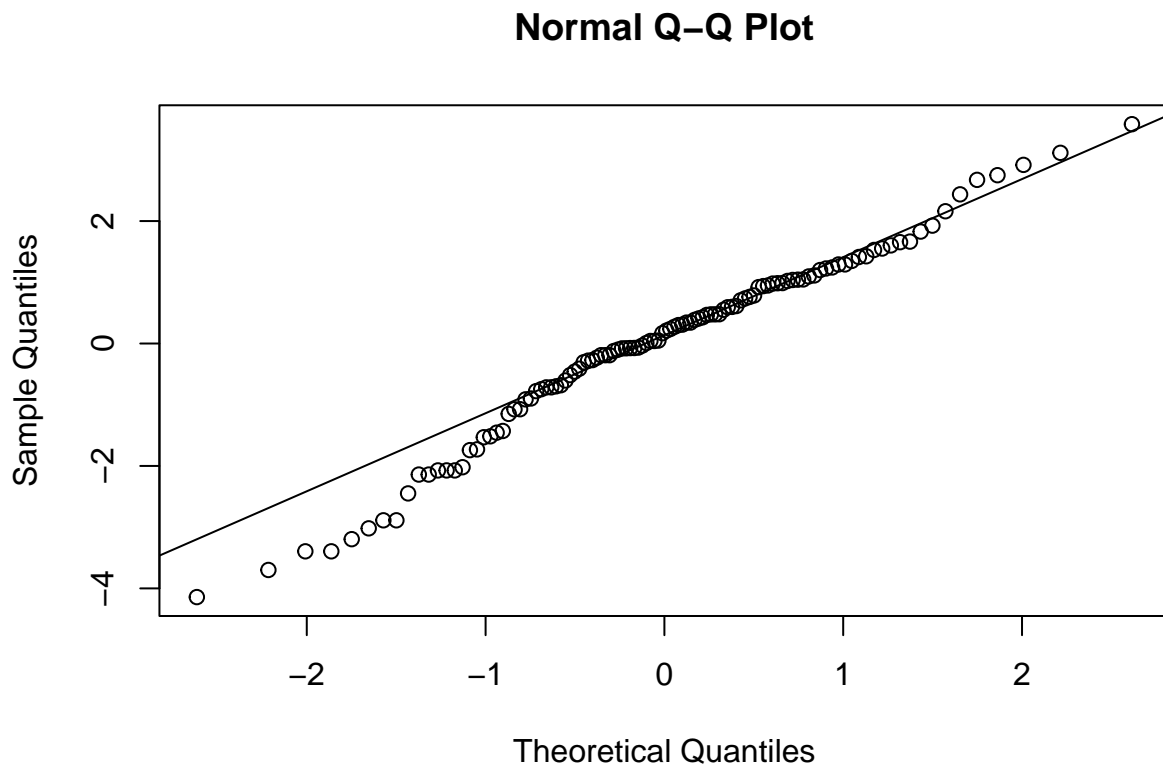
Thus, the model is now only with the main effects such that $Y = \text{lengths} + \text{putters} + \text{breaks} + \text{slopes} + \epsilon$

```
reduced1<-aov(ystar ~ as.factor(lengths)+as.factor(putters)
              +as.factor(breaks)+as.factor(slopes),data=data)
summary(reduced1)
```

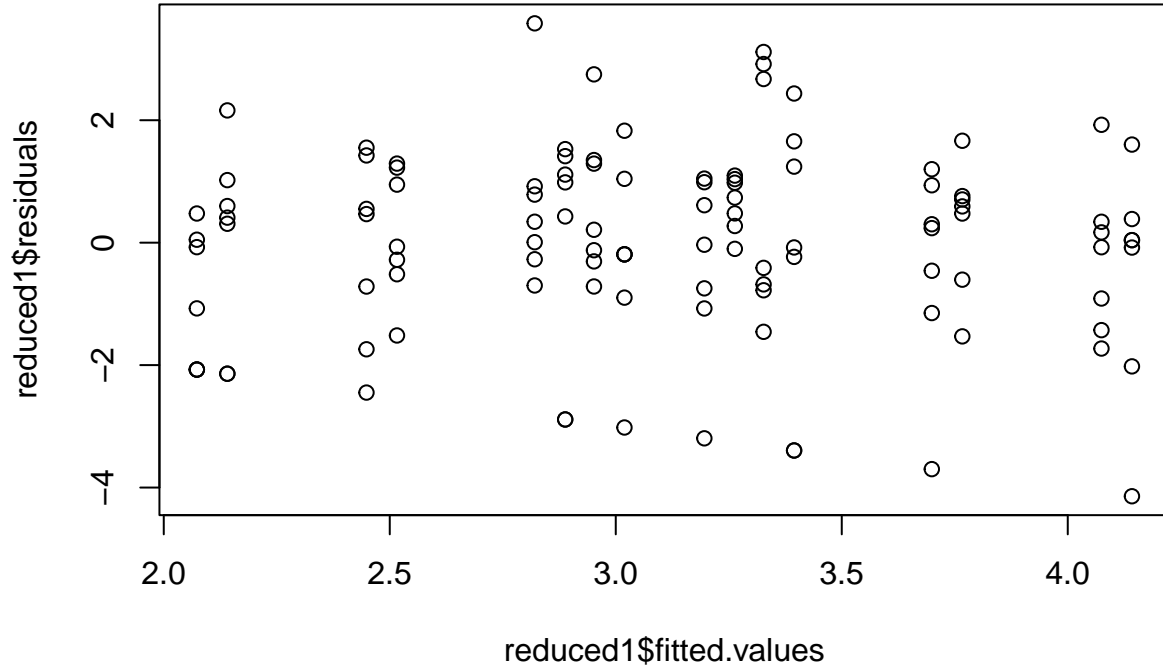
```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(lengths)    1   21.61   21.613    9.085 0.00322 **
## as.factor(putters)    1   15.64   15.644    6.576 0.01173 *
## as.factor(breaks)     1    3.94    3.944    1.658 0.20068
## as.factor(slopes)     1    0.13    0.127    0.054 0.81743
## Residuals           107 254.56    2.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We should check assumptions again to ensure the validity of the test results.

```
qqnorm(reduced1$residuals)
qqline(reduced1$residuals)
```



```
plot(reduced1$fitted.values, reduced1$residuals)
```



The assumption of normality is generally met because the data-points approximately fall onto the qq-line. The residuals have constant vertical distances across x and predicted values. Therefore, we can continue variable screening.

$$H_0: \beta_{breaks} = 0$$

$$H_a: \beta_{breaks} \neq 0$$

The test statistic is 1.658 and the p-value is 0.20068. Since the p-value is greater than 0.05 significance level, there is not sufficient evidence against the null hypothesis. We failed to reject the null hypothesis. Factor slopes is not helpful for the model. The p-value of 0.20068 indicates that the probability of observing the given test statistic, or a more extreme one, is 20.068% when assuming the breaks factor has no effect on the response variable.

$$H_0: \beta_{slopes} = 0$$

$$H_a: \beta_{slopes} \neq 0$$

The test statistic is 0.054 and the p-value is 0.81743. Since the p-value is greater than 0.05 significance level, there is not sufficient evidence against the null hypothesis. We failed to reject the null hypothesis. Factor breaks is not helpful for the model. The p-value of 0.81743 indicates that the probability of observing the given test statistic, or a more extreme one, is 81.743% when assuming the slopes factor has no effect on the response variable.

However, we can see that for factor breaks and slopes, the p-values are greater than the significance level at $\alpha=0.05$. There is not sufficient evidence against the null hypothesis that the factor is helpful for the model. As we failed to reject the null hypothesis, we conclude that both breaks and slopes do not have a statistically significant impact on the response variable (putting accuracy) in the current analysis. We can drop them from the model.

Thus, the reduced model should include the two significant main effects: Length and Putter. The reduced model will be simpler and easier to interpret, focusing on the most important factors influencing putting accuracy.

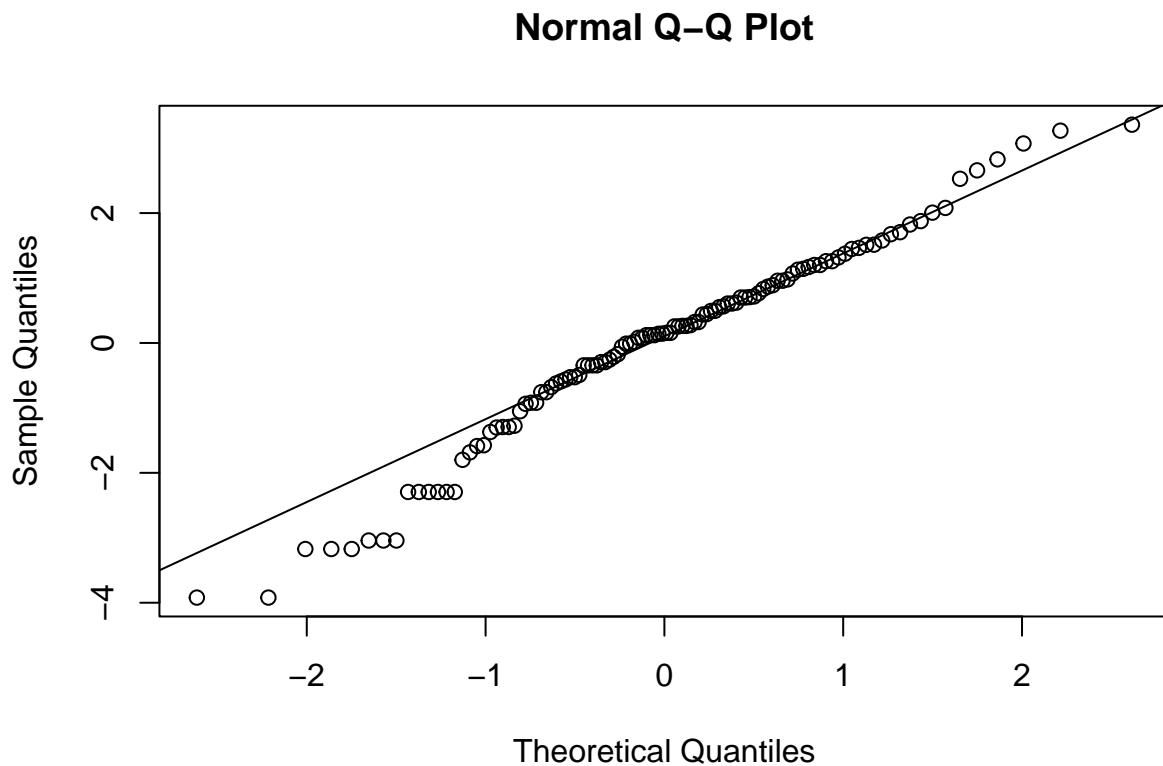
Thus, the reduced model is:

```
reduced2<-aov(ystar ~ as.factor(lengths) + as.factor(putters),data=data)
summary(reduced2)
```

```
##               Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(lengths)    1   21.61   21.613    9.109 0.00317 **
## as.factor(putters)    1   15.64   15.644    6.593 0.01159 *
## Residuals           109  258.63    2.373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

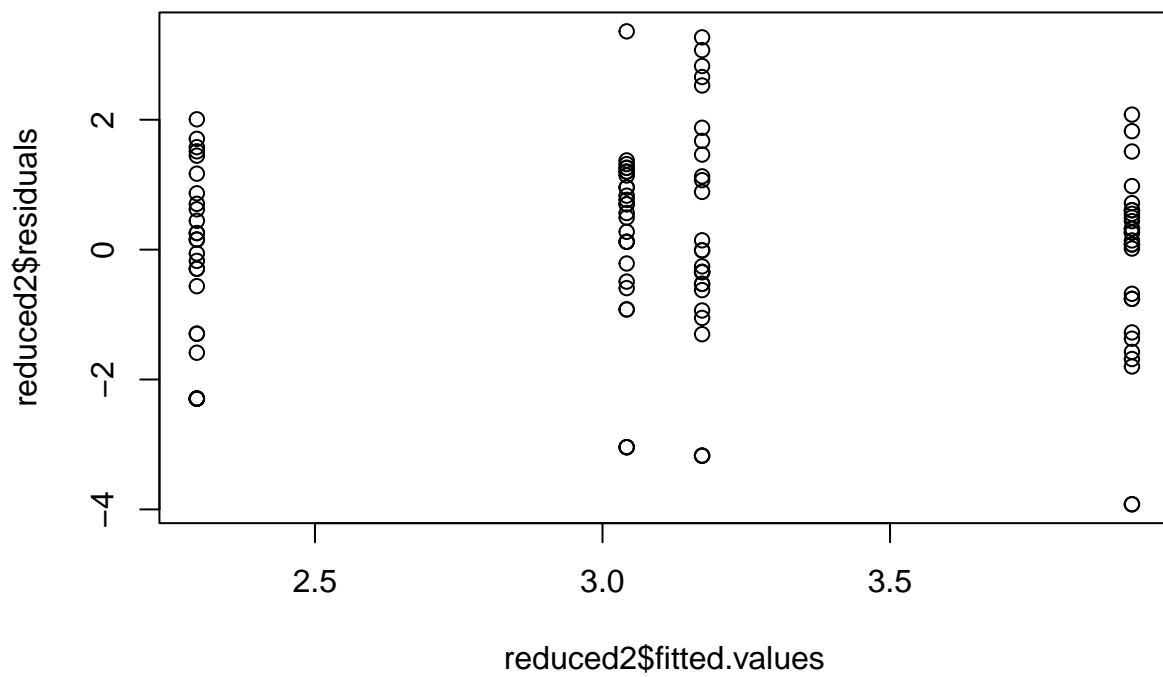
Then we check the assumptions:

```
qqnorm(reduced2$residuals)
qqline(reduced2$residuals)
```



The assumption of normality is generally met because the data-points approximately fall onto the qq-line.

```
plot(reduced2$fitted.values, reduced2$residuals)
```



The constant variance assumption is met.

```
length_means_ystar <- aggregate(ystar ~ Length, data = data, FUN = mean)
length_means_ystar
```

```
##   Length   ystar
## 1      1 2.668327
## 2      2 3.546898
```

```
putter_means_ystar <- aggregate(ystar ~ Putter, data = data, FUN = mean)
putter_means_ystar
```

```
##   Putter   ystar
## 1      1 3.481345
## 2      2 2.733879
```

```
mse_ystar <- 2.373
```

```
length_t_stat_ystar <- abs((2.668327-3.546898) / sqrt(mse_ystar * (2 / 56)))
length_t_stat_ystar
```

```
## [1] 3.017915
```

```
putter_t_stat_ystar <- abs((3.481345-2.733879) / sqrt(mse_ystar * (2 / 56)))
putter_t_stat_ystar
```

```
## [1] 2.567566
```

```
length_t_critical_ystar <- qt(0.975, df = (2 * 2 * 7) - 2 - 2 + 1)
putter_t_critical_ystar <- qt(0.975, df = (2 * 2 * 7) - 2 - 2 + 1)
length_t_stat_ystar > length_t_critical_ystar
```

```
## [1] TRUE
```

```
putter_t_stat_ystar > putter_t_critical_ystar
```

```
## [1] TRUE
```

Length T-statistic: 3.017915 Putter T-statistic: 2.567566 Length Significance: TRUE Putter Significance: TRUE

The T-statistics for both Length and Putter factors are greater than their corresponding critical t-values at a 95% confidence level. As a result, both tests indicate that there is a significant difference in the means of the transformed response variable ystar between the different levels of Length and Putter factors.

In other words, the square root transformation of the Response variable still shows significant differences between the two lengths (10ft and 30ft) and the two types of putters (Mallet and Cavity-Back) in terms of putting accuracy.