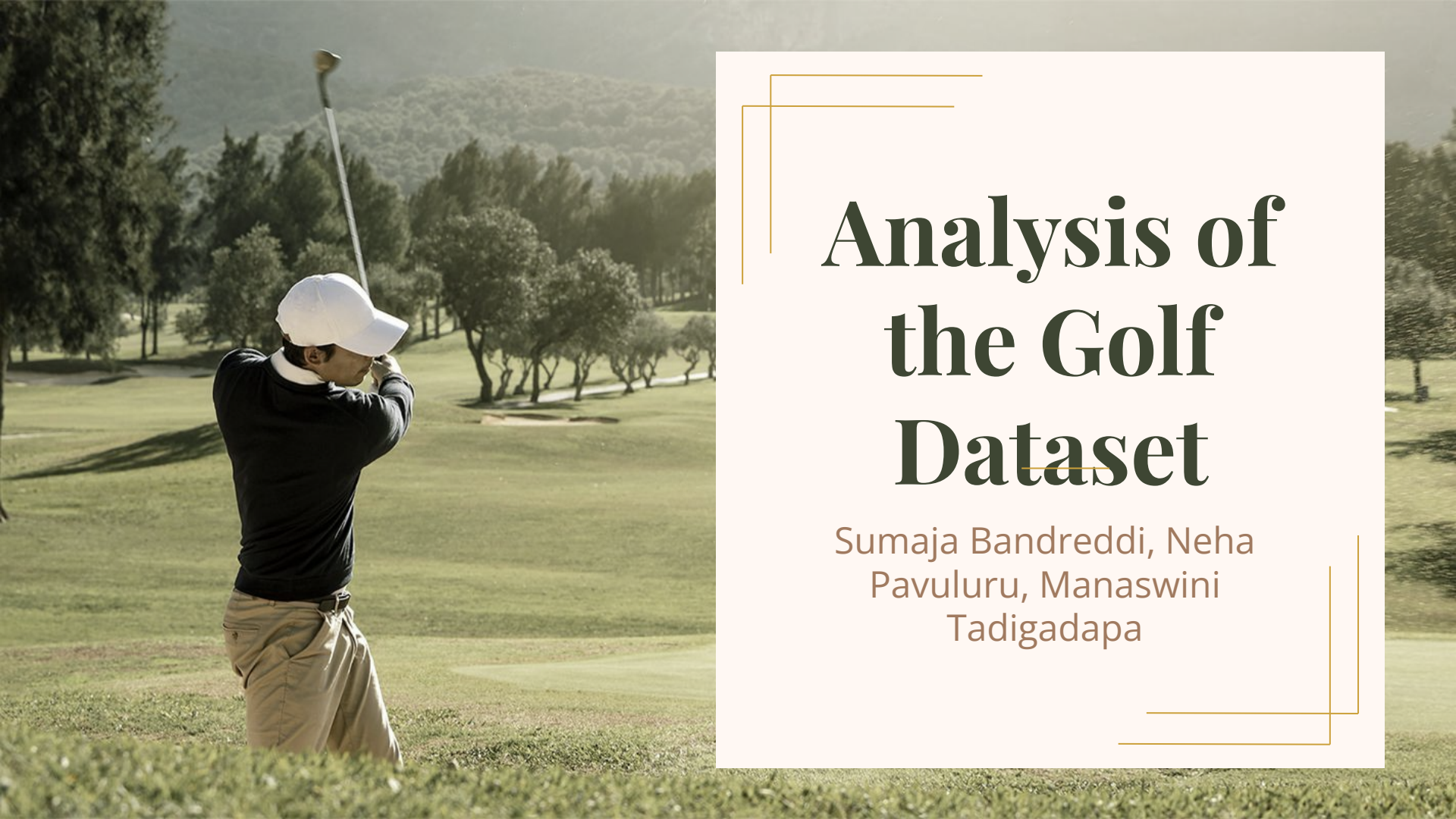


Analysis of the Golf Dataset

Sumaja Bandreddi, Neha Pavuluru, Manaswini Tadigadapa

Introduction

Goal:

Study the effects of 4 factors on putting accuracy

- 1. length of putt (10 or 30 feet),
- 2. type of putter (mallet or cavity-back),
- 3. break of putt (breaking or straight),
- 4. and slope of putt (level or downhill)

Response Variable:

- Putting accuracy is measured by the distance from the ball to the center of the cup (in inches) after the ball comes to rest

Design:

2^k Factorial Experiment with 7 replications

- We conducted a multifactor and multi-step ANOVA to find the significant interactions and main effects on putting accuracy
- We eliminated insignificant effects at each step to reach our final reduced model

Results and Discussion

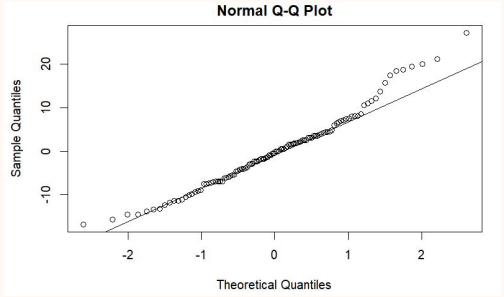
Step 1

- Uploaded the data in R making sure it followed the same order
- Ran an ANOVA test on the full model with all 15 tests
  - The corresponding model:
  - $$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + (\alpha\beta\gamma\delta)_{ijkl} + \epsilon_{ijk}$$
  - Table:

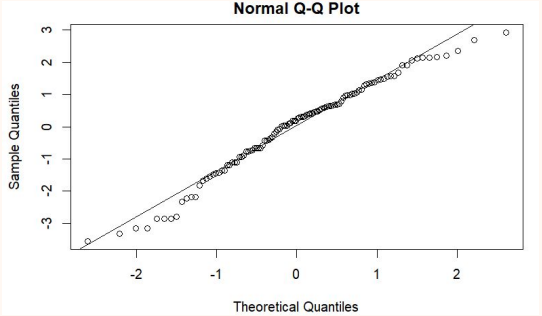
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	917	917.1	10.588	0.00157	**
as.factor(putter)	1	388	388.1	4.481	0.03686	*
as.factor(Break)	1	145	145.1	1.676	0.19862	
as.factor(slope)	1	1	1.4	0.016	0.89928	
as.factor(length):as.factor(putter)	1	219	218.7	2.525	0.11538	
as.factor(length):as.factor(Break)	1	12	11.9	0.137	0.71178	
as.factor(putter):as.factor(Break)	1	115	115.0	1.328	0.25205	
as.factor(length):as.factor(slope)	1	94	93.8	1.083	0.30066	
as.factor(putter):as.factor(slope)	1	56	56.4	0.651	0.42159	
as.factor(Break):as.factor(slope)	1	2	1.6	0.019	0.89127	
as.factor(length):as.factor(putter):as.factor(Break)	1	7	7.3	0.084	0.77294	
as.factor(length):as.factor(putter):as.factor(slope)	1	113	113.0	1.305	0.25623	
as.factor(length):as.factor(Break):as.factor(slope)	1	39	39.5	0.456	0.50121	
as.factor(putter):as.factor(Break):as.factor(slope)	1	34	33.8	0.390	0.53386	
as.factor(length):as.factor(putter):as.factor(Break):as.factor(slope)	1	96	95.6	1.104	0.29599	
Residuals	96	8316	86.6			

Assumptions for Step 1

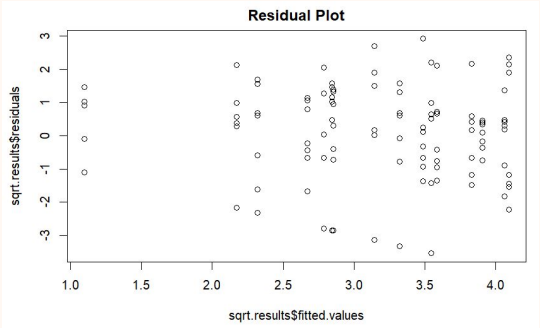
First QQ-plot



Second QQ-plot



Residual plot



We checked the normality assumption through a qq- plot The values seem to be curved upward from the normal distribution line, so we decided to transform by a lower power. We chose to transform by ½ (square root). The second qq-plot fits a normal distribution more closely

We also checked the constant variance assumption and since the points are roughly scattered we concluded that the assumption is met

Table After the Sqrt Transformation

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	21.61	21.613	9.082	0.0033	**
as.factor(putter)	1	15.64	15.644	6.574	0.0119	*
as.factor(Break)	1	3.94	3.944	1.657	0.2011	
as.factor(slope)	1	0.13	0.127	0.054	0.8175	
as.factor(length):as.factor(putter)	1	5.94	5.943	2.497	0.1173	
as.factor(length):as.factor(Break)	1	0.74	0.735	0.309	0.5796	
as.factor(putter):as.factor(Break)	1	2.05	2.051	0.862	0.3556	
as.factor(length):as.factor(slope)	1	4.31	4.308	1.810	0.1816	
as.factor(putter):as.factor(slope)	1	0.62	0.618	0.260	0.6115	
as.factor(Break):as.factor(slope)	1	0.02	0.018	0.007	0.9315	
as.factor(length):as.factor(putter):as.factor(Break)	1	0.08	0.079	0.033	0.8559	
as.factor(length):as.factor(putter):as.factor(slope)	1	4.55	4.554	1.914	0.1698	
as.factor(length):as.factor(Break):as.factor(slope)	1	2.02	2.023	0.850	0.3589	
as.factor(putter):as.factor(Break):as.factor(slope)	1	1.05	1.052	0.442	0.5078	
as.factor(length):as.factor(putter):as.factor(Break):as.factor(slope)	1	4.73	4.734	1.989	0.1616	
Residuals	96	228.45	2.380			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

FDR Control Test

Order	Effect	Test stat	P-value	FDR value	Significant (Y/N)
1	as.factor(length)	9.082	0.0033	0.0066	Y
2	as.factor(putter)	6.574	0.0119	0.0133	Y
3	as.factor(length):as.factor(putter)	2.497	0.1173	0.02	N
4	as.factor(length):as.factor(putter):as.factor(Break):as.factor(slope)	1.989	0.1616	0.02667	N
5	as.factor(length):as.factor(putter):as.factor(slope)	1.914	0.1698	0.04	N
6	as.factor(length):as.factor(slope)	1.810	0.1816	0.04	N
7	as.factor(Break)	1.657	0.2011	0.04667	N
8	as.factor(putter):as.factor(Break)	0.862	0.3556	0.0533	N
9	as.factor(length):as.factor(Break):as.factor(slope)	0.850	0.3589	0.06	N
10	as.factor(putter):as.factor(Break):as.factor(slope)	0.442	0.5078	0.0667	N
11	as.factor(length):as.factor(Break)	0.309	0.5796	0.0733	N
12	as.factor(putter):as.factor(slope)	0.260	0.6115	0.08	N
13	as.factor(slope)	0.054	0.8175	0.0867	N
14	as.factor(length):as.factor(putter):as.factor(Break)	0.033	0.8559	0.0933	N
15	as.factor(Break):as.factor(slope)	0.007	0.9315	0.1	N

Step 2

- Our FDR control test shows that even after controlling for type 1 error, the only two significant factors are length and putter. We wanted to see if the significance of effects would increase if unimportant interactions were removed one at a time. For this reason, we decided to use a step-by-step process using the heredity principle and hierarchy principle to remove unimportant factors/interactions.
- In the transformed ANOVA we found that only length and putter were significant as they had p-values less than alpha (0.05). We decided remove the 4 way interaction as the p-value was way above 0.05, and is not likely to be important according to the principle of hierarchy.

Model 2

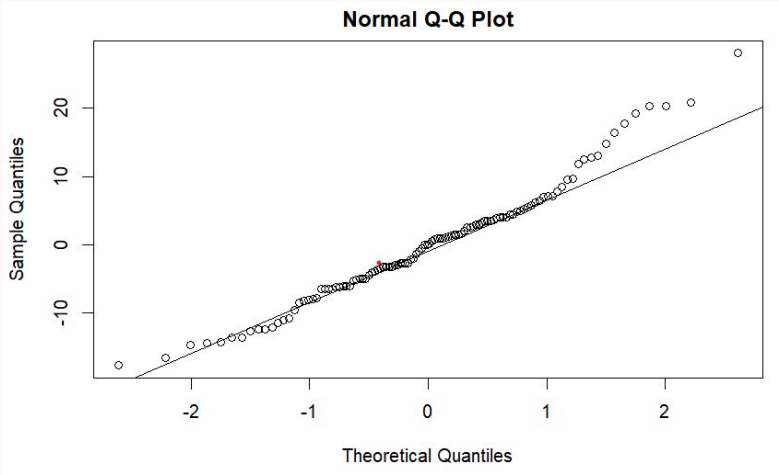
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	21.61	21.613	8.991	0.00345	**
as.factor(putter)	1	15.64	15.644	6.508	0.01230	*
as.factor(Break)	1	3.94	3.944	1.641	0.20329	
as.factor(slope)	1	0.13	0.127	0.053	0.81840	
as.factor(length):as.factor(putter)	1	5.94	5.943	2.472	0.11913	
as.factor(length):as.factor(Break)	1	0.74	0.735	0.306	0.58154	
as.factor(putter):as.factor(Break)	1	2.05	2.051	0.853	0.35796	
as.factor(length):as.factor(slope)	1	4.31	4.308	1.792	0.18380	
as.factor(putter):as.factor(slope)	1	0.62	0.618	0.257	0.61326	
as.factor(Break):as.factor(slope)	1	0.02	0.018	0.007	0.93182	
as.factor(length):as.factor(putter):as.factor(Break)	1	0.08	0.079	0.033	0.85660	
as.factor(length):as.factor(putter):as.factor(slope)	1	4.55	4.554	1.894	0.17188	
as.factor(length):as.factor(Break):as.factor(slope)	1	2.02	2.023	0.841	0.36127	
as.factor(putter):as.factor(Break):as.factor(slope)	1	1.05	1.052	0.437	0.50993	
Residuals	97	233.18	2.404			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + \epsilon_{ijkl}$$

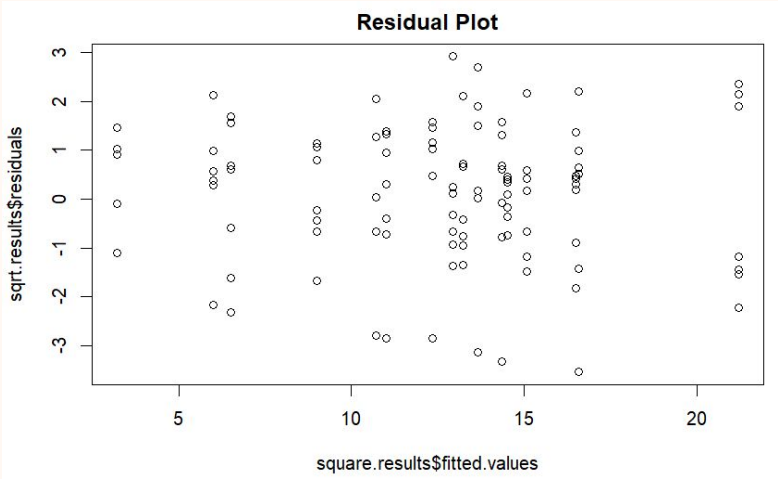
Assumptions for Model 2

QQ-plot



For the normality assumption we ran a qq plot and concluded that the assumption was met as the residuals followed a roughly linear line

Residual plot



We also checked the constant variance assumption and the random scatter of points showed the assumption is met

FDR Control Test

Order	Effect	Test stat	P-value	FDR value	Significant (Y/N)
1	as.factor(length)	8.991	0.00345	0.0071	Y
2	as.factor(putter)	6.508	0.01230	0.0143	Y
3	as.factor(length):as.factor(putter)	2.472	0.11913	0.02143	N
4	as.factor(length):as.factor(putter):as.factor(slope)	1.894	0.17188	0.02857	N
5	as.factor(length):as.factor(slope)	1.792	0.35796	0.03571	N
6	as.factor(Break)	1.641	0.20329	0.04286	N
7	as.factor(putter):as.factor(Break)	0.853	0.35796	0.05	N
8	as.factor(length):as.factor(Break):as.factor(slope)	0.841	0.36127	0.0571	N
9	as.factor(putter):as.factor(Break):as.factor(slope)	0.437	0.50993	0.0642	N
10	as.factor(length):as.factor(Break)	0.306	0.58154	0.0714	N
11	as.factor(putter):as.factor(slope)	0.260	0.61326	0.0786	N
12	as.factor(slope)	0.053	0.81840	0.08571	N
13	as.factor(length):as.factor(putter):as.factor(Break)	0.033	0.85660	0.09286	N
14	as.factor(Break):as.factor(slope)	0.007	0.93182	0.1	N

We noticed that the p-values for all effects stayed the same or increased slightly after removing the 4-way interaction. We believe this is because the 4 way interaction was actually more significant than some of the other lower order 2 way and 3 way interactions. However, length and putter are still significant after controlling for type 1 error. To see if any p-values would decrease and become more significant, we wanted to follow the step-by-step process and remove insignificant 3 way interactions next.

Step 3

Based on the previous ANOVA and FDR control test, we see that the 3 way interactions are still insignificant. These p-values were much greater than 0.05, so we decided to remove them from the model. The principle of hierarchy also says that these interactions are less likely to be important than 2 way interactions or main effects, so we wanted to remove the 3 way interactions and see if it increased significance of the 2 way interactions.

Model 3

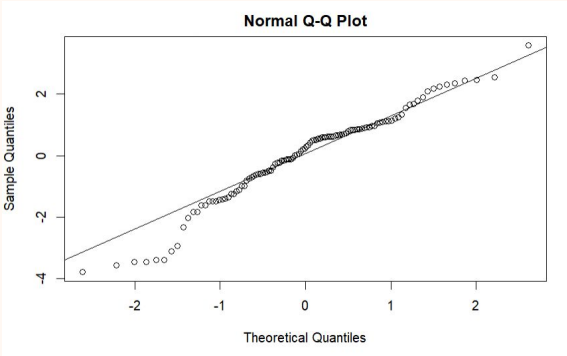
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	21.61	21.613	9.062	0.0033	**
as.factor(putter)	1	15.64	15.644	6.559	0.0119	*
as.factor(Break)	1	3.94	3.944	1.654	0.2014	
as.factor(slope)	1	0.13	0.127	0.053	0.8177	
as.factor(length):as.factor(putter)	1	5.94	5.943	2.492	0.1176	
as.factor(length):as.factor(Break)	1	0.74	0.735	0.308	0.5800	
as.factor(putter):as.factor(Break)	1	2.05	2.051	0.860	0.3560	
as.factor(length):as.factor(slope)	1	4.31	4.308	1.806	0.1820	
as.factor(putter):as.factor(slope)	1	0.62	0.618	0.259	0.6118	
as.factor(Break):as.factor(slope)	1	0.02	0.018	0.007	0.9315	
Residuals	101	240.89	2.385			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + \epsilon_{ijkl}$$

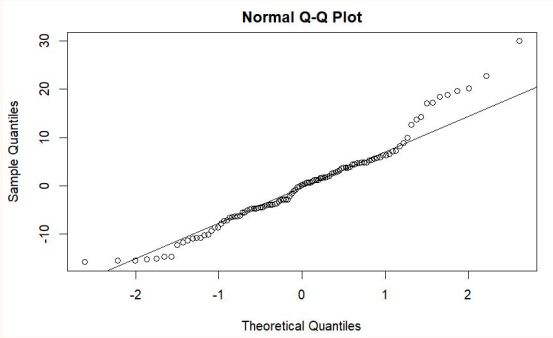
Assumptions for Model 3

First QQ-plot



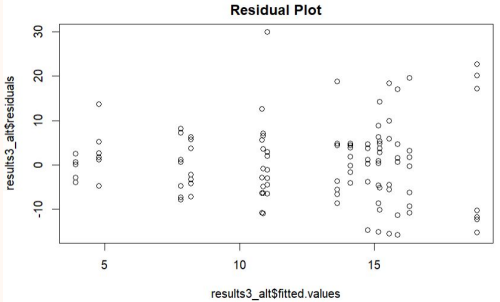
Since the qq plot seems to be curved downward from the normal distribution line, we decided to transform by a higher power. We chose to transform by squaring the values.

Second QQ-plot



With the new transformation, the residuals are following the normal distribution line more closely, so we decided the normality assumption was met.

Residual plot



We then checked our variance assumption and the assumptions seem to be met because there is no evident fanning out pattern within the plot.

New Transformed Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	917	917.1	10.765	0.00142	**
as.factor(putter)	1	388	388.1	4.556	0.03523	*
as.factor(Break)	1	145	145.1	1.704	0.19478	
as.factor(slope)	1	1	1.4	0.016	0.89843	
as.factor(length):as.factor(putter)	1	219	218.7	2.567	0.11225	
as.factor(length):as.factor(Break)	1	12	11.9	0.140	0.70944	
as.factor(putter):as.factor(Break)	1	115	115.0	1.350	0.24801	
as.factor(length):as.factor(slope)	1	94	93.8	1.101	0.29654	
as.factor(putter):as.factor(slope)	1	56	56.4	0.662	0.41764	
as.factor(Break):as.factor(slope)	1	2	1.6	0.019	0.89036	
Residuals	101	8605	85.2			

This is our transformed ANOVA table with 4 way and 3 way interactions removed. After removing 3 way interactions, we can see that a few of the p-values decreased, and a few p-values increased:

FDR Control Test

Order	Effect	Test stat	P-value	FDR value	Significant (Y/N)
1	as.factor(length)	10.765	0.00142	0.0071	Y
2	as.factor(putter)	4.556	0.03523	0.0143	N
3	as.factor(length):as.factor(putter)	0.016	0.11225	0.02143	N
4	as.factor(Break)	1.704	0.19478	0.04286	N
5	as.factor(putter):as.factor(Break)	1.350	0.24801	0.05	N
6	as.factor(length):as.factor(slope)	1.101	0.29654	0.03571	N
7	as.factor(putter):as.factor(slope)	0.662	0.41764	0.0786	N
8	as.factor(length):as.factor(Break)	0.140	0.70944	0.0714	N
9	as.factor(Break):as.factor(slope)	0.007	0.89036	0.1	N
10	as.factor(slope)	0.053	0.89843	0.08571	N

We decided to run an FDR control test again to make sure that we were controlling for type 1 error rate after running 10 hypothesis tests. We used a false discovery rate of 0.1.

Step 4

- Since two way interactions are shown to be insignificant, we wanted to continue with the step-by-step process of removal to see if the p-values of main effects would decrease and become more significant. From this FDR control, we also see that putter is now insignificant after controlling for type 1 error. However, we don't want to remove it just yet, as it has been significant up until now and we believe it warrants further testing.

- Based on the previously reduced ANOVA and FDR control test, we see that 2 way interactions are all insignificant and much greater than 0.05, so we decided to remove them from the model. All we have left now is the 4 original factors.

Model 4

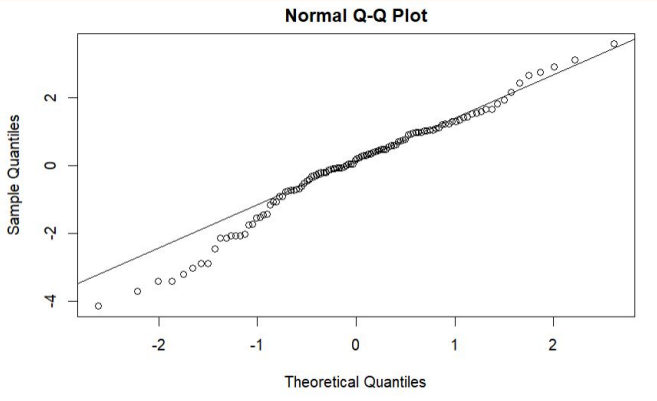
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	21.61	21.613	9.085	0.00322	**
as.factor(putter)	1	15.64	15.644	6.576	0.01173	*
as.factor(Break)	1	3.94	3.944	1.658	0.20068	
as.factor(slope)	1	0.13	0.127	0.054	0.81743	
Residuals	107	254.56	2.379			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl}$$

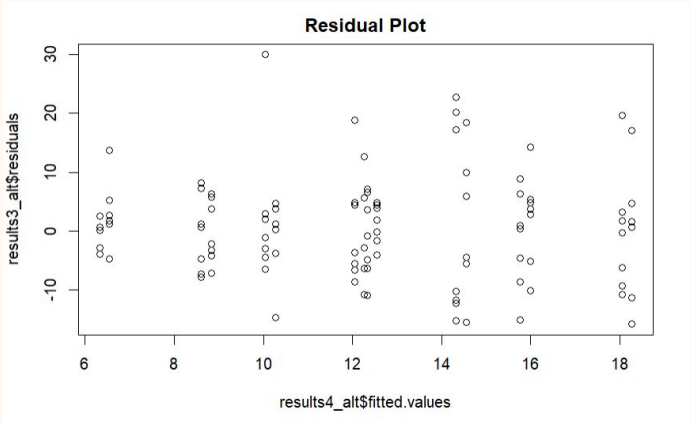
Assumptions for Model 4

QQ-plot



We then checked the normality assumption. The points are following relatively closely along the normal distribution line. There is some deviation in the bottom left, but since the rest of the plot is within the normal distribution, we decided the normality assumption is met

Residual Plot



We checked the constant variance assumption. There is no evident fanning out pattern so this assumption is satisfied.

FDR Control Test

Order	Effect	Test stat	P-value	FDR value	Significant (Y/N)
1	as.factor(length)	9.085	0.00322	0.025	Y
2	as.factor(putter)	6.576	0.01173	0.05	Y
3	as.factor(Break)	1.658	0.20068	0.075	N
4	as.factor(slope)	0.054	0.81743	0.1	N

By looking at the p-values in the reduced ANOVA and FDR control test, can see that break and slope are still insignificant as their p-values are greater than 0.05. At this point, we found that none of the interactions are important to the model, so we could freely remove the break and slope factors according to the principle of heredity.

Final Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(length)	1	21.61	21.613	9.109	0.00317	**
as.factor(putter)	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

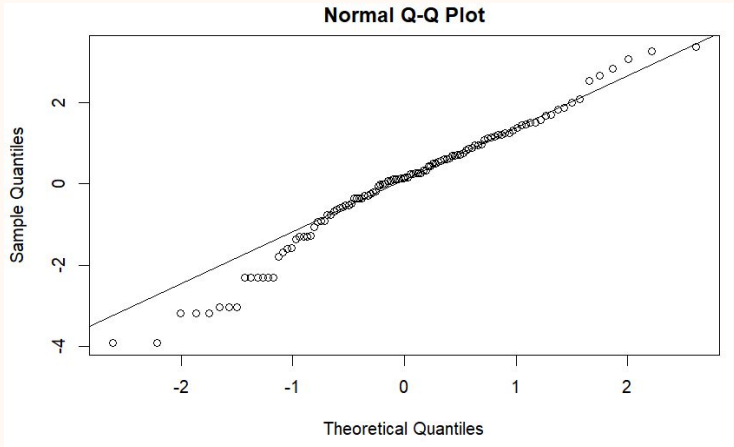
$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \epsilon_{ijkl}$$

Length p-value interpretation: The corresponding p-value is 0.00317. If the mean of the response variable is the same across all treatment levels in this factor, then there is 0.317% chance of treatment level differences in the sample mean response as large as observed in our experiment.

Putter p-value interpretation: The corresponding p-value is 0.01159. If the mean of the response variable is the same across all treatment levels in this factor, then there is 1.159% chance of treatment level differences in the sample mean response as large as observed in our experiment.

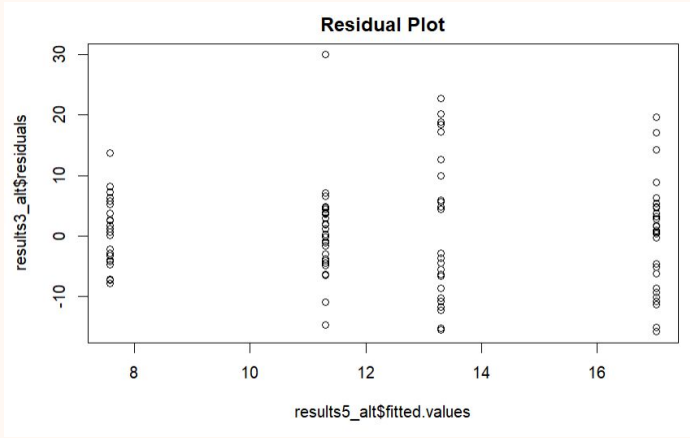
Assumptions for Final Model

QQ-plot



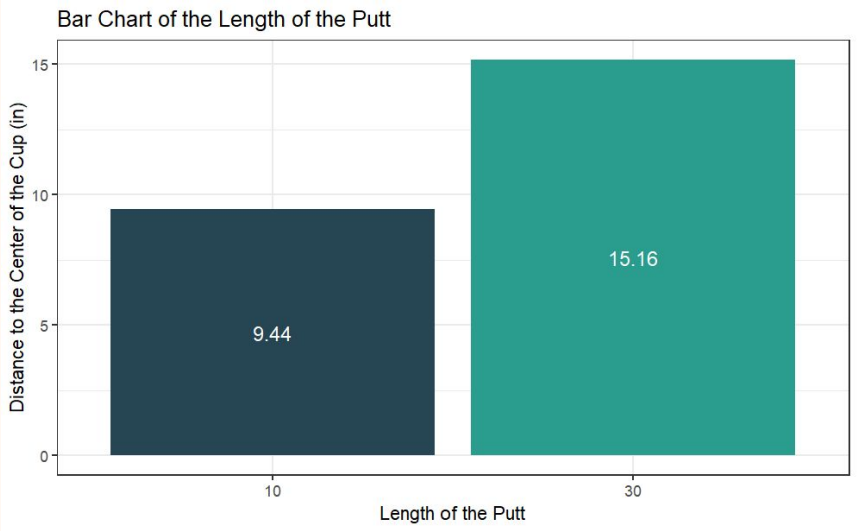
We checked our normality assumption and decided the assumption was satisfied because the residuals were following relatively closely to the normal distribution line.

Residual Plot

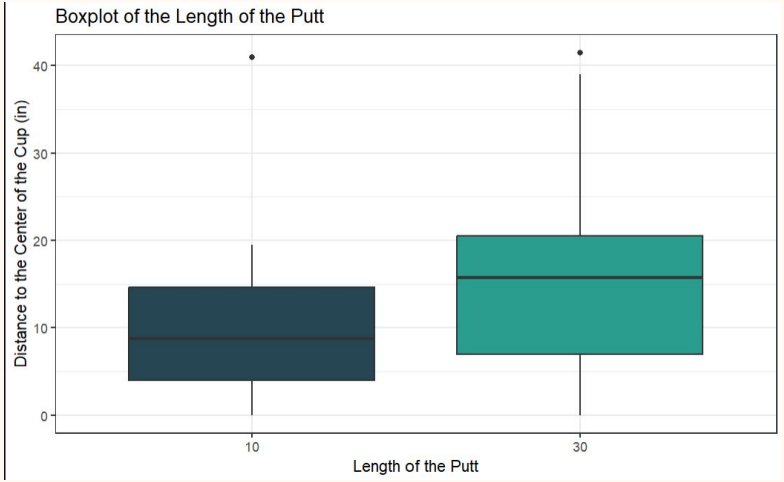


We then checked our constant variance assumption. There is no evident fanning out pattern and residuals are falling within horizontal bands, so this assumption is satisfied

Exploratory Analysis - Length

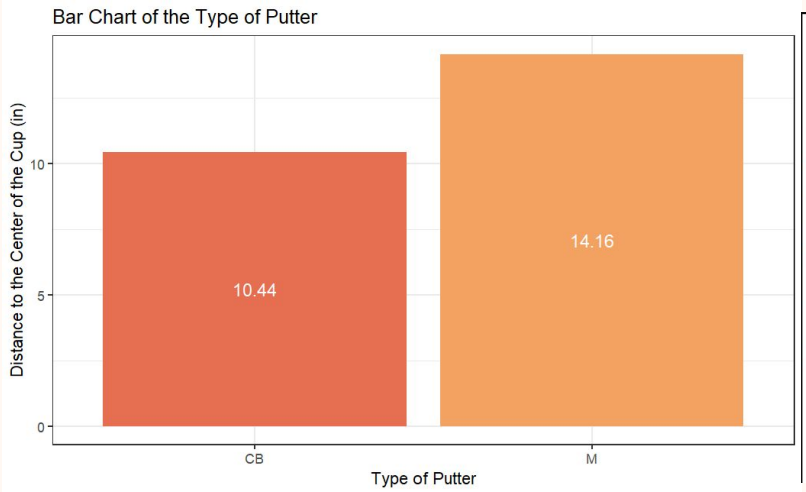


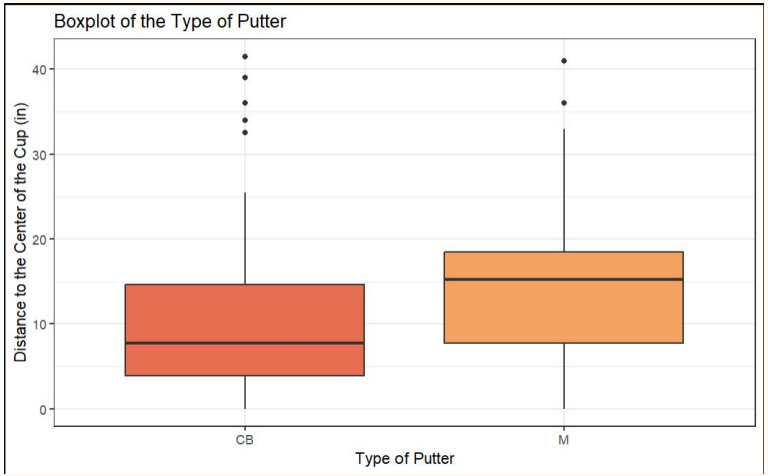
This is the bar chart for the length factor. According to this plot, putt lengths of 10 feet generally have a better-putting accuracy, with the average distance to the cup being 9.44 inches. The putt length of 30 has worse putting accuracy, with the distance to the center of the cup being 15.16 inches



According to the boxplot, the putt length of 10 only has one outlier, with the distance to the center of the cup being greater than 40. The putt length of 10 also has a much smaller deviation. The putt length of 30 also only has one outlier, with the distance to the center of the cup being greater than 40. However, this putt length has much more deviation than the length of 10.

Exploratory Analysis - Putter





According to bar chart, putter types CB feet have an average putting accuracy of 10.44 inches. Putter type M has an average putting accuracy of 14.16 inches

According to the boxplot, putter type CB has less deviation, but more outliers compared to putter type M. Putter type CB has 4 outliers after a putting accuracy of around 30 inches. Putter type M has 2 outliers after a putting accuracy of around 35 inches.

Limitations and Future Work

The Design

- A 2^k factorial design can only test two levels per factor, which limits the range of values that can be tested for each factor.
- The design also assumes that the effects of each factor are independent of each other. However, the effects of one factor may have depended on the level of another factor but there is no way to test for dependencies in a 2^k factorial design.
- We can better improve our model by conducting more replications as our model only had 7. Increasing the number of replications will yield more accurate results.

The Analysis

- Explore different post-hoc tests
  - We decided to use the FDR Control test as a way of limiting type 1 error because it isn't as conservative as Bonferroni's and is useful for large scale analysis
  - However, we could have used Bonferroni or Tukey's for our interaction terms
  - We could have also done pairwise testing within each factor for our final model
- Instead of removing the interactions step by step, we could remove them all at once
  - This would simplify our process

Future Work

We can increase the number of levels in the length and putter factors. Length can have 10, 15, 20, 25, and 30 as levels. Putter type can have CB, M, and other types as levels.

We can increase the number of factors to study. There are possibly other factors that interact with the current factors more, so that could be explored. Examples of other factors include golf course ground material, golf ball type, and wind level.

We can conduct a full factorial experiment on the two factors we found significant (length and putter) to see which levels have the most effect on putting accuracy.