Stat 361 Final Project Report

Amy Bertelsman, Pearl Muensterman, & Ethan Morlock

Intro

Our project models wait times for amusement park attractions. This information is essential for park managers. Wait times demonstrate how popular a ride is compared to others. If there is one ride in particular with a long wait, a manager could decide there should be more rides built that are similar to cut down the wait time. If some attractions have little to no wait time, is it worth continuing operation if patrons are not enjoying it? Do some attractions have longer waits than others based on the type of ride? Park visitors are also interested in wait times, as they would assume them to be accurate and would be upset if they were underestimated. Inaccurate wait times could lead to unhappy customers which park managers wish to avoid.

Initially looking at this data, we had some hypotheses. Does the length of an attraction cause longer wait times? If there is a higher interest rating among most age groups for an attraction, is there a longer wait time? Do main attractions grab people's attention more than other attractions and therefore have a longer wait? Similar to main attractions, are new shiny ones more appealing causing a wait longer than older attractions? Do roller coasters have longer wait times than other attractions? With fewer building restrictions outdoors, do the larger outdoor attractions entice more customers to wait? We investigate these theories throughout our project.

Data

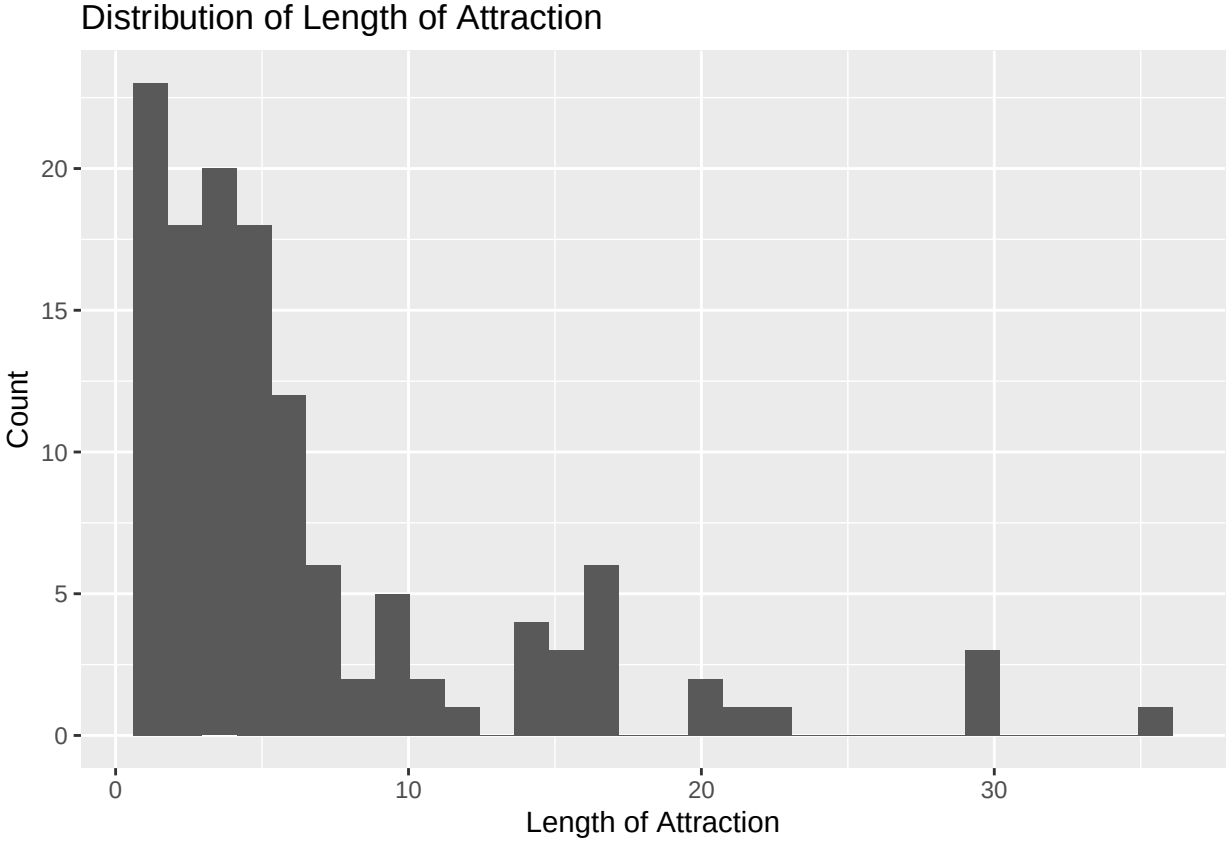
The data used in this project was obtained from a published senior thesis done by Danica Dvorachek, a student at Liberty University, in 2018. She collected the data the same year using the official park websites, park guides, and participant surveys. The data consists of 128 observations, each corresponding to an attraction from three major U.S. theme parks: Disney World, Disneyland, and Universal Studios Orlando. Seventeen variables were recorded for each observation. Theses 17 variables are name of the attraction, whether the attraction is a main ticket attraction, the attraction's age in years, the length of the attraction in minutes, the attraction's height requirement (if any) in inches, the interest level of the six different age groups (preschool aged children, grade school aged children, teenagers, young adults, adults over 30, and seniors), the type of vehicle used in the attraction, the average wait time in minutes per 100 people ahead of an individual, whether the attraction is based on a popular story, whether an attraction is indoor, outdoor, or both, the average thrill level of the attraction, and the average sense level of the attraction. The interest level of the age group was a scaled rating from one to five with one indicating little interest and a five indicating high interest. The age groups were defined as follows: preschool ages 0-5, grade school ages 5-12, teenager ages 12-18, young adult ages 18-30, adult ages 30-55, and senior ages 55 and over. The type of vehicle was used to indicate the type of attraction and had 15 levels: boat, drop tower, fast guided track, merry-go-round, midway, omnimover, raft, roller coaster, show (stationary seats), spinner, train, truck, variable track, and water flume. The average thrill and sense level were calculated by averaging the responses of a participant survey conducted by the author. Both are scaled ratings. Thrill was on a scale from zero to three and dealt with the intensity of attraction's track (whether it had big drops or spins and how many), and sensory was a scale of zero to four indicating how many of the human senses were stimulated to enhance the attraction experience (taste was not utilized in any attraction thus the scale ends at four).

To analyze the data any categorical variables were changed to indicator variables. Whether or not the

attraction was a main ticket attraction and whether or not the attraction was based on a popular story were simply changed from "Yes" or "No" to 1 and 0. For attraction environment (indoor, outdoor, or both), two indicator variables were created: indoor and outdoor. A one in the outdoor column indicated that it was an outdoor attraction while a one in the indoor column indicated an indoor attraction. A zero in both columns indicated that the attraction had elements that were both indoor and outdoor. The final categorical variable, vehicle type, was not used in the analysis as to turn it into an indicator variable would have required the creation of 14 dummy variables which could create multicollinearity issues and too many variables for too few observations. Thus, it was dropped from the data.

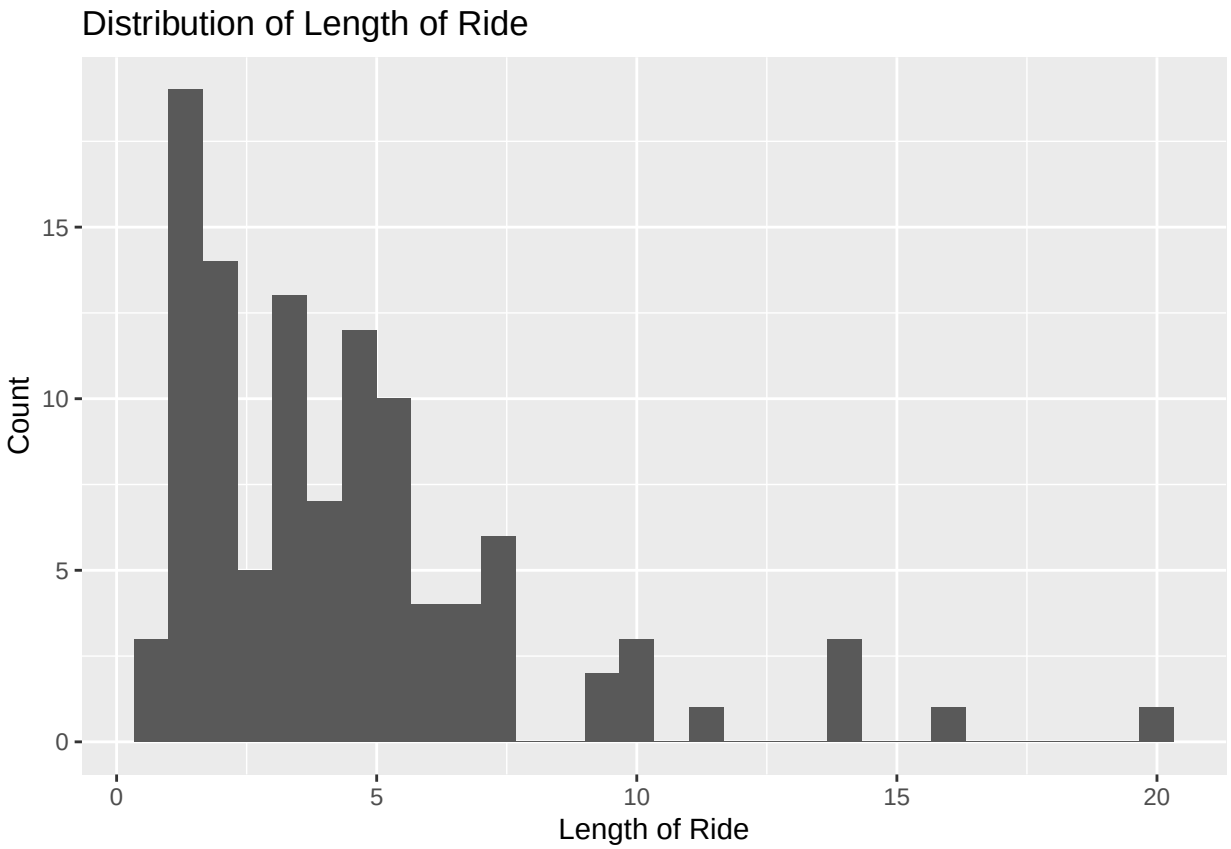
Models and Analysis

For the first attempt at finding our best model, we compared the models that resulted from the variable selection methods: forward selection, backward elimination, stepwise regression, and all-possible regression. The metric of comparison was the adjusted R-squared value. The backward elimination and stepwise regression gave us the same models. Also, all four models had common variables. They were the interest of young adults, whether the attraction was based on a popular story, the length of attraction, and outdoor attractions. We found our best model by using the all-possible regression which had length of attraction, interest level of teens, interest level of young adults, interest level of seniors, average thrill level, whether the attraction is based on a popular story, and outdoor attractions as the variables. This model gave us an adjusted R-squared of 0.4633, which was not great. We ran an influence measures test and saw that 7 out of 10 of the influential points were shows rather than other attractions. When investigated, we saw that typically shows had longer lengths than all other attractions. Looking at a distribution of attraction length it was heavily right skewed



This raised questions of the impact of shows on the model, and whether or not shows were fundamentally different from other attractions, from here on referred to as rides.

Since we saw the difference in rides and shows from the first attempt we decided to remove all the data points that were shows for our second attempt at finding our best model. When looking at the distribution of ride length after removing shows, it is still right skewed but not as extreme. This skew comes from rides which are not clearly a ride (such a roller coaster or drop tower) or show and thus have longer wait times, such as Kilimanjaro Safari or simulator rides like Soarin'.

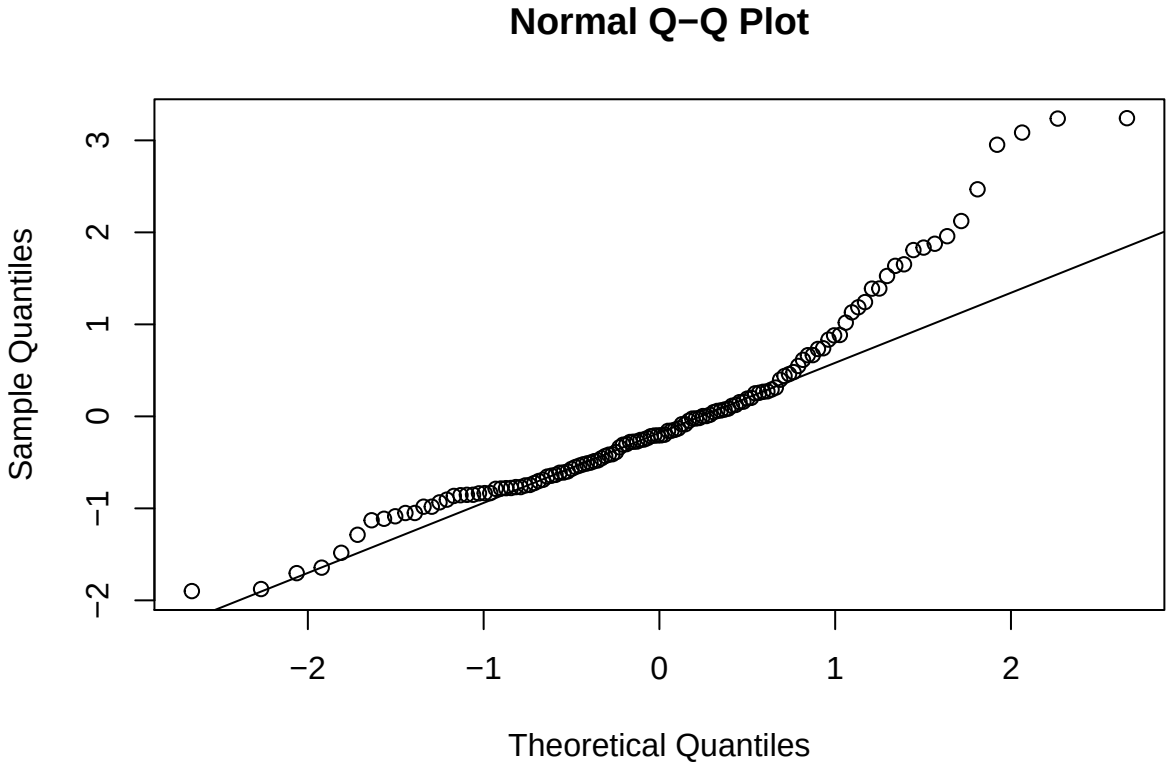


The forward selection, backward elimination, and stepwise regression all gave us the same model. However, the forward selection had a different order of the variables. The all-possible regression gave us our best model again with an adjusted R-squared of 0.4923 this time. It was better than our first model. This adjusted R-squared was only slightly better than the model the other three selection methods gave us which was 0.4878. The regressors of this model were whether it was a main ticket ride, age of the ride, length of the ride, interest level of teens, interest level of young adults, interest level of seniors, whether the ride was based on a popular story, average sense level, indoor rides, and outdoor rides. When we looked at this model we also noticed the variables interest level of teens, interest level of young adults, and interest level of adults were all very similar for all attractions. This raised the questions of having some possible collinearity occurring in our model.

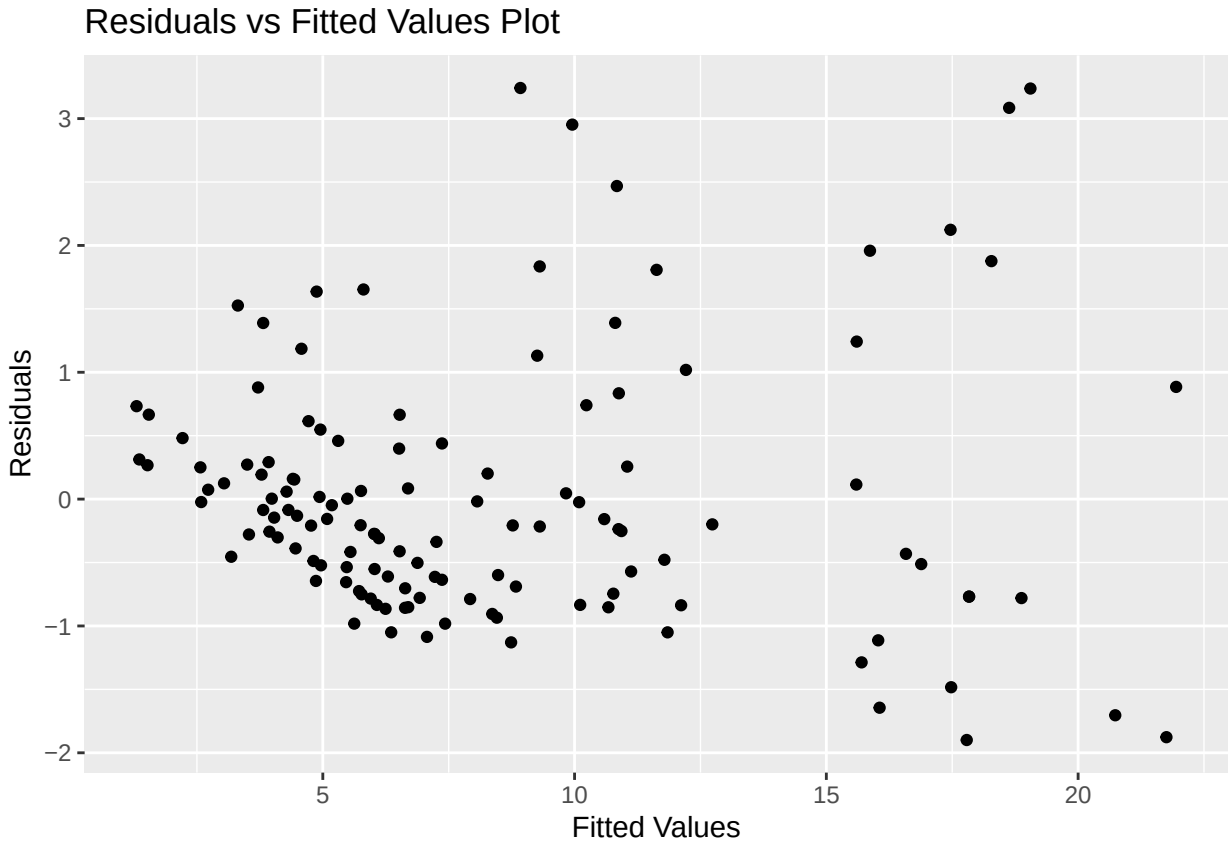
This led us to our third attempt at finding our best model. We computed the average of the three variables (interest level of teens, interest level of young adults, interest level of adults) and made a variable of the average interest level of adults. We did this to try and get rid of the collinearity we thought was occurring in the model. We also kept using the data without the shows because our model did improve when we got rid of the shows. We had all the same dummy variables from the first two attempts as well as not having the vehicle variable, just like the first two attempts. We did the same process and saw that the forward selection, backward elimination, and stepwise regression gave us the same models again with forward selection having a different order of the variables. Once again, the all-possible regression gave us our best model with the regressors whether the ride was a main ticket ride, age of the ride, length of the ride, whether the ride was based on a popular story, average sense level, indoor rides, outdoor rides, and average interest level of adults. This gave us an adjusted R-squared of 0.4970 which was better than our second attempt and the

other selection methods for our third attempt, but not by very much. This adjusted R-squared value was still not good.

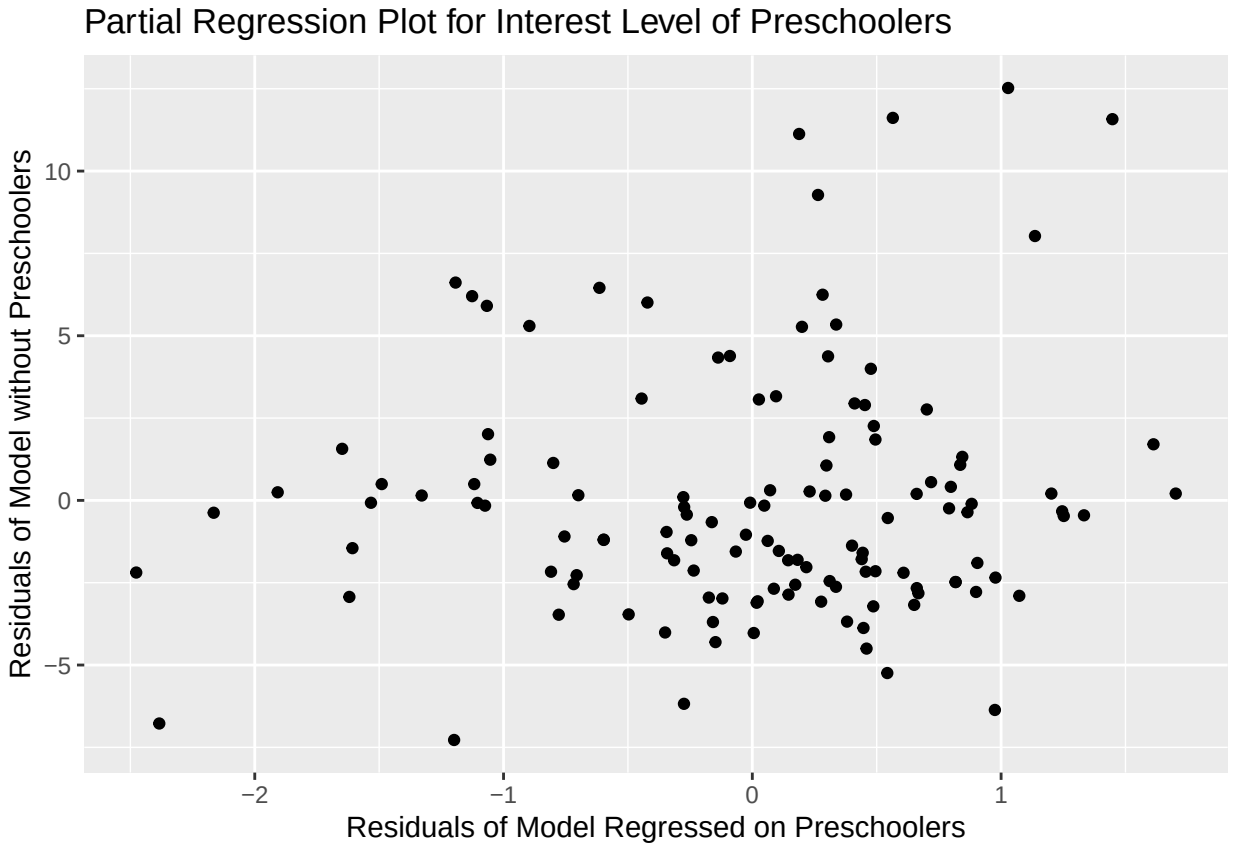
After our third attempt at finding our best model we decided to put shows back into our data and make a dummy variable out of the vehicle variable. Instead of having 15 different descriptions for vehicles, we made it into the dummy variable ride which said whether an attraction was a ride (1) or a show (0). After we made this dummy variable we then dropped the vehicle variable again. We performed the same selection methods as the first three attempts and again the forward selection, backward elimination, and stepwise regression gave us the same model with forward having the variables in a different order. The all-possible regression method gave us our best model of all the attempts with an adjusted R-squared of 0.6141. This was a good improvement over our other three attempts. This model had the regressors length of attraction, interest level of preschoolers, whether the attraction was based on a popular story, average thrill level, outdoor attraction, average interest level of adults, and whether the attraction was a ride. Since this was our best model, we performed residual analysis upon this model.



The normality plot showed the data to deviate from the upper tail and have a right skew. The residuals vs fitted values plot showed a slight negative relation but not a strong pattern. Thus, we determined a transformation would not improve the results greatly. Instead, we believed the real issue was we did not have the appropriate data to accurately predict wait times.



We also did the partial regression plots for the quantitative variables. The plot below shows the partial regression plot for interest level of preschoolers.



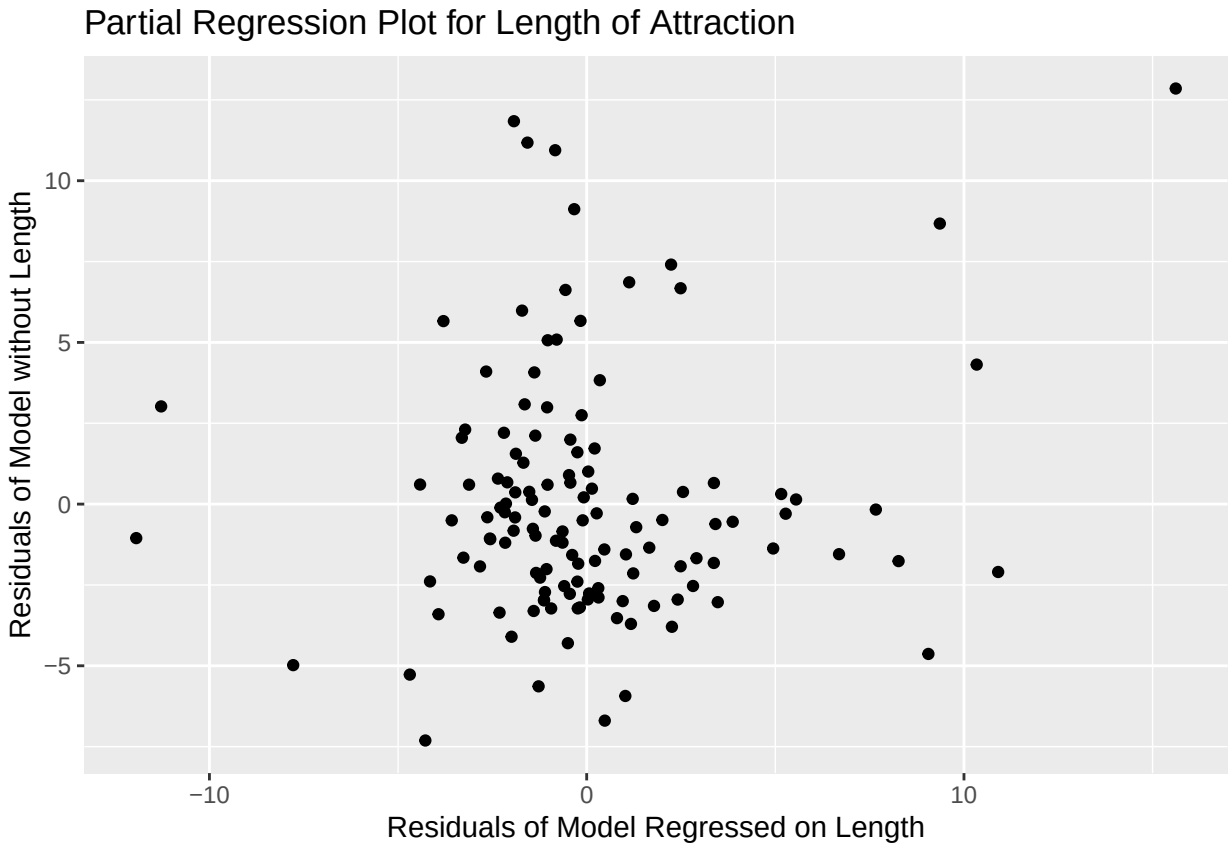
The plot shows a very slight positive relationship but generally has no pattern. This indicates that there is not a strong linear relationship between the regressor and the response and implies that the interest level of preschoolers is probably not a good regressor. The rest of the partial regression plots showed the same thing (these plots can be viewed in the appendix). Thus, none of the regressors indicated a strong linear relationship with the response variable.

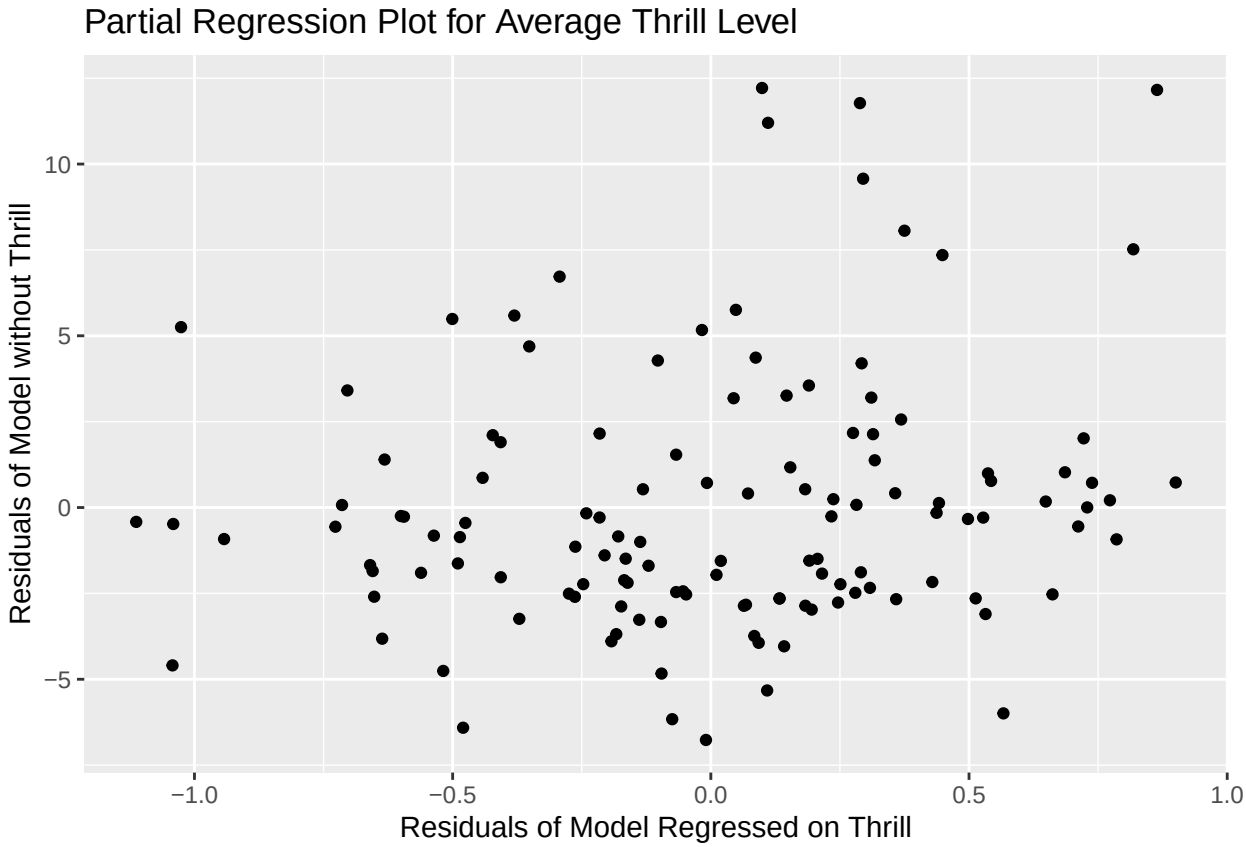
Conclusion

We had less than stellar results however we believe this is because we simply were not given the appropriate variables to model wait time. There were many variables we desired to analyze but were not in the dataset. A variable given the location within the park relative to the entrance or other main ticket attractions were thought to be able to give us an inside scoop on clusters of rides. Often customers have a ride in mind and visit neighboring attractions after the first in which this could determine a relationship with wait times. Given the number of people an attraction could service in an hour could be useful as this could give a better representation of wait times. This leads to more variables like time of day, time of the year (season), and the number of people in the park in a day. Rides could even have data for each day of the year. The amount of time and money spent on advertising specifically for a new attraction would be beneficial as it would indicate the knowledge of a certain ride to patrons and therefore impact the wait time. As a final statement, if ticket sales were applicable to each ride, this could also help determine wait time.

Our best model only explained about 61% of the variability in the wait times. This is not a strong relationship and thought to not predict well. We are inclined to think that the additional variables would aid tremendously in increasing the explained variation in wait time. In our residual analysis, the partial regression plots did not show any obvious relationship between any of the regressors and response. Therefore, we have evidence of a non linear relationship and think a different modeling technique may fit the data better.

Appendix


Partial Regression Plot for Length of Attraction
Residuals of Model without Length
10
5
0
−5
−10
0
10
Residuals of Model Regressed on Length


Partial Regression Plot for Average Thrill Level
Residuals of Model without Thrill
10
5
0
−5
−1.0
−0.5
0.0
0.5
1.0
Residuals of Model Regressed on Thrill


Partial Regression Plot for Average Interest Level of Adults
Residuals of Model without Average Adults
10
5
0
−5
−1.0
−0.5
0.0
0.5
Residuals of Model Regressed on Average Adults