Mengjie Shen

Jeffrey Simonoff

Regression and Multivariate Data Analysis

March 29, 2021

What influences the Total Tourists' Attendance of Amusement Parks?

Everyone loves amusement parks. There are many famous theme park brands, such as Walt Disney Attractions and Merlin Entertainment Group. These companies built lots of amusement parks all around the world in order to make a profit. However, the total attendance for different amusement parks varies. Even for the parks that are operated by the same companies, there is a significant difference between the total attendance of the whole year. For example, for Disneyland, the attendance of 2019 varies from 20.96 million (by Magic Kingdom (Walt Disney World), USA) to 5.7 million people (by Hong Kong Disney Land). For the operators of the companies, total attendance of the year can directly affect the company's profit. Therefore, it would be very interesting and meaningful to know what factors can influence the total attendance of the whole year. It will improve the decision-making of those operators of entertainment companies.

I gathered the total attendance data of amusement parks from the website Statista. Since the impact that the covid-19 brought to tourism varies from countries, I select the data of 2019 for simplification. I select two different kinds of amusement park data: the world's top theme park data (https://www.statista.com/statistics/194247/worldwide-attendance-at-theme-and-amusement-parks/) and the world top water park data(https://www.statista.com/statistics/194343/attendance-figures-of-waterparks-worldwide/). I use an indicator variable to indicate the type of park with water parks represent as 0 and theme

parks represent as 1. The variables given are the park size and the adult ticket price of the amusement parks. They are available on the website The Park Database (https://www.theparkdb.com/). I also get the ratings for each individual park on TripAdvisor(https://www.tripadvisor.com/). The graph down below is the head of the data. We have 20 theme park observations and 16 water park observations in total.

	C1-T	C2	C3	C4	C5	C6
		total attendance(in millions)	Waterpark(0)/Themepark(1)	Size(square miles)	Ticket Price($)	Rate on TripAdvisor
1	Magic Kingdom (Walt Disney World), USA	20.96	1	380000	122	4.5
2	Disneyland Anaheim, USA	18.66	1	350000	149	4.5
3	Tokyo Disneyland, Japan	17.91	1	480000	68	4.5
4	Tokyo Disney Sea, Japan	14.65	1	390000	68	4.5
5	Universal Studios, Japan	14.50	1	360000	71	4.0

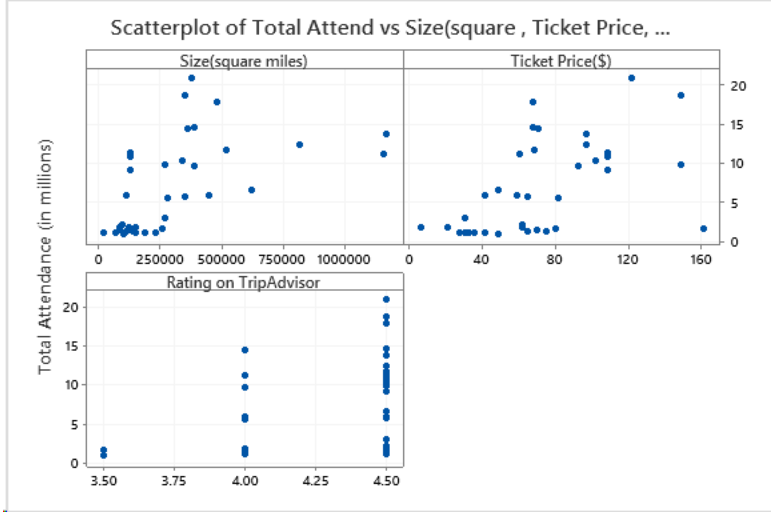
First, let's take a look at the data. Here are the summary statistics:

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Total Attendance (in millions)	36	0	7.040	0.984	5.906	1.120	1.600	5.850	11.412
Size(square miles)	36	0	306194	45650	273901	20000	120000	245000	387500
Ticket Price($)	36	0	72.86	6.14	36.86	7.00	43.75	68.00	97.00
Rating on TripAdvisor	36	0	4.3056	0.0499	0.2995	3.5000	4.0000	4.5000	4.5000

Variable	Maximum
Total Attendance (in millions)	20.960
Size(square miles)	1170000
Ticket Price($)	161.00
Rating on TripAdvisor	4.5000

We can see that the average total attendance is 984 thousand people per year. The average size of the theme park is 45650 square miles, the average ticket price is 72.86$ and the average rating on TripAdvisor is 4.3 out of 5. The highest total attendance of 20.96 million people is achieved by Magic Kingdom (Walt Disney World) in the USA. It was operated by the Walt Disney Company.



Although the relationship is not very distinguishable, we seem to identify the pattern that larger size leads to higher total attendance. However, it's surprising that the higher ticket price leads to higher total attendance. This goes against our common sense since we usually assume people would prefer places with a lower ticket price. There is an obvious point that is on the lower right corner. This is Bahamas Aquaventure Water Park with a relatively high ticket-price and low total attendance. Since the ratings on Trip Advisor are only accurate to 0.5. It's hard to identify the direct relationship between the rating and the total attendance. But we can still see that the higher the rating is, the maximum total attendance the park is higher.

Here's the regression of total attendance on the three variables:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	636.62	212.206	11.63	0.000
Size(square miles)	1	226.61	226.609	12.41	0.001
Ticket Price($)	1	210.35	210.353	11.52	0.002
Rating on TripAdvisor	1	19.95	19.953	1.09	0.304
Error	32	584.09	18.253		
Lack-of-Fit	30	581.13	19.371	13.10	0.073
Pure Error	2	2.96	1.479		
Total	35	1220.71			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.27234	52.15%	47.67%	37.93%

Coefficients

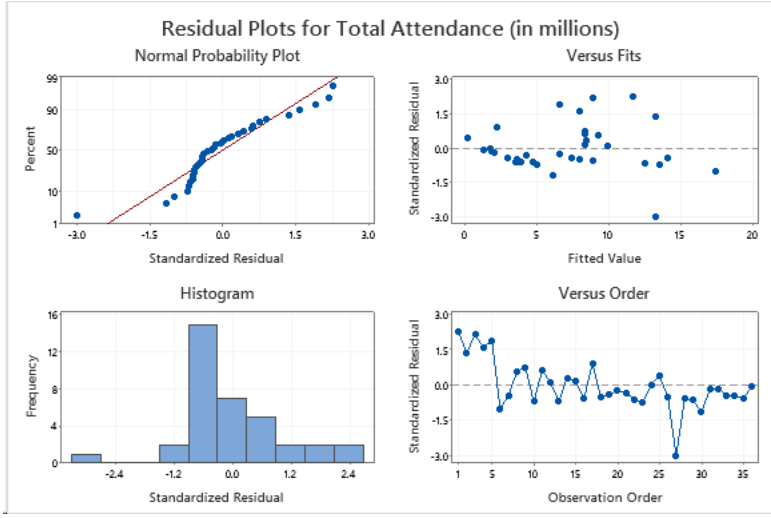
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-12.2	10.5	-1.17	0.253	
Size(square miles)	0.000010	0.000003	3.52	0.001	1.05
Ticket Price($)	0.0696	0.0205	3.39	0.002	1.09
Rating on TripAdvisor	2.62	2.51	1.05	0.304	1.08

Regression Equation

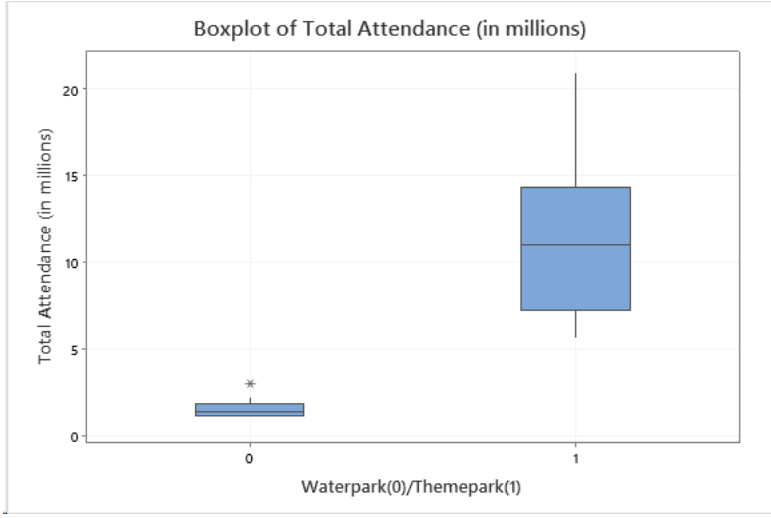
Total Attendance (in millions) = -12.2 + 0.000010 Size(square miles) + 0.0696 Ticket Price($)
+ 2.62 Rating on TripAdvisor

R^2 is 52.15%, which means that this model accounts for 52.15% of the variability in the target variable. It's moderately strong. Since we got 36 observations with only three predictors, the adjusted R^2 doesn't change a lot from R^2. VIFs are pretty small, so the collinearity is not a problem. Since the p-value is less than 0.01, the overall F test is highly statistically significant. The t-test in multiple linear regression means that given the other two variables fixed, whether the new variable can add predictive power. Here we can see that the p-value for size and ticket price is less than 0.01, which means that they can add significant predictive power given the two other variables fixed. However, for a rating on TripAdvisor, it has a very high p-value, we have no evidence to think that the slope for the rating on TripAdvisor isn't zero. The coefficient for size says that given the other two variables are held fixed a one square mile increase in size is associated with an estimated expected increase in a total attendance of 10 people. The coefficient for Ticket Price says that given the two other variables are held fixed, increasing one dollar in the ticket price is associated with an estimated expected increase in total attendance of a 69.6 thousand people. The intercept is not meaningful here in the model. The value of s implies that a rough 95% prediction interval for the total attendance is $\pm$8.54 million. However, the total attendance in our dataset ranges from 1.12 to 20.96 million, the regression is not practically important.

Next, we use the residual plots to check the assumption. The residuals versus fitted values plot and normal plot of the residuals show that Bahamas Aquaventure Water Park is an outlier. The normal probability plot shows the non-normality of residuals and the residuals versus fitted values plot also indicates the evidence of nonconstant variance.



Since in our dataset, we have two different types of parks, one is the theme park and the other is the water park. Next, we take the type of parks into consideration. The graph down below is the side-by-side boxplots. We can see that there is a significant difference in total attendance based on the type of parks, with the theme parks having a significantly higher total attendance.



Therefore, let's form a constant shift model and run the regression again. We use 0 and 1 as the indicator variable to represent water parks and theme parks respectively. Here is the result for the regression:

Regression Equation

Total Attendance (in millions) = -6.39 + 7.50 Waterpark(0)/Themepark(1)
+ 0.000003 Size(square miles) + 0.0320 Ticket Price($)
+ 1.40 Rating on TripAdvisor

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-6.39	7.95	-0.80	0.428	
Waterpark(0)/Themepark(1)	7.50	1.47	5.09	0.000	1.88
Size(square miles)	0.000003	0.000002	1.24	0.226	1.48
Ticket Price($)	0.0320	0.0170	1.88	0.070	1.35
Rating on TripAdvisor	1.40	1.89	0.74	0.466	1.10

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.20301	73.95%	70.58%	65.47%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	902.67	225.668	22.00	0.000
Waterpark(0)/Themepark(1)	1	266.06	266.056	25.93	0.000
Size(square miles)	1	15.66	15.655	1.53	0.226
Ticket Price($)	1	36.10	36.101	3.52	0.070
Rating on TripAdvisor	1	5.59	5.595	0.55	0.466
Error	31	318.04	10.259		
Lack-of-Fit	29	315.08	10.865	7.34	0.127
Pure Error	2	2.96	1.479		
Total	35	1220.71			

The p-value for the F test is less than 0.01, so the overall statistics are highly statistically significant. The ticket price variable is marginally statistically significant, while the size variable and rating on TripAdvisor are not statistically significant. The newly added variable Park Type is highly statistically significant because the t-test for the Park type has a p-value less than 0.01. Also, R^2 improves greatly from 52.15% to 73.95%. So, we can say that there is a significant improvement for the constant shift model over the pooled model.

The standard error of the estimate is 3.2, which means that a rough 95% prediction interval for the total attendance is $\pm$6.4 million, less than the original pooled model. So, there is an improvement in terms of practical importance. The coefficient estimate for the park type is 7.5. This means that given the other variables size, ticket price and ratings fixed, the estimated expected difference between the theme park and the water park is 7.5 (million people) in total attendance. Therefore, we can see that there is a huge gap in the attendance between the theme park and the water park.

Next, we try to use the full model to see whether the regression will further improve.

Regression Equation

Total Attendance (in millions) = -0.03 - 5.2 Waterpark(0)/Themepark(1)
+ 0.000002 Size(square miles) - 0.0005 Ticket Price($)
+ 0.32 Rating on TripAdvisor + 0.000002 typesize
+ 0.0687 typeprice + 1.68 typerating

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.10923	77.83%	72.28%	67.80%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.03	9.43	-0.00	0.997	
Waterpark(0)/Themepark(1)	-5.2	17.3	-0.30	0.767	274.25
Size(square miles)	0.000002	0.000012	0.19	0.849	40.18
Ticket Price($)	-0.0005	0.0233	-0.02	0.984	2.68
Rating on TripAdvisor	0.32	2.25	0.14	0.887	1.65
typesize	0.000002	0.000012	0.14	0.893	55.07
typeprice	0.0687	0.0348	1.97	0.058	10.94
typerating	1.68	4.12	0.41	0.687	300.74

Analysis of Variance

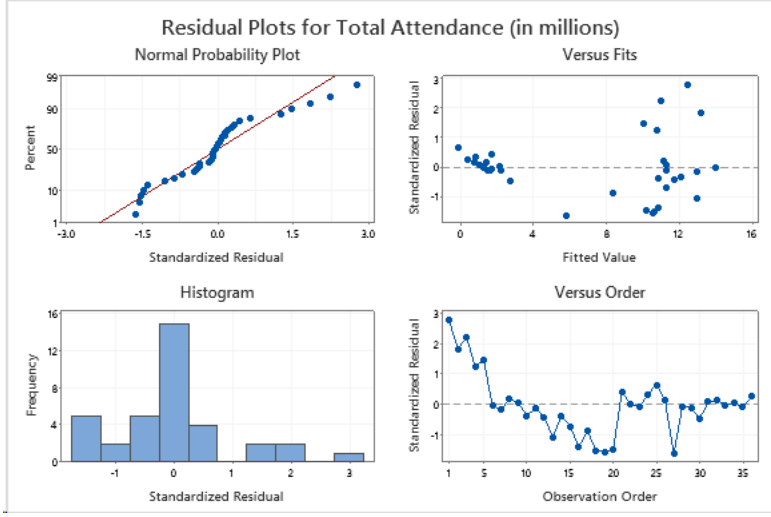
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	950.03	135.718	14.04	0.000
Waterpark(0)/Themepark(1)	1	0.87	0.866	0.09	0.767
Size(square miles)	1	0.36	0.357	0.04	0.849
Ticket Price($)	1	0.00	0.004	0.00	0.984
Rating on TripAdvisor	1	0.20	0.198	0.02	0.887
typesize	1	0.18	0.179	0.02	0.893
typeprice	1	37.64	37.639	3.89	0.058
typerating	1	1.60	1.603	0.17	0.687
Error	28	270.69	9.667		
Lack-of-Fit	26	267.73	10.297	6.96	0.133
Pure Error	2	2.96	1.479		
Total	35	1220.71			

We can conduct the partial F test to test the performance of the fit of the full model compared to the constant shift model:

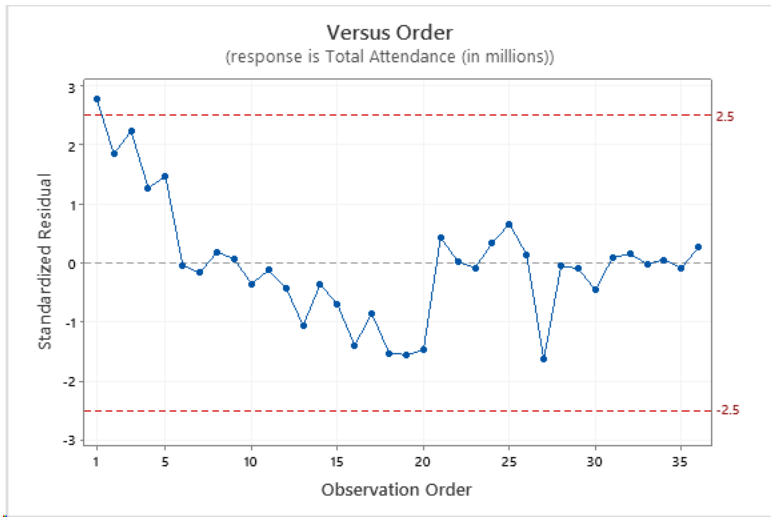
$$F = \frac{(318.04 - 270.69)/3}{270.69/(36 - 7 - 1)} = 1.63$$

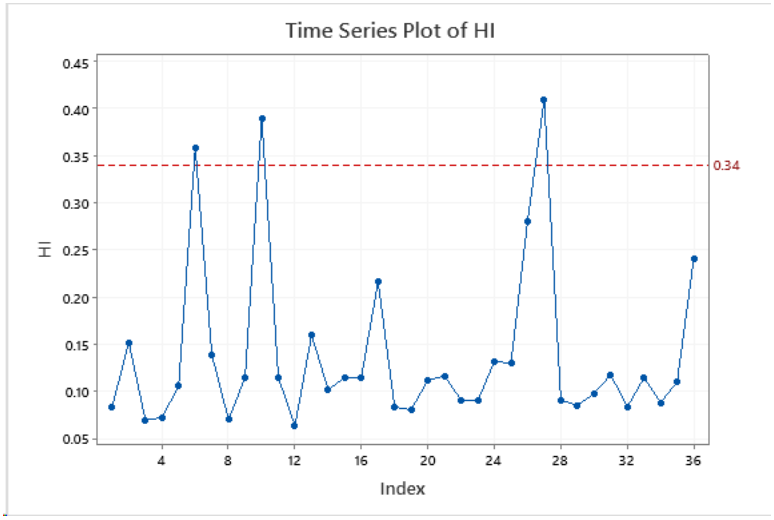
It has a tail probability of roughly 0.19, it's not statistically significant. Therefore, we can say that the constant shift model outperforms the full model (or we can say the constant shift model is the best among these three).

Here is the residual plot for the constant shift model:

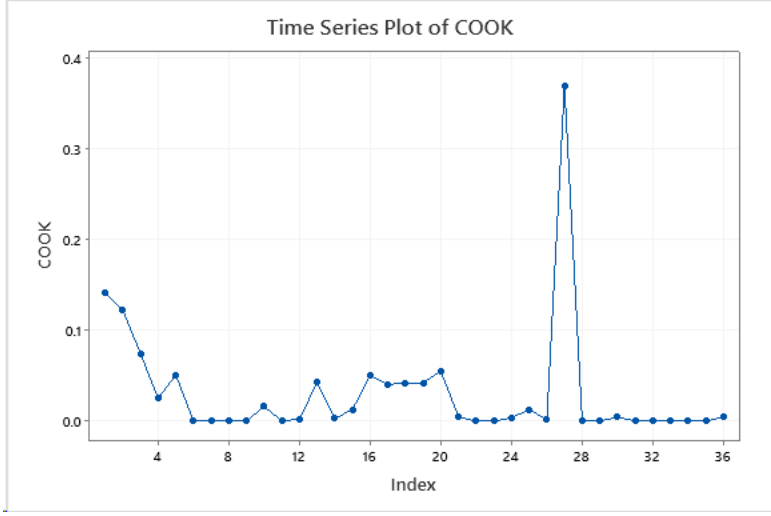


The assumptions are violated here, we witness non-normality and non-constant variance in our data. There are also several leverage points and outliers. We calculate the standardized residuals, leverages and Cook's distance to identify the outlier and leverage point.





As we add the reference line to both plots, we can see clearly that Magic Kingdom (Walt Disney World), USA appears as the outlier; Disney's Animal Kingdom, USA; Shanghai Disneyland, China and Bahamas Aquaventure Water Park, Bahamas appears as the leverage point.



There is no observation that is over 1 in terms of the Cook's distance, but there is one point that seems unusual from the other and has the Cook's distance of around 0.4. That's Bahamas Aquaventure Water Park. Therefore, Bahamas Aquaventure Water Park does indeed have a strong influence on the result of our regression.

Bahamas Aquaventure Water Park is located on Atlantis Paradise Island. It is a part of the Atlantis Resort Center. It focuses on offering tourists a luxurious vacation experience, which is probably why Bahamas Aquaventure Water Park is the most expensive park in the entire dataset in terms of ticket prices. Since the number of people going to the island for vacation is relatively smaller than other tourism places, so the total attendance is relatively small. As most of the other parks in my dataset are for all tourists, this park targets the tourists that go for a luxurious vacation. This is probably the reason why it becomes an unusual point.

Let's see what will happen if we remove the leverage points (Disney's Animal Kingdom, USA; Shanghai Disneyland, China and Bahamas Aquaventure Water Park, Bahamas):

WITHOUT LEVERAGE

Regression Analysis: Total Attendance (in millions) versus Size(square miles), Ticket Price($), Rating on TripAdvisor

Regression Equation

Total Attendance (in millions) = -11.94 + 0.000014 Size(square miles)
+ 0.0935 Ticket Price($) + 2.04 Rating on TripAdvisor

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-11.94	9.25	-1.29	0.207	
Size(square miles)	0.000014	0.000004	3.62	0.001	1.16
Ticket Price($)	0.0935	0.0192	4.86	0.000	1.10
Rating on TripAdvisor	2.04	2.22	0.92	0.367	1.13

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.61073	66.48%	63.02%	55.49%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	749.95	249.984	19.17	0.000
Size(square miles)	1	171.03	171.026	13.12	0.001
Ticket Price($)	1	308.06	308.065	23.63	0.000
Rating on TripAdvisor	1	10.97	10.970	0.84	0.367
Error	29	378.08	13.037		
Lack-of-Fit	27	375.12	13.894	9.39	0.101
Pure Error	2	2.96	1.479		
Total	32	1128.03			

WITHOUT LEVERAGE

Regression Analysis: Total Attendance (in millions) versus Size(square miles), Ticket Price($), Rating on TripAdvisor, Waterpark(0)/Themepark(1)

Regression Equation

Total Attendance (in millions) = -7.87 + 0.000006 Size(square miles) + 0.0542 Ticket Price($)
+ 1.48 Rating on TripAdvisor
+ 5.60 Waterpark(0)/Themepark(1)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-7.87	8.31	-0.95	0.352	
Size(square miles)	0.000006	0.000004	1.43	0.164	1.84
Ticket Price($)	0.0542	0.0215	2.52	0.018	1.75
Rating on TripAdvisor	1.48	1.98	0.75	0.461	1.14
Waterpark(0)/Themepark(1)	5.60	1.87	3.00	0.006	2.79

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.19715	74.63%	71.00%	65.77%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	841.83	210.456	20.59	0.000
Size(square miles)	1	20.88	20.883	2.04	0.164
Ticket Price($)	1	64.94	64.935	6.35	0.018
Rating on TripAdvisor	1	5.71	5.707	0.56	0.461
Waterpark(0)/Themepark(1)	1	91.87	91.874	8.99	0.006
Error	28	286.21	10.222		
Lack-of-Fit	26	283.25	10.894	7.36	0.126
Pure Error	2	2.96	1.479		
Total	32	1128.03			

WITHOUT LEVERAGE

Regression Analysis: Total Attendance (in millions) versus Size(square miles), Ticket Price($), Rating on TripAdvisor, Waterpark(0)/Themepark(1), typesize, typeprice, typerating

Regression Equation

Total Attendance (in millions) = -0.5 + 0.000003 Size(square miles) + 0.0027 Ticket Price($)
+ 0.39 Rating on TripAdvisor - 2.5 Waterpark(0)/Themepark(1)
+ 0.000004 typesize + 0.0710 typeprice + 0.79 typerating

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.5	11.2	-0.05	0.964	
Size(square miles)	0.000003	0.000014	0.20	0.840	19.89
Ticket Price($)	0.0027	0.0425	0.06	0.951	6.61
Rating on TripAdvisor	0.39	2.46	0.16	0.877	1.70
Waterpark(0)/Themepark(1)	-2.5	20.2	-0.12	0.904	314.64
typesize	0.000004	0.000015	0.27	0.792	33.94
typeprice	0.0710	0.0509	1.39	0.176	20.15
typerating	0.79	4.79	0.16	0.871	342.26

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.25524	76.52%	69.94%	62.88%

Analysis of Variance

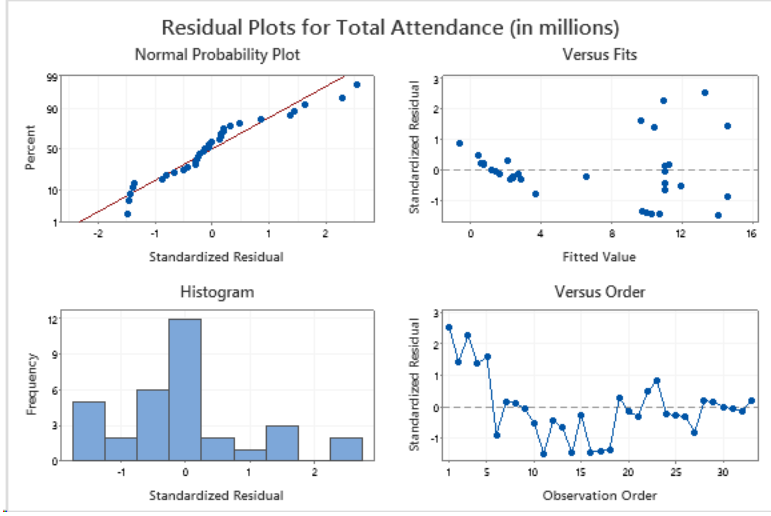
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	863.12	123.303	11.64	0.000
Size(square miles)	1	0.44	0.443	0.04	0.840
Ticket Price($)	1	0.04	0.042	0.00	0.951
Rating on TripAdvisor	1	0.26	0.261	0.02	0.877
Waterpark(0)/Themepark(1)	1	0.16	0.158	0.01	0.904
typesize	1	0.76	0.755	0.07	0.792
typeprice	1	20.58	20.576	1.94	0.176
typerating	1	0.29	0.285	0.03	0.871
Error	25	264.91	10.597		
Lack-of-Fit	23	261.96	11.389	7.70	0.121
Pure Error	2	2.96	1.479		
Total	32	1128.03			

It's apparent that the constant shift model is better than the pooled model. We then check whether dropping out leverage point will make the full model better than the constant shift model. The partial-F test for the full model compared with the constant shift model now is:

$$F = \frac{(286.21 - 264.91)/3}{264.91/(33 - 7 - 1)} = 0.67$$

It has a tail probability of around 0.58. This is not statistically significant. The constant shift model is still the best model. The fitted regression coefficients and regression output haven't changed a lot. R^2 improves a little bit from 73.95% to 74.63%.

Let's then check the assumptions for the chosen model (constant shift model). Here are the 4-in-1 residual plots. The normal probability plot shows the non-normality of residuals and the residuals versus fitted values plot also indicates some nonconstant variance. There are two points that appear to be outliers. This is Magic Kingdom (Walt Disney World), the USA and Tokyo Disneyland, Japan.



Then let's take a look if we omit both outliers: Magic Kingdom (Walt Disney World), USA and leverage points: Disney's Animal Kingdom, USA; Shanghai Disneyland, China and Bahamas Aquaventure Water Park, Bahamas. We run our regression again on the three models:

WITHOUT LEVERAGOUTLIER

Regression Analysis: Total Attendance (in millions) versus Size(square miles), Ticket Price($), Rating on TripAdvisor

Regression Equation

Total Attendance (in millions) = -10.39 + 0.000013 Size(square miles)
+ 0.0837 Ticket Price($) + 1.80 Rating on TripAdvisor

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-10.39	8.71	-1.19	0.243	
Size(square miles)	0.000013	0.000004	3.77	0.001	1.15
Ticket Price($)	0.0837	0.0186	4.50	0.000	1.08
Rating on TripAdvisor	1.80	2.09	0.86	0.396	1.11

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.38891	65.17%	61.43%	53.68%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	601.580	200.527	17.46	0.000
Size(square miles)	1	163.520	163.520	14.24	0.001
Ticket Price($)	1	232.589	232.589	20.25	0.000
Rating on TripAdvisor	1	8.535	8.535	0.74	0.396
Error	28	321.572	11.485		
Lack-of-Fit	26	318.614	12.254	8.28	0.113
Pure Error	2	2.958	1.479		
Total	31	923.153			

WITHOUT LEVERAGOUTLIER

Regression Analysis: Total Attendance (in millions) versus Size(square miles), Ticket Price($), Rating on TripAdvisor, Waterpark(0)/Themepark(1)

Regression Equation

Total Attendance (in millions) = -5.97 + 0.000005 Size(square miles) + 0.0414 Ticket Price($)
+ 1.19 Rating on TripAdvisor
+ 5.90 Waterpark(0)/Themepark(1)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-5.97	7.45	-0.80	0.429	
Size(square miles)	0.000005	0.000004	1.40	0.172	1.82
Ticket Price($)	0.0414	0.0197	2.10	0.045	1.72
Rating on TripAdvisor	1.19	1.77	0.67	0.507	1.12
Waterpark(0)/Themepark(1)	5.90	1.67	3.53	0.002	2.73

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.85436	76.17%	72.64%	67.25%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	703.174	175.794	21.58	0.000
Size(square miles)	1	16.046	16.046	1.97	0.172
Ticket Price($)	1	35.944	35.944	4.41	0.045
Rating on TripAdvisor	1	3.684	3.684	0.45	0.507
Waterpark(0)/Themepark(1)	1	101.594	101.594	12.47	0.002
Error	27	219.978	8.147		
Lack-of-Fit	25	217.020	8.681	5.87	0.156
Pure Error	2	2.958	1.479		
Total	31	923.153			

WITHOUT LEVERAGOUTLIER

Regression Analysis: Total Attendance (in millions) versus Size(square miles), Ticket Price($), Rating on TripAdvisor, Waterpark(0)/Themepark(1), typesize, typeprice, typerating

Regression Equation

Total Attendance (in millions) = -0.5 + 0.000003 Size(square miles) + 0.0027 Ticket Price($)
+ 0.39 Rating on TripAdvisor - 0.0 Waterpark(0)/Themepark(1)
+ 0.000003 typesize + 0.0546 typeprice + 0.54 typerating

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.5	10.1	-0.05	0.961	
Size(square miles)	0.000003	0.000013	0.23	0.823	19.59
Ticket Price($)	0.0027	0.0384	0.07	0.945	6.13
Rating on TripAdvisor	0.39	2.23	0.17	0.864	1.68
Waterpark(0)/Themepark(1)	-0.0	18.3	-0.00	0.998	307.26
typesize	0.000003	0.000013	0.23	0.819	33.24
typeprice	0.0546	0.0465	1.17	0.252	19.15
typerating	0.54	4.33	0.12	0.902	332.55

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.94170	77.50%	70.94%	63.24%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	715.466	102.209	11.81	0.000
Size(square miles)	1	0.443	0.443	0.05	0.823
Ticket Price($)	1	0.042	0.042	0.00	0.945
Rating on TripAdvisor	1	0.261	0.261	0.03	0.864
Waterpark(0)/Themepark(1)	1	0.000	0.000	0.00	0.998
typesize	1	0.464	0.464	0.05	0.819
typeprice	1	11.945	11.945	1.38	0.252
typerating	1	0.133	0.133	0.02	0.902
Error	24	207.686	8.654		
Lack-of-Fit	22	204.728	9.306	6.29	0.146
Pure Error	2	2.958	1.479		
Total	31	923.153			

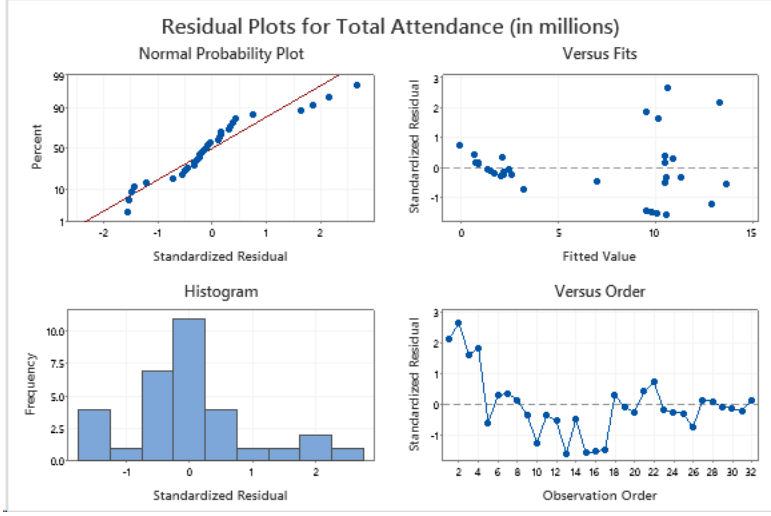
Since the constant shift model is apparently better than the pooled model. Let's check whether the full model or constant shift model is better without outliers. The partial-F test for the full model compared with the constant shift model is:

$$F = \frac{(219.978 - 207.686)/3}{207.686/(32 - 7 - 1)} = 0.473$$

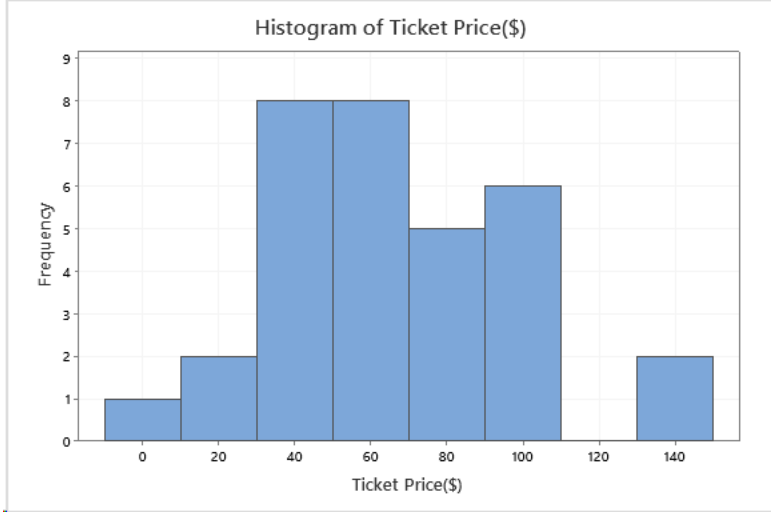
The p-value is around 0.706. It's even higher. This is not statistically significant. The constant shift model still outperforms the rest two.

The fitted regression coefficients and regression output haven't changed a lot. R^2 improves a little bit from 74.73% to 76.17%. The standard error of the estimate is down to 2.85, much smaller than our original model's.

Let's check the assumptions for the constant shift model that we choose. We can see a right-tailed distribution in the histogram. There is still a non-normality pattern in our residuals and heteroscedasticity still occurs.



In order to satisfy the assumption of regression more completely, we try to implement transformation for our model. Since there is money data (ticket price) in our model.



Because the number of our observations is relatively small, it's hard to identify the right-tailed distribution. But we will still take log on ticket price and run the regression again to see what will happen:

Regression Equation

Total Attendance (in millions) = -9.72 + 6.84 Waterpark(0)/Themepark(1)
+ 0.000004 Size(square miles) + 1.36 Rating on TripAdvisor
+ 3.18 logticketprice

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-9.72	8.85	-1.10	0.282	
Waterpark(0)/Themepark(1)	6.84	1.68	4.07	0.000	2.51
Size(square miles)	0.000004	0.000004	1.07	0.294	1.76
Rating on TripAdvisor	1.36	1.85	0.73	0.470	1.12
logticketprice	3.18	2.52	1.26	0.218	1.59

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.99184	73.82%	69.94%	64.44%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	681.472	170.368	19.03	0.000
Waterpark(0)/Themepark(1)	1	148.460	148.460	16.59	0.000
Size(square miles)	1	10.256	10.256	1.15	0.294
Rating on TripAdvisor	1	4.816	4.816	0.54	0.470
logticketprice	1	14.242	14.242	1.59	0.218
Error	27	241.681	8.951		
Lack-of-Fit	25	238.722	9.549	6.46	0.143
Pure Error	2	2.958	1.479		
Total	31	923.153			

The overall F test is still statistically significant. The coefficient of log of ticket price means that holding everything else is fixed, multiplying the indicator variable park type by 10 is associated with an expected increase of 3.18 million in total attendance. However, as we can see the p-value for the log of the ticket price is greater than 0.1, which means that it is not statistically significant. So, it's not meaningful to taking logs. We will abandon this model.

Lastly, let's perform the model selection to find the best balance between fit and simplicity. We use the best subset regression and here is the output:

Response is Total Attendance (in millions)

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Waterpark(0)/Themepark(1)
1	70.8	69.8	304.2	67.0	5.1	2.9969	165.848	169.388	1.000	X
1	42.8	40.9	588.0	36.3	36.8	4.1947	187.368	190.908	1.000	
2	73.7	71.9	309.6	66.5	3.8	2.8918	165.103	169.484	4.236	X
2	71.8	69.8	311.5	66.3	6.0	2.9969	167.389	171.770	4.516	X
3	75.8	73.2	298.7	67.6	3.5	2.8263	165.341	170.362	8.839	X
3	74.4	71.7	312.7	66.1	5.0	2.9033	167.062	172.083	4.795	X
4	76.2	72.6	302.3	67.3	5.0	2.8544	167.861	173.296	9.602	X

Vars	Size (square miles)	Rating on TripAdvisor	logticketprice
1			
1		X	
2		X	
2	X		
3	X	X	
3		X	X
4	X	X	X

The model with three predictors park type, size and the ticket price is the best choice, in terms of both the criteria of choosing the model that maximizes R^2_{Pred}, or minimizing AIC_c. However, if we go with the criteria of choosing the one that $C_p \approx p + 1$, the model with four predictors is the best choice. Here, I am going to go for simplicity and choose the model with three predictors park type, size and ticket price.

Let's run our regression again with variable park type, size and ticket price.

Regression Equation

Total Attendance (in millions) = -1.05 + 6.01 Waterpark(0)/Themepark(1)
+ 0.000006 Size(square miles) + 0.0419 Ticket Price($)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.05	1.33	-0.79	0.437	
Waterpark(0)/Themepark(1)	6.01	1.65	3.65	0.001	2.70
Size(square miles)	0.000006	0.000004	1.54	0.136	1.78
Ticket Price($)	0.0419	0.0195	2.15	0.041	1.71

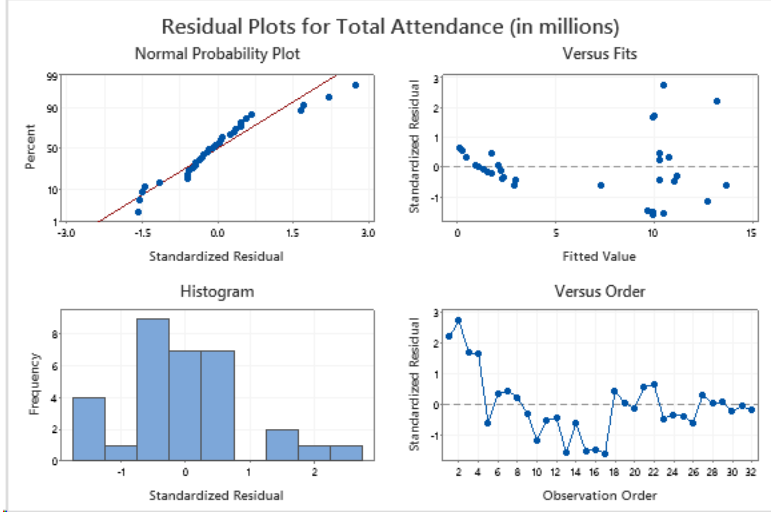
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.82629	75.77%	73.18%	67.64%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	699.490	233.163	29.19	0.000
Waterpark(0)/Themepark(1)	1	106.445	106.445	13.33	0.001
Size(square miles)	1	18.843	18.843	2.36	0.136
Ticket Price($)	1	36.803	36.803	4.61	0.041
Error	28	223.662	7.988		
Lack-of-Fit	26	220.704	8.489	5.74	0.159
Pure Error	2	2.958	1.479		
Total	31	923.153			

The p-value for the F test is less than 0.01, so the overall statistics are highly statistically significant. The variable Park Type is high statistically significant, the ticket price variable is statistically significant, and the size variable is not significant. VIF is small, so the collinearity is not a problem. The model has a rough 95% prediction interval of ± 5.64. The practical importance improves a lot from our very first model.



The assumptions are still violated. There is still a non-normality pattern in the normal probability plot of residuals. The residuals still follow the right-tailed distribution in the histogram of the residuals. Heteroscedasticity still occurs in the plot of residuals versus fitted value. The park with a smaller number of total attendances tends to have fewer residuals.

In conclusion, the simpler constant shift model with three variables park type, size and ticket price is the model I choose in the end. When I made the assumption, I was hoping to see a stronger relationship. As a result, we only see the indicator variable of the park type is statistically significant in predicting the total attendance. However, ratings on TripAdvisor and size variables are not significant at all. I think this is partly because the ratings on TripAdvisor only have an accuracy of 0.5 and most of the parks' ratings are similar to each other. So, people won't take it as an important reference to decide whether to go to the park or not. Moreover, it's interesting to observe that the relationship between total attendance and the ticket price is positive. Because logically speaking, the higher the ticket price, the less likely people want to afford the ticket. I guess this is because the ticket price might vary from countries and may

depend on the price level of each country. So simply using the ticket price as a variable is not enough.

For further improvement, I am interested to see whether the number of parks in the nearby region can have any influence on the total attendance. For example, for Walt Disney World Resort in Orlando, it has four theme parks in the resort in total, which are Magic Kingdom, Disney's Animal Kingdom, Epcot and Disney's Hollywood Studios. Tourists who travel there usually won't visit them all, so these theme parks are in a relationship of substitution. This might influence the total attendance of each of these four parks.