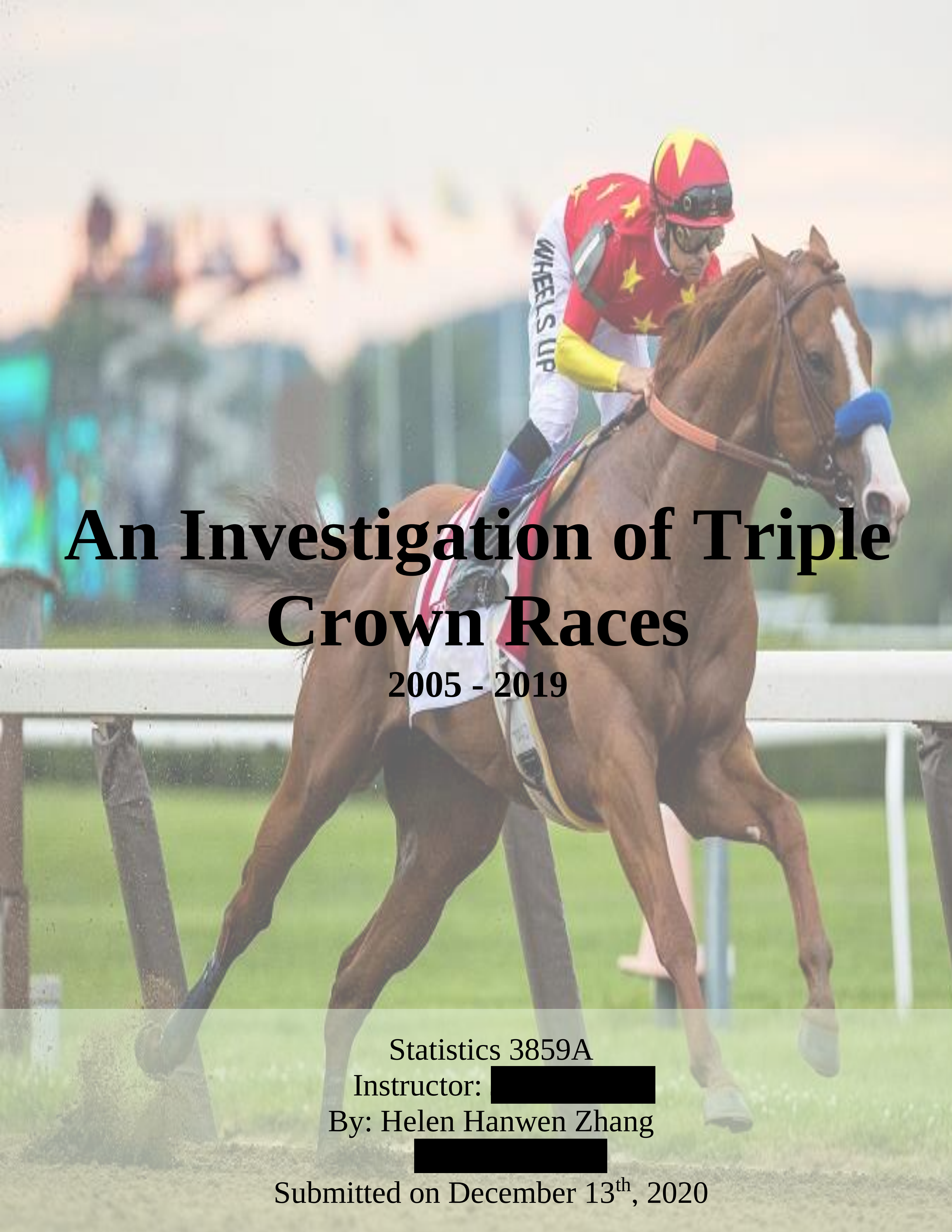

An Investigation of Triple Crown Races

2005 - 2019

Statistics 3859A
Instructor: [redacted]
By: Helen Hanwen Zhang
[redacted]
Submitted on December 13th, 2020

Introduction

To be crowned the coveted Triple Crown title, a horse would need to win the Kentucky Derby, the Preakness Stakes, and the Belmont Stakes. Beginning in 1919, this annual tradition has only seen 13 Triple Crown winners including the most recent winner, Justify, back in 2018. Traditionally run in May and early June of each year, winning the Triple Crown has been considered one of the most elusive and celebrated achievements in all sports. Although the Triple Crown races are restricted to 3-year-old thoroughbred horses, there are other factors that may affect their final place in a race. In this report, factors such as their pole position, odds of winning, and payout of a: win bet, place bet, and show bet will be scrutinized. Due to insufficient data, it is important to recognize that other factors such as the track condition, weather condition, and track length all have potential to influence a horse's final place as well.

A similar Triple Crown analysis was investigated by a group of Harvard students who wanted to see if there was a relationship between a horse's final place and their place bet. After conducting their data analysis, they concluded that the horse who had a mediocre place bet had the highest likelihood of winning a Triple Crown. On the other hand, the horse who had an extremely high place bet was more susceptible to be placed near the bottom half of the race. Their conclusion indicates a strong betting strategy moving forward (Meyer, 2005).

Kentucky Derby

The Kentucky Derby is held annually in Louisville, Kentucky. Nicknamed the "The Most Exciting Two Minutes in Sports", this race is a 2 km Grade I stake at Churchill Downs. From 2005 to 2019, the data set features 294 horses who competed in the Kentucky Derby. Once the fitted model was obtained, appropriate graphical and testing approaches were applied to see if any contraventions existed. Referring to *Figure 1*, it is evident to see that the linearity, equal variance, and normality was violated without any model corrections. As the residual plot did not exhibit a zero mean, linearity did not hold and thus, a parabolic relationship was observed between the residuals and the fitted values. Furthermore, the residual plot showed that a constant variance did not exist as the left side was significantly narrower and had fewer points. By conducting the Breusch-Pagan (BP) test, it confirmed this violation by displaying a small p-value of 0.254. While the Q-Q plot should show that the residuals correspond to a normal distribution, the left tail of the distribution of the residuals was shorter than a standard normal distribution. The small p-value of 0.0015 from the Shapiro-Wilk (SW) test supported this conclusion.

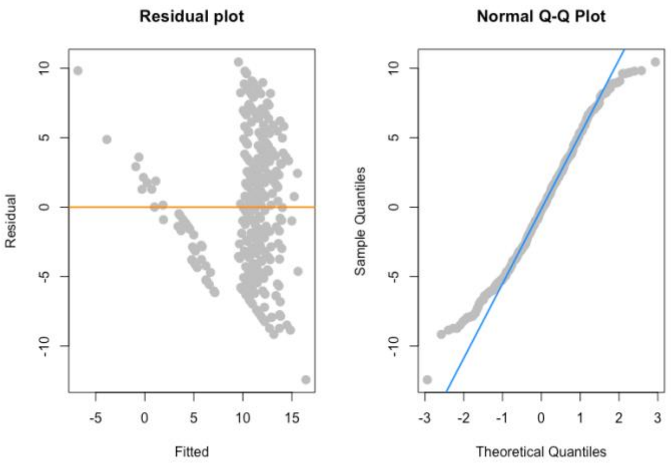

Figure 1
(Residual and QQ Plot Without Any Model Corrections)

To attempt to fix these model assumptions, influential points were removed by using Cook's distance with a threshold of 4/n. Out of the 4 influential points, 2 outliers were observed. Under the presumption that the influential points were simple measurement errors, a new residual plot and Q-Q plot was created (*Figure 2*). Despite eliminating the influential points, the plots showed that the linearity, equal variance, and normality was still violated. Although the residual plot was more scattered than the previous plot, it still did not display a zero mean. The equal variance remains uneven on the right and left sides. In terms of equal variance, the smaller p-value of 0.0316 indicates the model was better off without any corrections. Lastly, the Q-Q plot still has a left tail, but it is closer to the standard normal distribution than the plot without any corrections. While the elimination of influential points did not fix the model assumptions, there was an evident improvement.

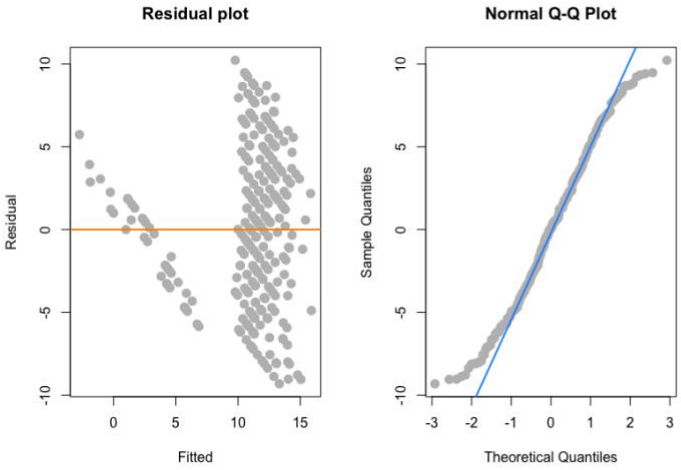

Figure 2
(Residual and QQ Plot Without Any Influential Points)

The next method of trying to correct the plots is to conduct a Box-Cox to find the best transformation for the response variable final_place. According to *Figure 3*, the log-likelihood was maximized at a lambda around 0.52. With this information, the residual and Q-Q plot can be remodelled once the parabolic trend of the fitted values was removed.

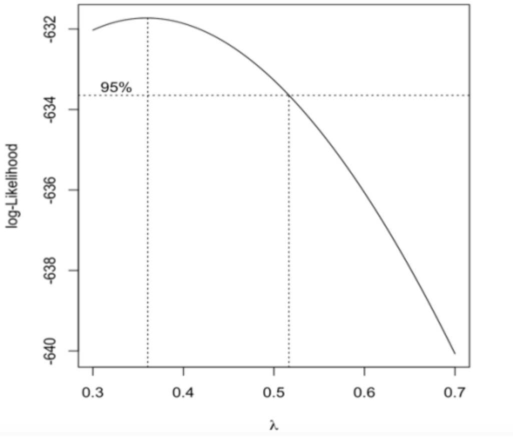

Figure 3
(Box-Cox Log-Likelihood)

After investigating *Figure 4*, the linearity, equal variance, and normality still do not hold. It showed a significant improvement from the model with no corrections (*Figure 1)*. The residual plot was more widely disperse and the Q-Q line was closer to the standard normal distribution. By performing the Box-Cox method, the highest p-value at 0.607 was observed for equal variance. Other than that, compared to the plots where the influential plots were removed *(Figure 2)*, it did not show any notable changes. The exceedingly high number of zeros in the win bet, place bet, and show bet columns of the data set may be responsible for the lack of changes in the residual plot and the Q-Q plot.

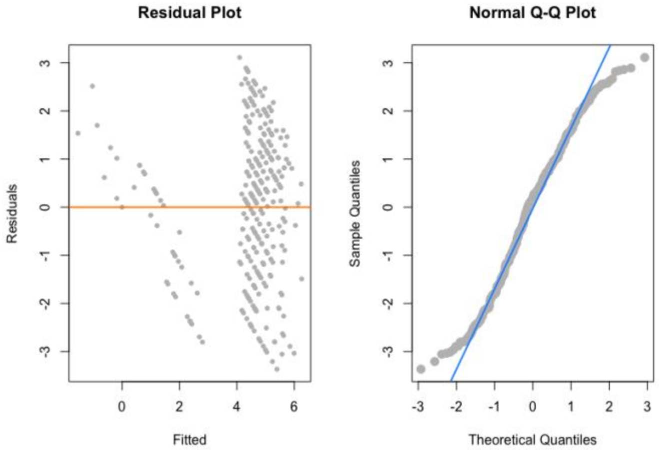

Figure 4
(Residual and Q-Q Plot Box-Cox Method)

The last method for correcting the model assumptions was to create a polynomial model *(Figure 5)*. However, similar to the previous graphs, everything was still violated. At first glance, equal variance does not hold as the width of the right side is far greater than the left side. This was confirmed by the small p-value of 0.0006789 from the BP test. Although the normality was also violated, the p-value from the SW test of 0.003087 showed that it had the highest p-value out of all the correctional methods.

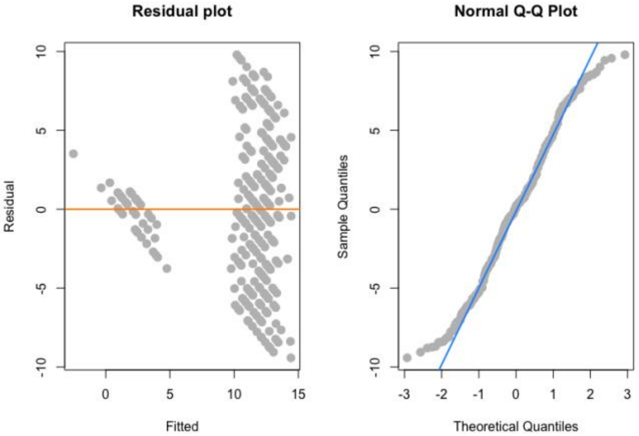

Figure 5
(Residual and Q-Q Plot Polynomial)

	No Model Corrections	Influential Points Removed	Box-Cox	Polynomial Model
Breusch-Pag P-Value	0.254	0.0316	0.607	0.0006789
Shaprio-Wilk P-Value	0.001482	0.001162	0.0001826	0.003087

Figure 6
(Kentucky P-Values Table)

Preakness Stakes

The Preakness Stakes is the second race of the Triple Crown and it is held at Pimlico Race Course in Baltimore, Maryland. This 1.9 km race termed, “The Run for the Black-Eyed Susans” because of the blanket of Maryland’s state flowers that are placed across the withers of the winning horse has 163 rows of data from 2005 to 2019. Like the Kentucky plot, the model assumptions were violated without any corrections *(Figure 7)*. Even though the linearity of the P reakness Stakes data does not hold, it is far more spread out than the Kentucky plot *(Figure 1)*. Both larger than the Kentucky p-values, the small p-values from the BP and SW test at 0.005294 and 0.008294, respectively, confirm the suspicion that equal variance and normality does not me et the linear regression standards *(Figure 12)*.

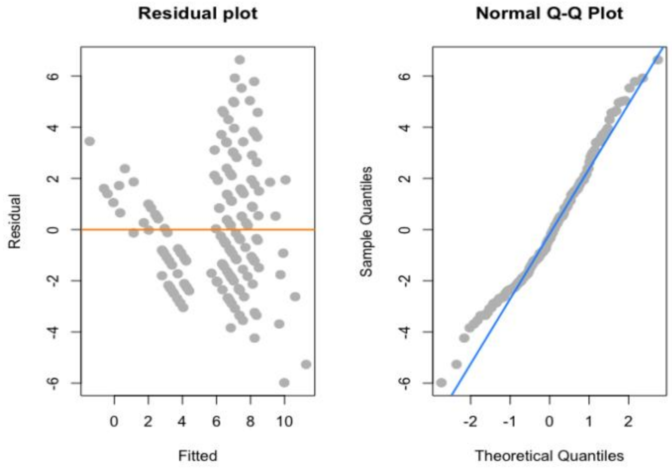

Figure 7
(Residual and Q-Q Without Any Corrections)

By using Cook's distance, 7 influential points and 3 outliers were found from the Preakness data. Once they were removed, *Figure 8* showed an improvement in equal variance by having a higher p-value at 0.04671 than the residual plot without any corrections *(Figure 12).* Additionally, the residual plot was much more scattered than the previous Preakness plot. In terms of normality, the elimination of influential plots hindered the Q-Q plot by resulting in a smaller p-value at 0.001382. Even with an enhancement in linearity and equal variance, none of the model assumptions have yet to hold.

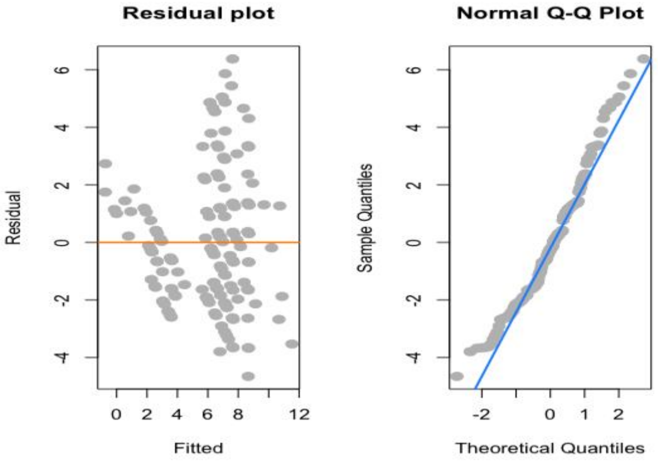

Figure 8
(Residual and QQ Plot Without Any Influential Points)

For the third method, a Box-Cox log likelihood was applied. By looking at *Figure 9*, a l ambda of 0.44 was used as the transformed response variable.

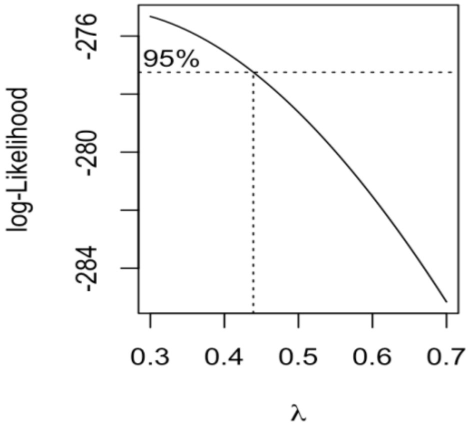

Figure 9
(Box-Cox Log-Likelihood)

By inspecting the residual plot and Q-Q plot in *Figure 10* as well as the p-values in *Figure 12*, the Box-Cox method has been the prominent correctional method thus far. It resulted in the highest p-values for equal variance and normality at 0.1296 and 0.3011, respectively. Looking at the Q-Q plot, the line appeared linearly which confirmed the idea that the normality holds. The linearity appeared to continue its pattern of not holding, and it does not appear to have a substantial improvement from the last method. However, it looks much more scattered than the model without any corrections which indicates the Box-Cox has been beneficial.

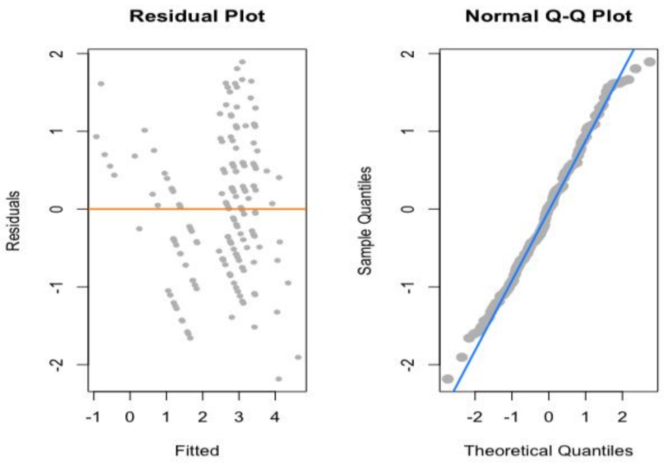

Figure 10
(Residual and Q-Q Plot Box-Cox Method)

Lastly, a polynomial model was plotted as an effort to fix the model assumptions *(Figure 11)*. The sizeable difference of data points on the left and right sides of the residual plot is a in dication that the linearity is violated. With the smallest p-value out of all of the methods at 0.0002215, the equal variance does not meet the linear regression requirement *(Figure 12).* After observing the QQ-plot, its right tail, and its small p-value at 0.04008, it too violates the model assumptions. After analyzing the Preakness data, the most advantageous method for correcting linearity, equal variance, and normality is the Box-Cox method. The weakest method for linearity and equal variance was the polynomial model and the weakest for normality was the model without any corrections.

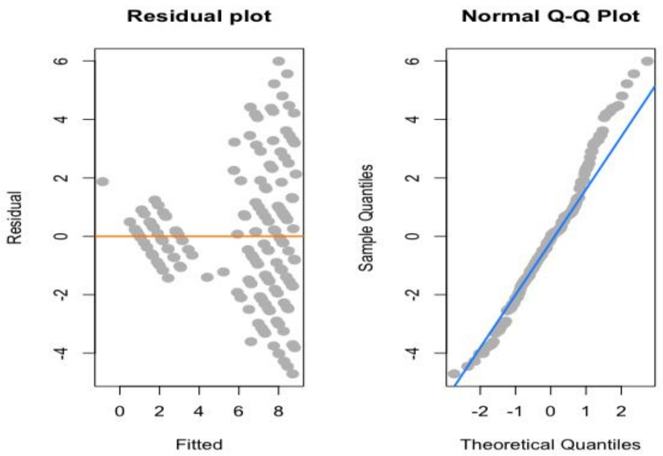

Figure 11
(Residual and Q-Q Plot Polynomial)

	No Model Corrections	Influential Points Removed	Box-Cox	Polynomial Model
Breusch-Pag P-Value	0.005294	0.04671	0.1296	0.0002215
Shaprio-Wilk P-Value	0.008294	0.001382	0.3011	0.04008

Figure 12
(Preakness P-Values Table)

Belmont Stakes

Labelled the "Test of the Champion" and "The Run for the Carnations", the Belmont Stakes is the final leg of the Triple Crown. With a purse of $1 million USD, the race is held in June in Elmont, New York. The Belmont Stakes is one of the top-attended events for American thoroughbred racing, and in 2004, it drew an audience of 21.9 million viewers. As a result of not making any model corrections, *Figure 13* showed that linearity and normality does not meet the linear regression benchmark. However, the residual plot showed a noticeable pattern for equal variance. To confirm this presumption, a BP test was conducted to show its large p-value at 0.6645 *(Figure 18).*

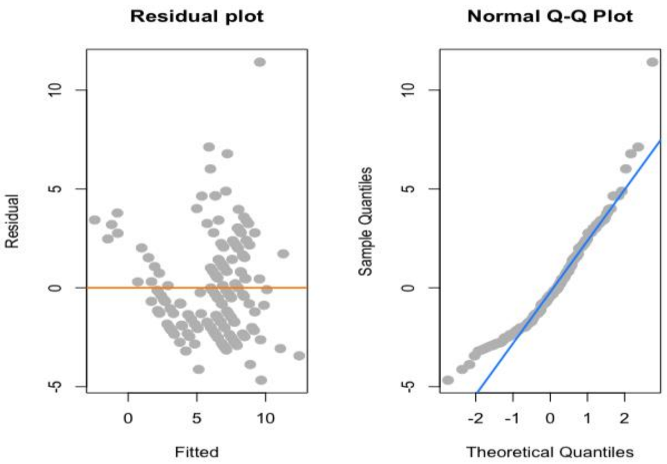

Figure 13
(Residual and QQ Plot Without Any Model Corrections)

Although equal variance was held in the previous plot without any corrections, linearity, and normality need to be fixed. After deleting 12 influential points from the model, only the normality appeared to be bettered. The left tail seemed to follow a more standard normal distribution than *Figure 13.* The p-value from the BP test decreased from 0.6645 to 0.03926, and the linearity remained unchanged. Thus, deleting the influential points turned out to be disadvantageous, and the model should have remained without any corrections.

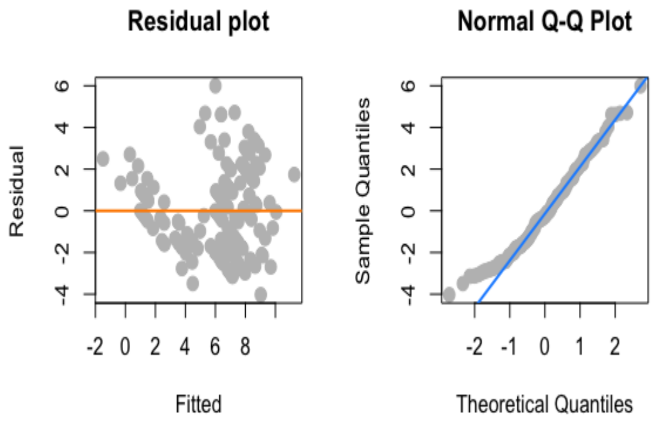

Figure 14
(Residual and QQ Plot Without Any Influential Points)

Though *Figure 14* did not show the desired results, the Box-Cox method proved useful by yielding higher than previously seen p-values from the BP and SW test. At a lambda equal to 0.41 *(Figure 15)*, the three assumptions were still violated, but compared to the influential point method, *Figure 16* showed a much more scattered residual plot. Granted the Q-Q plot continues to have a left tail, it has moved closer to the standard normal line than the plot without any correc tions. It is safe to say that the Box-Cox has made tremendous efforts to correct the model assumptions.

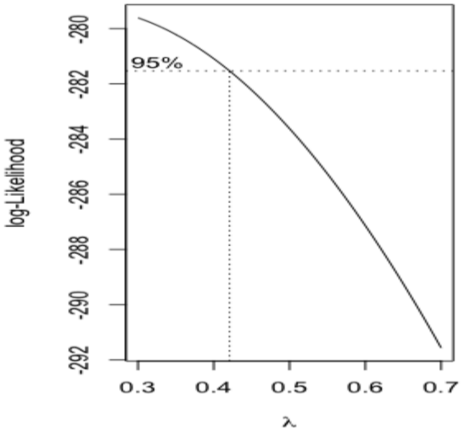

Figure 15
(Box-Cox Log-Likelihood)

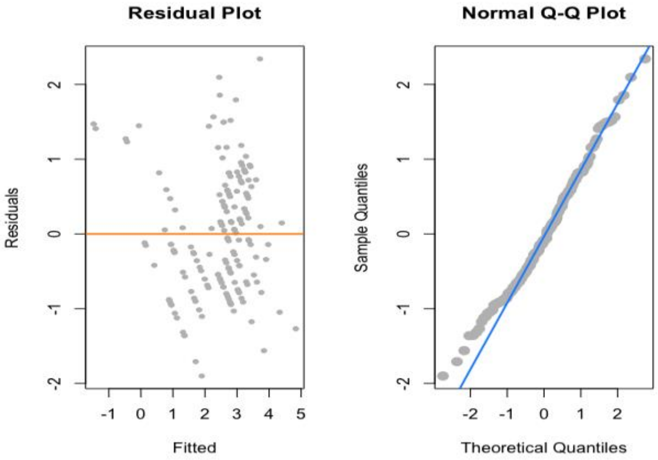

Figure 16
(Residual and Q-Q Plot Box-Cox Method)

Once again, the polynomial model had proven that it is the weakest correctional method for linearity as the BP test resulted in the smallest p-value than any other method. In the Belmont Stakes case, it also resulted in the smallest p-value for the normality assumption (*Figure 18).* Similar to the Preakness, the most helpful method for fixing linearity, equal variance, and normality for the Belmont Stakes data was the Box-Cox method. The weakest method was the polynomial model.

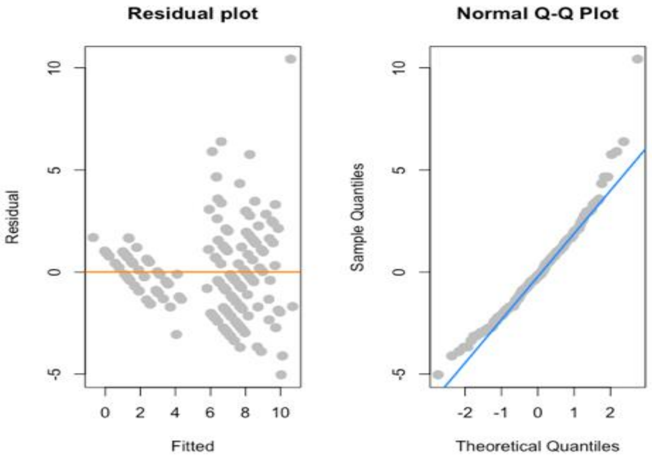

Figure 17
(Residual and Q-Q Plot Polynomial)

	No Model Corrections	Influential Points Removed	Box-Cox	Polynomial Model
Breusch-Pag P-Value	0.6645	0.03926	0.3223	0.03918
Shaprio-Wilk P-Value	1.32e-06	0.008393	0.1036	9.462e-05

Figure 18
(Belmont P-Values Table)

The Most Influential Predicator

To find the most influential predictor for the response variable, final_place, the Variance Inflation Factor (VIF) was used. The Akaike's Information Criteria (AIC) was conducted to find the best subset of predictors to predict final_place. First, a forward stepwise function with AIC found that the best predictors to anticipate final_place for the Kentucky Derby were odds of winning, place bet, a nd show bet. For the Preakness, the best predictors were odds of winning, win bet, and show bet. Lastly, for the Belmont Stakes, the pole position, odds of winning, and place bet were the most accurate predictors. As for the most influential predictor for each race, *Figure 19* showed that the most influential predictor for all three Triple Crown races was the show bet. The least influential predictor for the Kentucky Derby and the Preakness Stakes was the pole position and the odds bet for the Belmont Stakes. The importance of understanding the existence of collinearity between variables is crucial for people who want to correctly bet on a winning horse.

	Pole Position	**Odds Bet**	**Wining Odds**	**Place Bet**	**Show Bet**
Kentucky	1.010105	1.056439	1.061261	1.152655	1.344965
Preakness	1.006634	1.180716	1.146457	1.281744	1.632098
Belmont	1.11079	1.101545	1.135636	1.282019	1.640384

Figure 19
(VIFs)

Conclusion

In summation, the most helpful correctional method for linearity, equal variance, and normality is the Box-Cox method. This had been proven through embedding various methods to the data sets of the Kentucky Derby, the Preakness Stakes, and the Belmont Stakes. With regard to the most influential predictor for the response variable final_place, the lower the show bet variable, the more likely a horse and its jockey will win that particular race. If the race is either the Kentucky Derby or the Preakness Stakes, a change in pole position will not affect their chances of winning as it has a low collinearity relationship with the response variable. Similarly, an odds bet will not heavily influence which horse will win the Belmont Stakes as it is the least correlated variable to final_place.

Works Cited

Meyer, Breit. “Dataset - Racing The Odds: Horse Racing from Then to Now.” *Google Sites*, 2005, sites.google.com/a/college.harvard.edu/pretty_ponies/dataset.