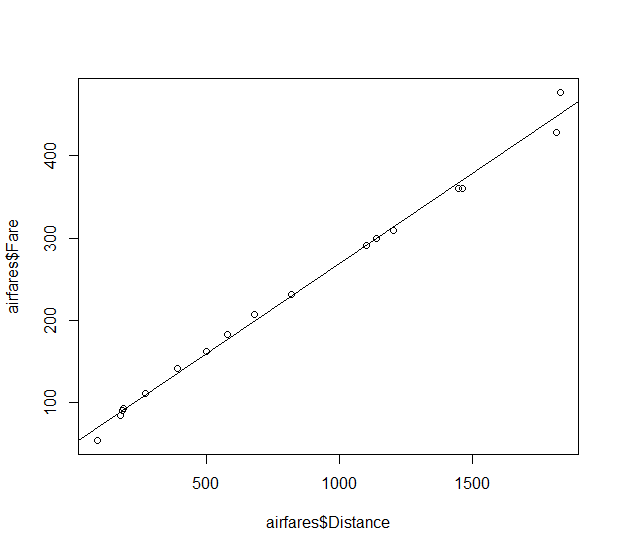
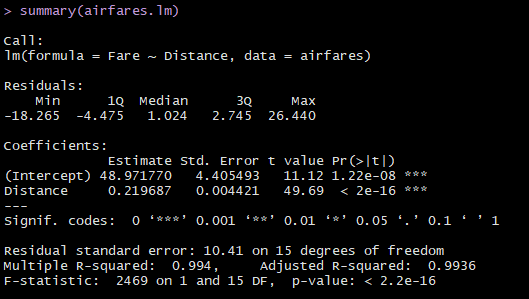
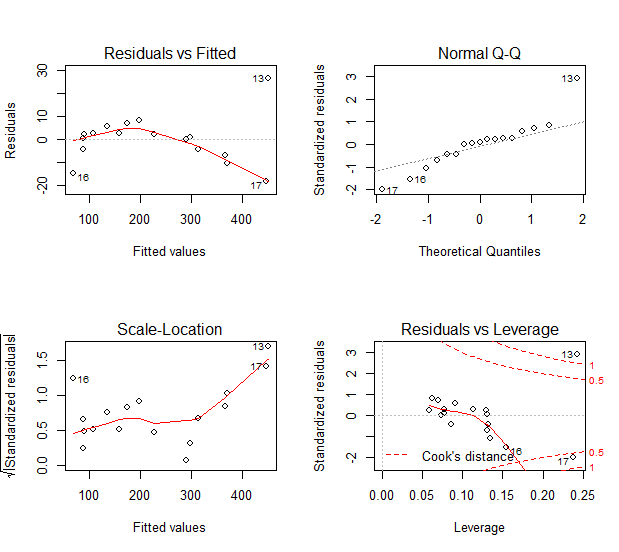
#1 Airfares



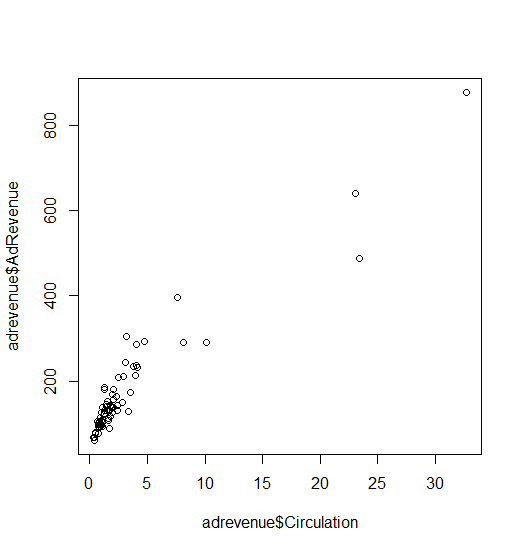


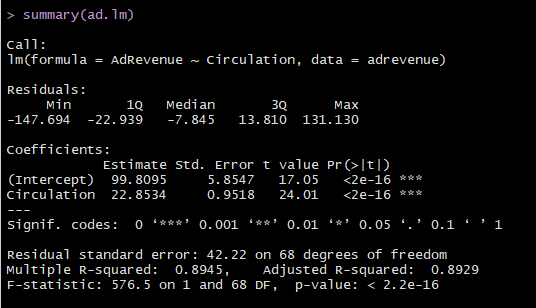


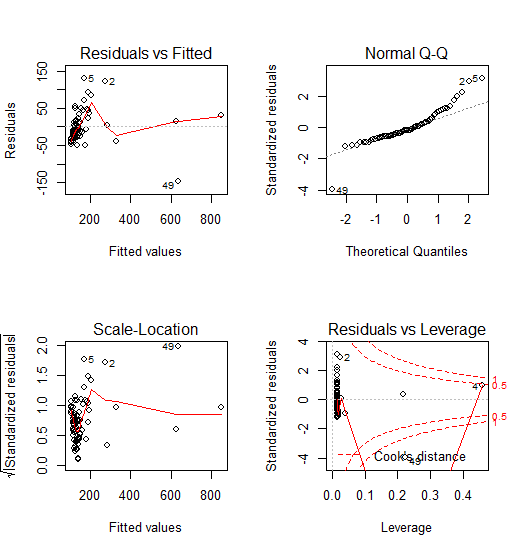
1. After looking at the residuals for the model, I would say that the model does well except for the long distance flights. The variance isn’t constant the farther your flight is.
2. It looks like the straight line regression model would fit well, for these data points, but I feel like if there were more long flights, it would start to look worse.

#3 AdRevenue

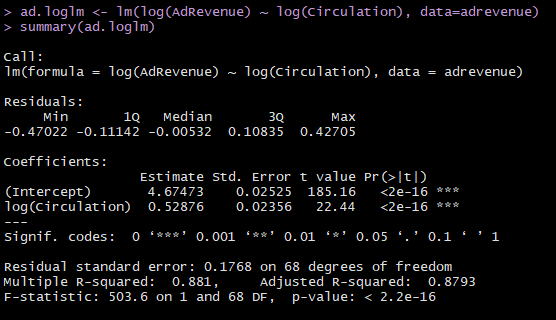
Part A

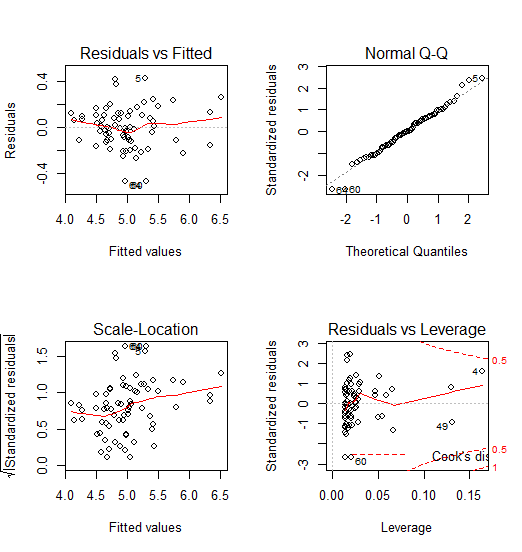


1. 

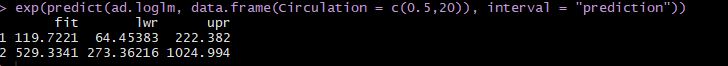


Above you will find the linear model that does not transform any of the variables. It doesn’t fit very well at all. I decided to log both the response and predictor variables and came up with the following model:





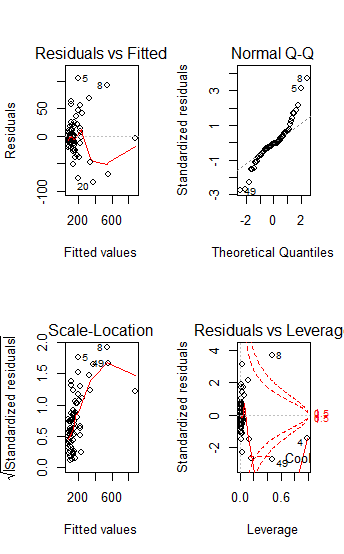
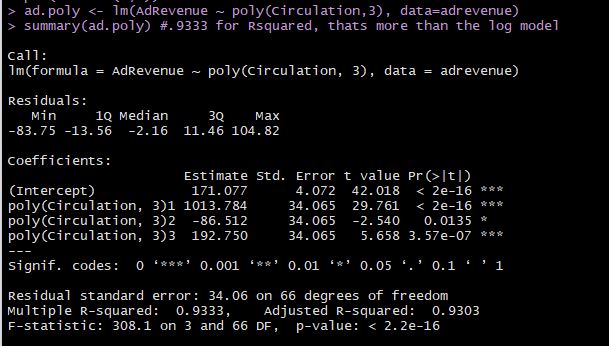
After doing a Log transform on the variables, it looks like the residuals are a little more normal, with the exception of one tail.



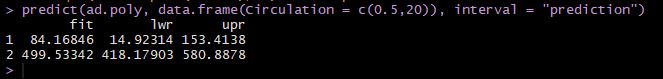
For .05 million in circulation, it’s predicted to have 119.72 million in revenue, with a lower and upper bound of 64.4 million and 222.38 million respectively. For 20 million in circulation, it’s predicted to have 529.33 million in revenue, with a lower and upper bound of 273.36 and 1024.99 million respectively.

1. One big weakness I see is that the confidence band for the prediction is really wide. Another one is that the residuals still don’t look all that great. But they do look better than a non-transformed model

Part B



As you can see, the residuals still don’t look that great, but the Rsquared value is higher.



But, the prediction for the revenue looks better. The band for the larger circulations is a lot tighter.

Part C

1. Both models don’t seem to be that great in fitting the data, so it would be hard to choose really. I think I might choose the log transform model, just because the residuals look a lot better.
2. For the prediction, I would probably choose the log transform for the smaller circulation and the polynomial model for larger circulations.

#4 Tonnages

1. From the output of the model, there is a big clump of points at the lower end, and then they start to spread out at higher values. I would assume that the variance is not constant for all values.
2. I can’t imagine any prediction for that amount of tonnage would be a good interval due to the variance growing bigger the higher your tonnage gets.

Second Model

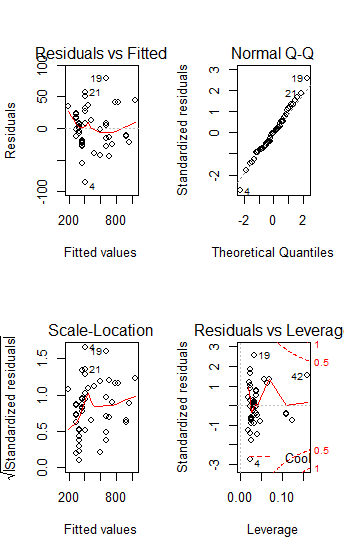
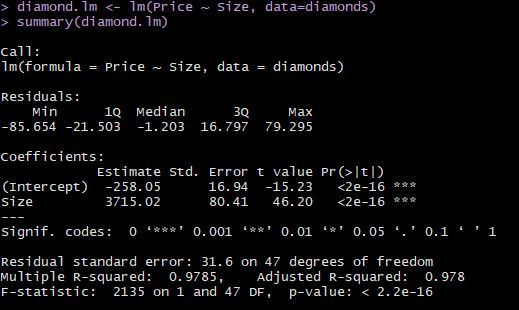
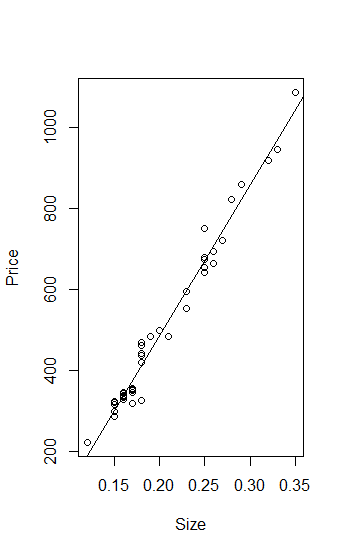
1. It looks like an improvement because the residual variance seems better. I would say that the prediction intervals for this model would be better than the first.
2. One issue is that the Rsquared value is lower in this model than in the first. And also, the residuals still don’t look as good as we would want.

#5 Cars

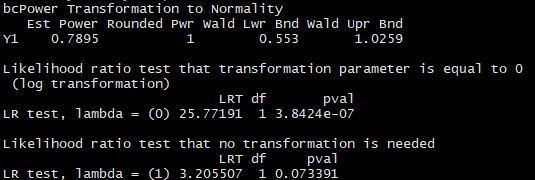
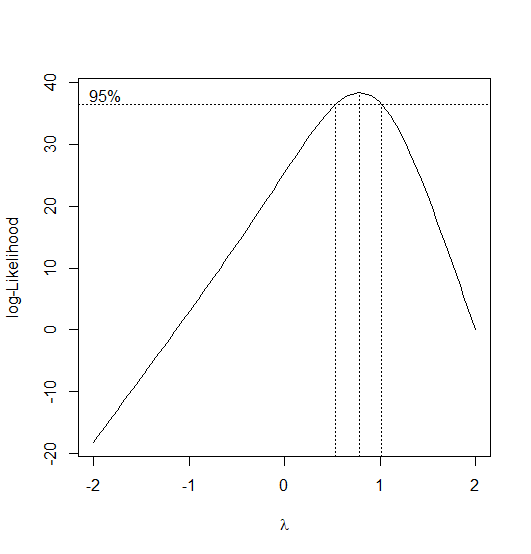
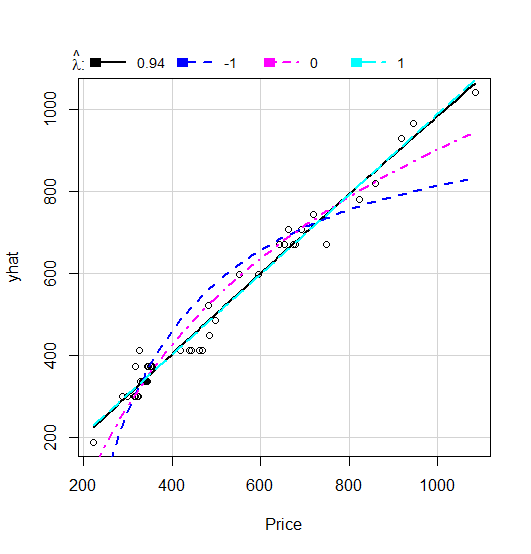
1. Based just on the summary of the lm call, that would be an accurate statement. However, looking at the residual plots you can tell that things start to fall apart for higher values of dealer cost.
2. First thing I would do is probably transform the data somehow, get those residuals back to random. There also looks to be a single really high point, this could cause a lot of leverage since it’s so far away from the other X’s.
3. The second model definitely looks like an improvement. The residuals look more random and the points are closer together on the X axis. Apart from the lower tail of the Q-Q plot, it looks better distributed.
4. It’s the % change of the dealer cost vs. the % change of the Suggested Price.
5. The errors are still not as random as you would like.

#8 Diamonds

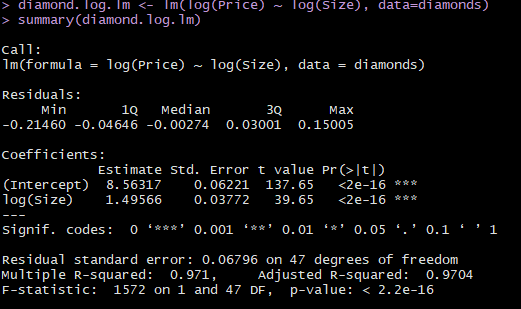
Part 1

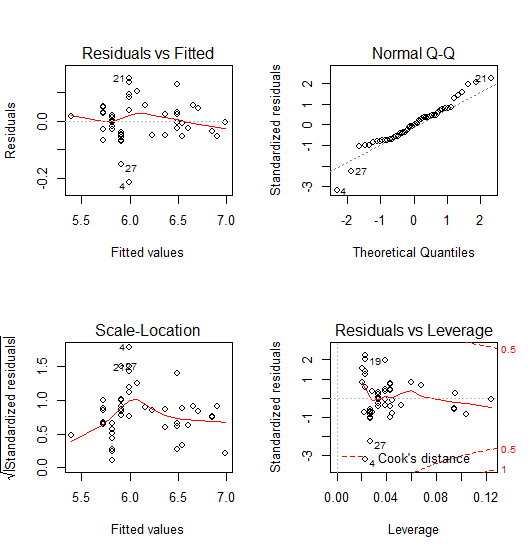


1. It told me to not transform anything
2. So I did the boxcox and powerTransform and both gave me a lambda value close to 1, so I don’t know if a transformation is needed. Images below.



I’ll try a log transform just for fun.





Part 3

Honestly, even after transform, it doesn’t look much different.