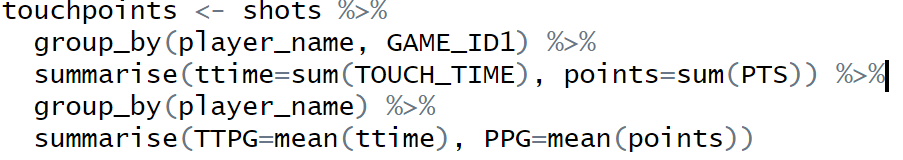
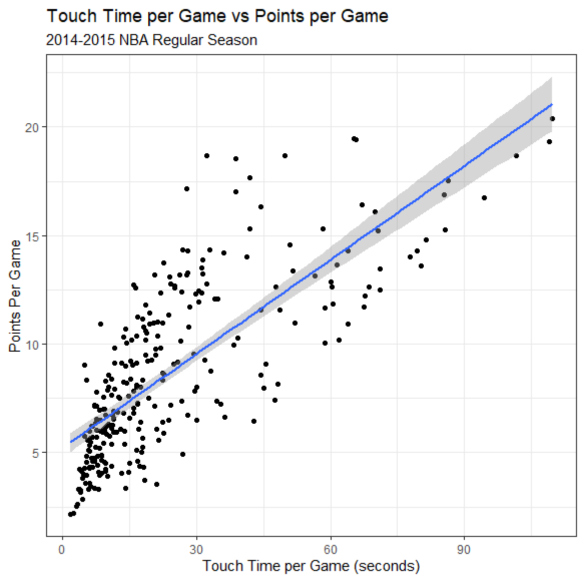
I have a dataset that includes data on every single shot taken in the 2014-2015 NBA regular season. This dataset has 128,069 observations. In other words, there were 128,069 shots taken in the 2014-2015 NBA regular season. Each observation (shot) contains 21 variables each possessing unique pieces of information. For this project, I wanted to test the following hypothesis: In the NBA, there is a direct relationship between the average amount of time per game a player touches the ball and the average number of points per game a player scores. To test this, I looked towards the variables titled “TOUCH\_TIME” and “PTS”, which describe the amount of time the player possessed the ball before taking that particular shot and the amount of points awarded as a result of that shot, respectively.

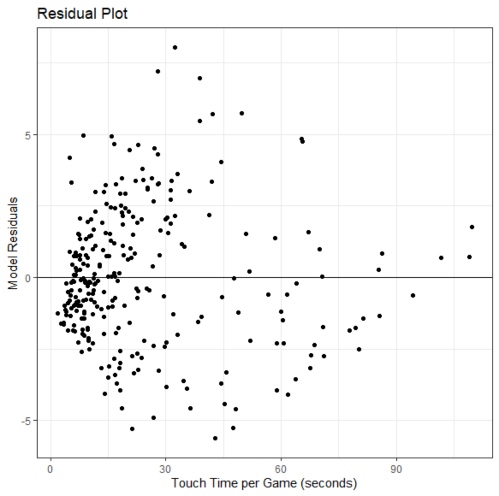
To prepare this data for use in a regression model, I had to filter all the observations and obtain unique observations that presented data about the average touch time per game per player as well as the average points per game per player. The R code I used to perform this data manipulation is below:



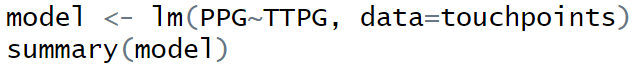
This is a plot showing the relationship between the average touch time per player per game and the average points per player per game. As it can be seen in this plot, there seems to be a general linear uptrend in this data. It can also be observed from this plot that we may have conflicts with constant variation in our regression model.

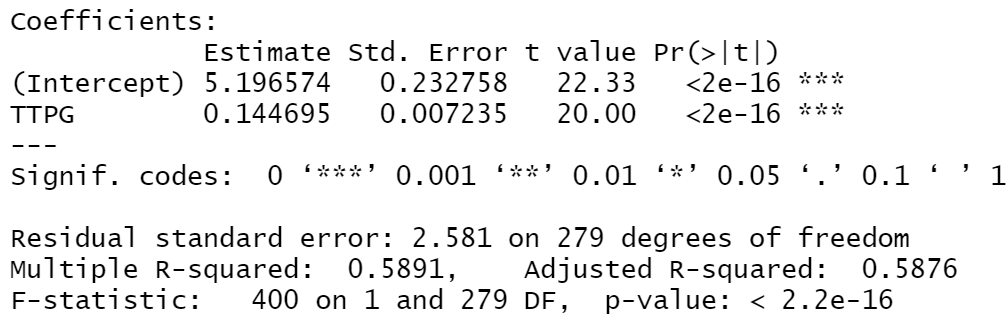


Let’s take a look at the residual plot to check assumptions:

Based on this residual plot, it can be seen very distinctly that there is a change in variance in our model. In general, the model variance seems to decrease as the touch time per game increases. Furthermore, this will need to be considered in the analysis because there were no transformations that was able to get rid of this change in variance in our model. We will proceed accordingly.

Here is the code and relevant output for this simple linear regression model:





From our model, we see that p < 0.5\*(2.2\*10-16) < 0.05. Therefore, we reject our null hypothesis that there is not linear relationship between these two variables in favor of our alternative hypothesis that there is a positive relationship between touch time per player per game and points scored per player per game. It should be noted though that this model clearly violates the constant variance assumption. So, although there is a linear association between these two variables, we do not have the capacity to predict with a certain level of confidence a certain points per game value given a touch time per game value.