Hursh Desai

Introduction

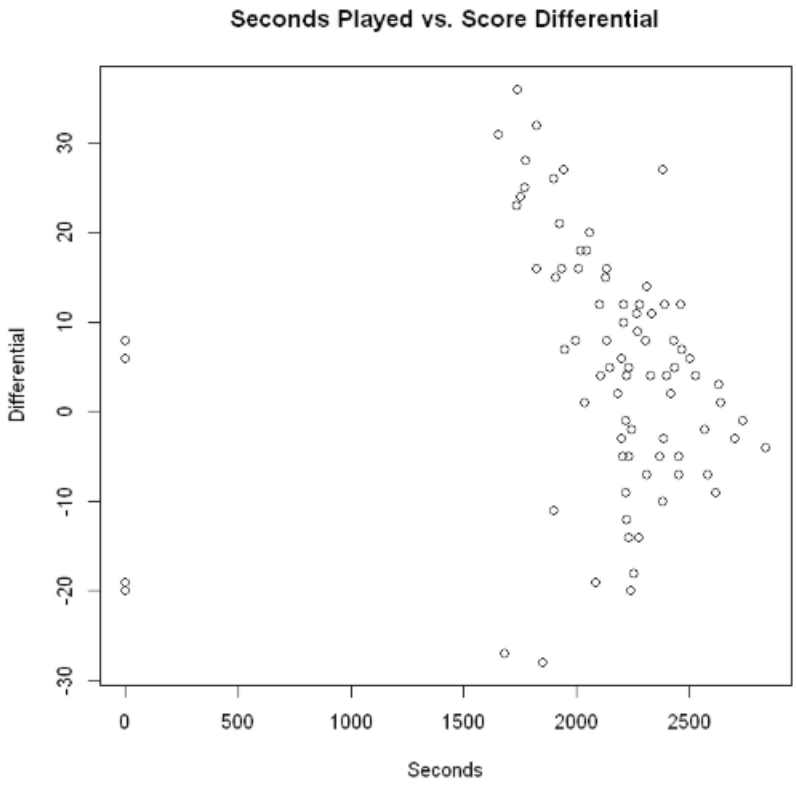
I was interested in seeing if James Harden was truly good for the Houston Rockets. A lot of people see him as the team's saving grace however others say that he makes the team worse whenever he is on the court. I wanted to compare two statistics that would best encapsulate all the things that player could do on the court to help the team win. I chose the minutes played by the player as the variable that would best express the player's ability to help the team. If his presence/skills were key in the Rockets winning, then the more minutes he plays in a game the more likely it is that the team should win. However, since we have to choose a numerical response variable, rather than whether the team lost or won, I chose the game's score differential (the amount of points the Rockets scored – the amount of points the opposing team scored) to show the team doing better as a higher score differential. In this way, since I actually think that James Harden is worse for the Rockets, I hypothesized that the more minutes James Harden played last season the lower the score differential of that game should be.

Gathering the Data

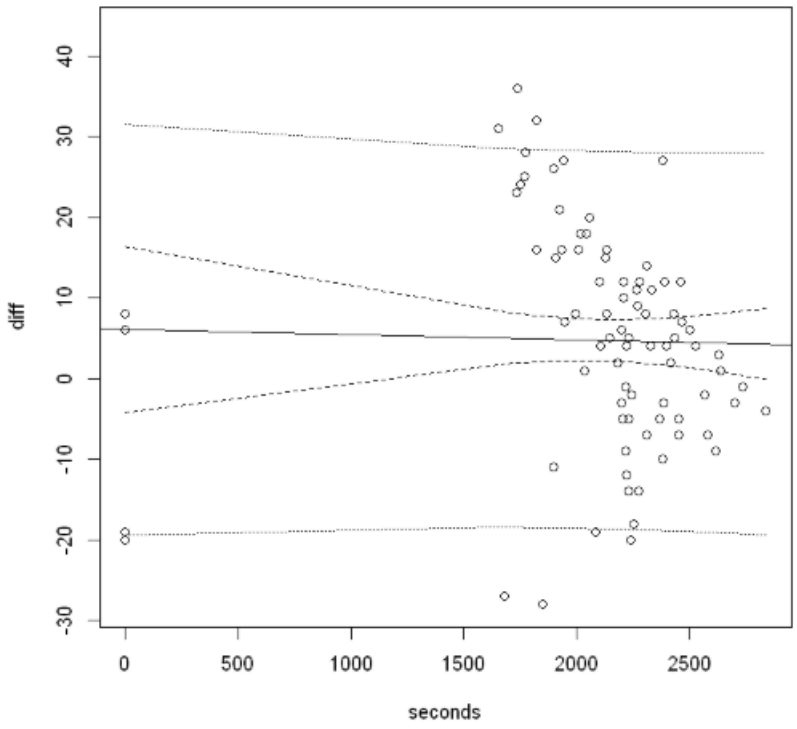
I found a website (basketball-reference.com) that housed all of core stats that I needed in order to do this analysis. I used the stats from the 2018-19 season primarily because since the season just started the Rockets haven't played enough games this season in order to have a big enough sample size. I scraped the amount of minutes Harden played for each game last season and converted them into seconds because I thought that would be the easiest to use for analysis. It turns out he did not play for 4 of the games last season so for those games I put 0 seconds. As for the score differential I used the final box score of each game to scrape the amount of points the Rockets scored and subtracted from that the amount of points the opposing team scored. This way if the Rockets did better the score differential should be higher and if the Rockets did worse the score differential would be lower and the times they lost their differential would be in the negatives.

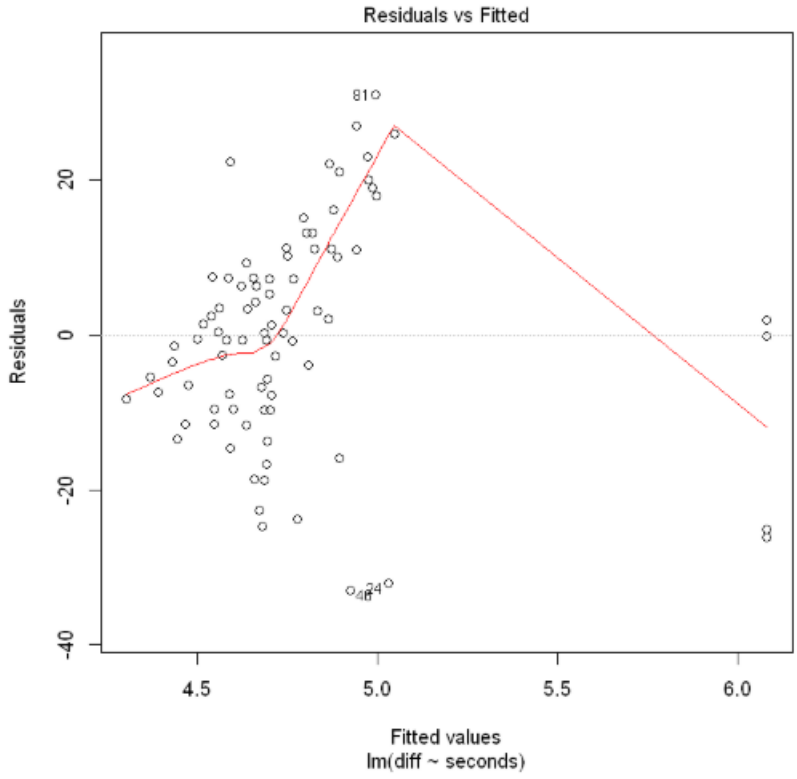
Data Analysis

I used R to do my data analysis and regression so that is where the images will be from.



Already looking at the scatter plot we can see a downwards trend in the data as the seconds amount of seconds he plays increases. The 4 games wherein he did not play also doesn't seem to tell us much since two of those times the team lost and twice they won both around the same amount of points lost and won by.


diff
seconds
40
30
20
10
0
-10
-20
-30
0
500
1000
1500
2000
2500


Residuals vs Fitted
Residuals
Fitted values
lm(diff ~ seconds)
81
46
24
20
0
-20
-40
4.5
5.0
5.5
6.0

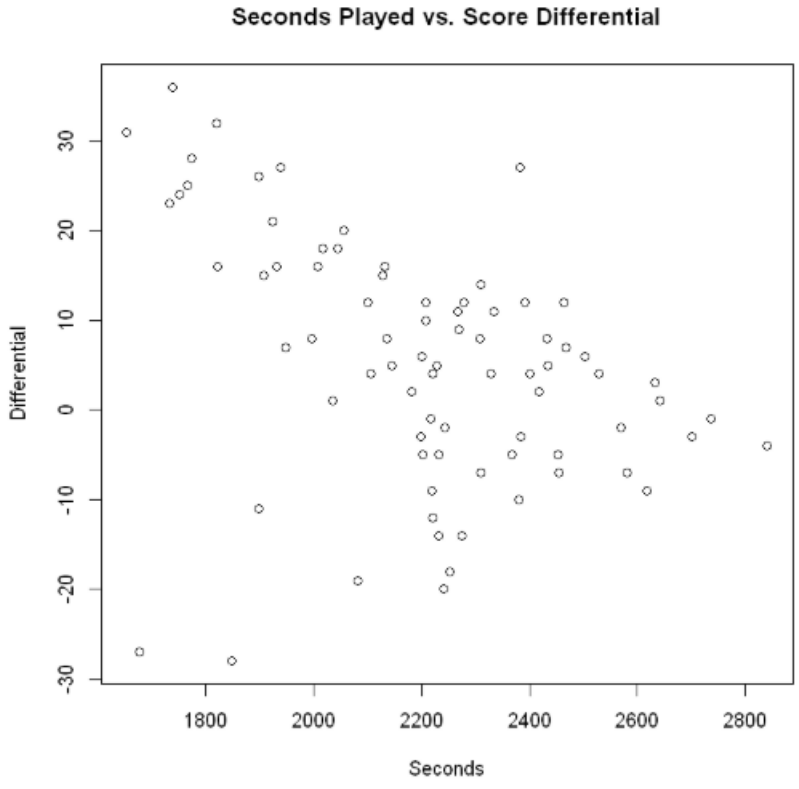
```
Call:
lm(formula = diff ~ seconds, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-32.924  -9.235   0.288   8.889  31.008

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.0806255  6.1845249   0.983    0.328
seconds     -0.0006256  0.0028550  -0.219    0.827

Residual standard error: 13.98 on 80 degrees of freedom
Multiple R-squared:  0.0005999, Adjusted R-squared:  -0.01189
F-statistic: 0.04802 on 1 and 80 DF,  p-value: 0.8271
```

As we can see from this fitted line plot the 4 games he missed is heavily skewing the data. And from the plot of the residuals vs. the fitted these four games are very unusual observations. Also from the regression analysis we can see these data points have severely affected the measure of fit like R2, t-statistics, and the F-statistic. A model that dropped these four data points will be much better at showing us if there is an underlying relationship between these two variables.



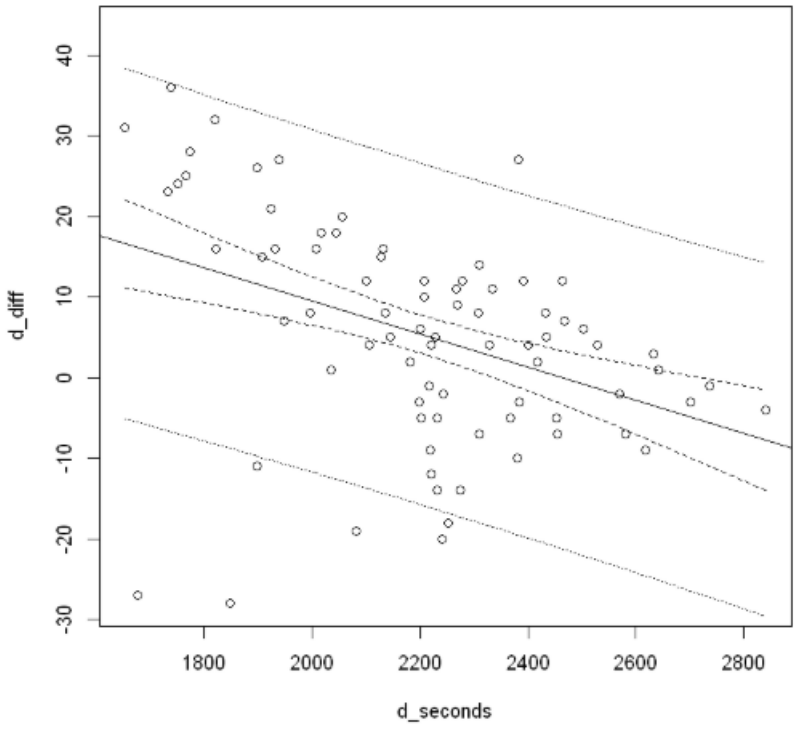
Taking out the 4 games that he didn't play in the downwards trend is even more apparent. It seems as the there are only about 4 outliers that show him playing a small amount of minutes and the team losing and 1 outlier where he played a large amount of minutes and that actually helped.

Confidence Interval

	fit	lwr	upr
1	7.83694007	4.69957407	10.9743061
2	4.00467983	1.07157137	6.9377883
3	1.64794224	-1.79548308	5.0913676

Prediction Interval

	fit	lwr	upr
1	7.83694007	-17.519894	33.19377
2	4.00467983	-21.327692	29.33705
3	1.64794224	-23.748576	27.04446



Fitting a line to this plot it is clear that there a negative correlation between minutes played and the score differential. As can be seen by the lines that indicate the prediction interval, the variance also seems to be very large which is why the lines are so much farther away from the

fitted line. It should also be noted that the confidence interval is much narrower in the middle of the graph and gets wider at the ends of the plot. This means that the prediction are less accurate the more extreme the predicting value gets. This is most probably due to the fact that there is many more data points clustered around the center than at the ends. The prediction interval is of course much wider than the confidence interval which is necessary because it must account for both the uncertainty in estimating the population mean, plus the random variation of the individual values.

```
Call:
lm(formula = d_diff ~ d_seconds, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-43.137  -6.021   3.196   8.062  25.291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.52463   11.99254   4.213 6.86e-05 ***
d_seconds   -0.02049    0.00540  -3.795 0.000294 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.63 on 76 degrees of freedom
Multiple R-squared:  0.1593,    Adjusted R-squared:  0.1483
F-statistic:  14.4 on 1 and 76 DF,  p-value: 0.0002944
```

```
[136]: cor(d_diff , d_seconds)
```

-0.399171333014705

```
[137]: vcov(paperlm)
```

A matrix: 2 × 2 of type dbl

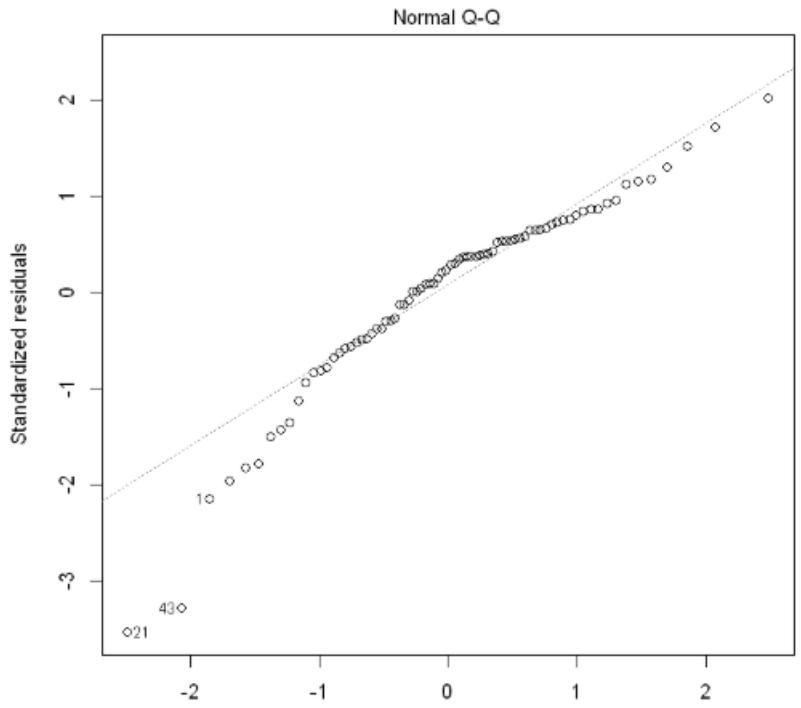
	(Intercept)	d_seconds
(Intercept)	143.82099991	-0.0642920772
d_seconds	-0.06429208	0.0000291552

```
[139]: (summary(paperlm)$sigma)
```

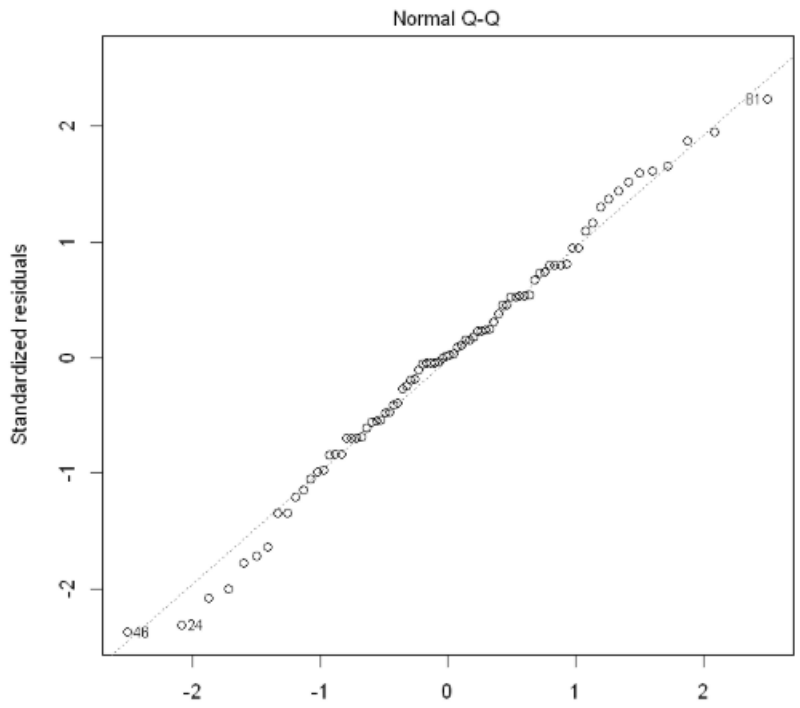
12.6335996148004

From the summary statistics we can see that because the p-value is very low and thus we can say that there is a relationship between the amount of minutes that James Harden played and the amount of points scored by the team. Looking at the correlation coefficient we can clearly see

that there is a somewhat strong negative correlation between these two variables. The coefficient of determination though between amount of time played by James Harden and score differential is quite weak, the Adjusted R-Squared is only 14.83%. This means that the amount of variation in score differential that can be explained by the amount of minutes that James Harden played is quite low. Interesting to note is that the residual standard error is 12.63 which is very far from the supposed assumed 0 that the expected value of all of the residuals is supposed to be. This means that there might be an underlying factor that I have not taken into account in this simple regression that is associated with the score differential. We can also see that this model is much more robust in the relationship between the two variables than the model that kept James Harden's missed games. The t-value here is also statistically significant at the Type I error levels. The association is much stronger and through comparing the coefficients we can see a tangible impact that the minutes he plays has on the score, now we can say that with every one second more that he plays in the game he is taking away .02 points from the final score.



The Q-Q plot shows us that most of the data points can be considered normally distributed. There are a few data points that waver away from the line which does mean that some of these data points do fall out of the normal distribution and so it could also be that a number of these observations are correlated with each other through some factor. My guesses for these unknown factors could be that when James Harden gets onto the court he brings the team chemistry down by through things like hogging the ball and so the team as whole scores less.



This is the Q-Q plot but with the games that James Harden missed which shows that for some reason the residuals are much more normally distributed when the games that he missed are added into the model.

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
d_seconds	1	2299.138	2299.1376	14.40492	0.0002943951
Residuals	76	12130.196	159.6078	NA	NA

Through this anova table we can see that our F value is significant, which means that our model is a good fit for the data. This compared to the anova table below shows that indeed those four data points of the games that James Harden missed made the previous model a very bad fit of the data.

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
seconds	1	9.384947	9.384947	0.04801954	0.8271041
Residuals	80	15635.212614	195.440158	NA	NA

Conclusion

There is a significant enough relationship between the amount of minutes that James Harden plays in a game and the end score differential of the game. However, it is important to note that the relationship is a somewhat weak one and the variability is high enough to make it so that making predictions using this model will not be accurate enough. It is also necessary to mention that missing games can ruin the fit of your model a lot, so in order to get a better fit it is best to drop any games that a player missed.