Mathematics: Application and Interpretation

Higher Level

Internal Assessment

Modeling the relationship between scoring Field Goals and Games Won in Basketball through Regression Models.

Page Count: 19, 22(including cover page, table of contents and bibliography)

Contents

Introduction

Growing up I was always involved in sports; not just playing different sports but also watching the matches and being well versed with the statistics of the sport. From knowing what each team scores in a game to the number of games they have won, and what chances they have at winning the trophy at the end of every season; knowing all the statistics and figures of the sport made watching sports more interesting to me. Throughout my life, I have continued to watch various new sports and one sport that piqued my interest was basketball.

evidence of some personal engagement.

Over the past two years of being in lockdowns due to the COVID-19 restrictions, I discovered basketball and, in particular, the National Basketball Association, the most recognized professional basketball league (NBA). This sport grabbed my attention as it displayed a really high level of athleticism along with a wide array of skills displayed which made the sport fun and entertaining to watch.

In the sport of basketball, scoring points occur mainly in three different ways, free throws, field goals, and three-pointers. The intensity of the sport has risen so much over the years that the most anticipated way of scoring points are three-pointers and they are often associated with winning games. But, I believe that field goals, the most fundamental method of scoring in basketball is often overlooked and isn't considered a 'game-winning' method of scoring points when they are the type of points that a team scores the most. Due to this, I wanted to see whether scoring field goals are statistically represented in a team winning games. This led me to my investigation as to *whether there is a relationship between scoring field goals and games won.*

Aim and rationale seen

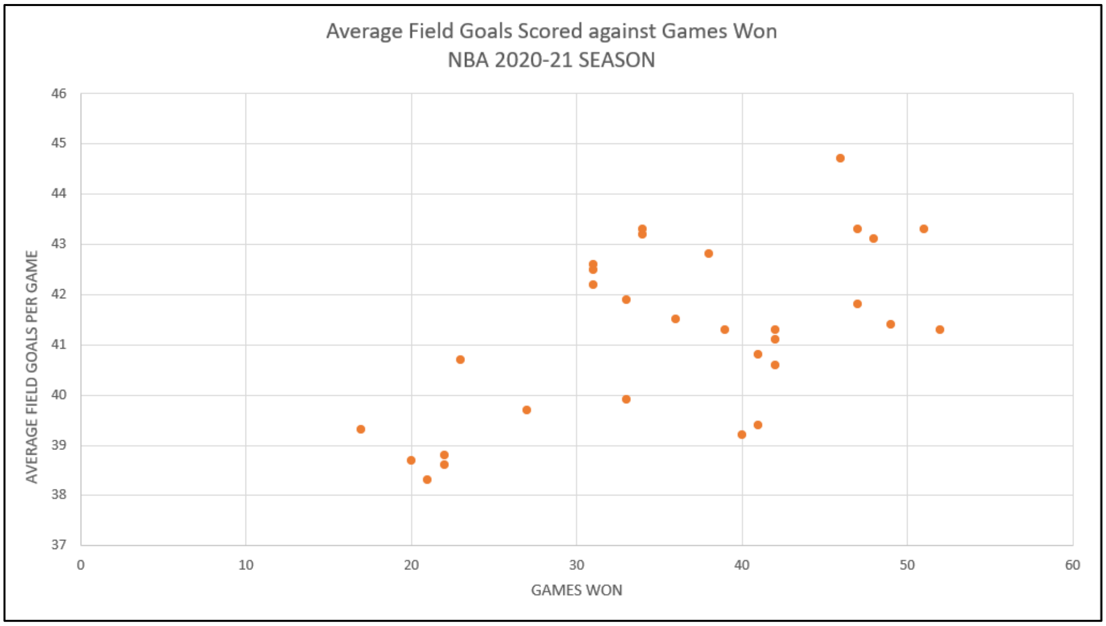
Collection of Data Used

Teams	***Games Won (x)***	***Average Field Goals-per game (y)***
Atlanta Hawks	41	40.8
Boston Celtics	36	41.5
Brooklyn Nets	48	43.1
Charlotte Hornets	33	39.9
Chicago Bulls	31	42.2
Cleveland Cavaliers	22	38.6
Dallas Mavericks	42	41.1
Denver Nuggets	47	43.3
Detroit Pistons	20	38.7
Golden State Warriors	39	41.3
Houston Rockets	17	39.3
Indiana Pacers	34	43.3
Los Angeles Clippers	47	41.8
Los Angeles Lakers	42	40.6
Memphis Grizzlies	38	42.8
Miami Heat	40	39.2
Milwaukee Bucks	46	44.7
Minnesota Timberwolves	23	40.7
New Orleans Pelicans	31	42.5
New York Knicks	41	39.4
Oklahoma City Thunder	22	38.8
Orlando Magic	21	38.3
Philadelphia 76ers	49	41.4
Phoenix Suns	51	43.3
Portland Trail Blazers	42	41.3
Sacramento Kings	31	42.6
San Antonio Spurs	33	41.9
Toronto Raptors	27	39.7
Utah Jazz	52	41.3
Washington Wizards	34	43.2

Table 1 – Number of Games Won and average field goals per game of the NBA season 2020/21[1]

[1] https://www.nba.com/stats/teams/traditional/?sort=TEAM_NAME&dir=-1&Season=2020-21&SeasonType=Regular%20Season

To visualize the relationship between the games won and the field goals scored per game that is displayed in Table 1 is plotted and displayed in Graph 1 below.



Graph 1 – Visual representation of the dataset

Aim of Investigation

To test the strength of correlation and model the variables, we will perform two main types of regressions on the dataset; linear regression and polynomial (quadratic regression). Through the use of linear regression and polynomial regression, this investigation aims to achieve three main things which are to:

1. Find the strength of the correlation between field goals scored and games won.
2. Model the relationship between field goals scored and games won.
3. Find out which model out of the linear and polynomial regression is more accurate in modeling the relationship.

Linear Regression Model

Linear regression is an approach used to model the relationship between two variables by fitting a linear equation to the observed data.[2] One variable is referred to as a dependent variable, while the other is referred to as an independent variable. A linear regression line has the following equation:

$$\boldsymbol{y = a + bx}$$

In the equation:

Presentation is approriate

- ***y*** - is the dependent variable; average field goals made per game.
- ***x*** - is the independent variable; games won by a team.
- ***a*** - y-intercept.
- ***b*** - slope of line.

The general equation that is used to calculate linear regression is[3] -

$$a = \frac{(\Sigma \mathrm{y})(\Sigma x^2) - (\Sigma \mathrm{x})(\Sigma xy)}{n(\Sigma x^2) - (\Sigma \mathrm{x})^2}$$

$$b = \frac{n(\Sigma \mathrm{xy}) - (\Sigma \mathrm{x})(\Sigma y)}{n(\Sigma x^2) - (\Sigma \mathrm{x})^2}$$

By using the dataset in Table 1, we can get all the data that is needed to calculate *'a'* and *'b'*.

[2] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
[3] https://www.andrews.edu/~calkins/math/edrm611/edrm06.htm

Variable needed for general equation	Value	What does it mean?
n	30	Total number of samples (Number of Teams)
(Σx)	1080	Sum of all x-values (All Games Won)
(Σx^2)	41788	Sum of all x^2-values
(Σy)	1236.6	Sum of all y-values (All Field Goals Scored)
(Σy^2)	51056.7	Sum of all y^2-values
(Σxy)	44806.4	Sum of x × y values

Table 2 – Variables calculated for general equation

Now that we have all the data that is required to calculate *'a'* and *'b'* we can input them into the equation as shown below.

$$\boldsymbol{a} = \frac{(1236.6)(41788) - (1080)(44806.4)}{30(41788) - (1080)^2} = 37.64475928 \approx 37.6448$$

$$\boldsymbol{b} = \frac{30(44806.4) - (1080)(1236.6)}{30(41788) - (1080)^2} = 0.09931224209 \approx 0.0993$$

From this we can form the linear regression equation which is:

$$y_L = 37.6448 + 0.0993\mathbf{x}$$

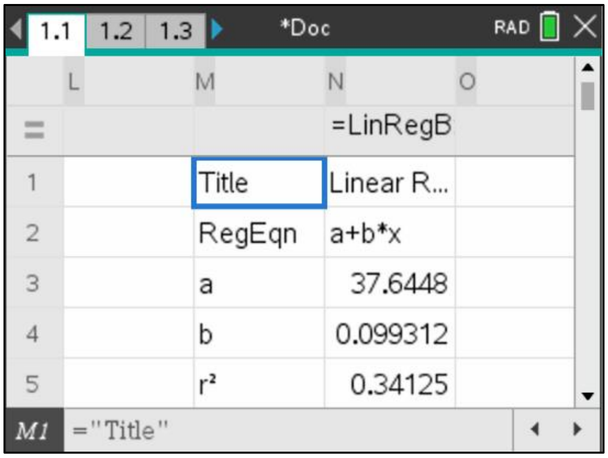
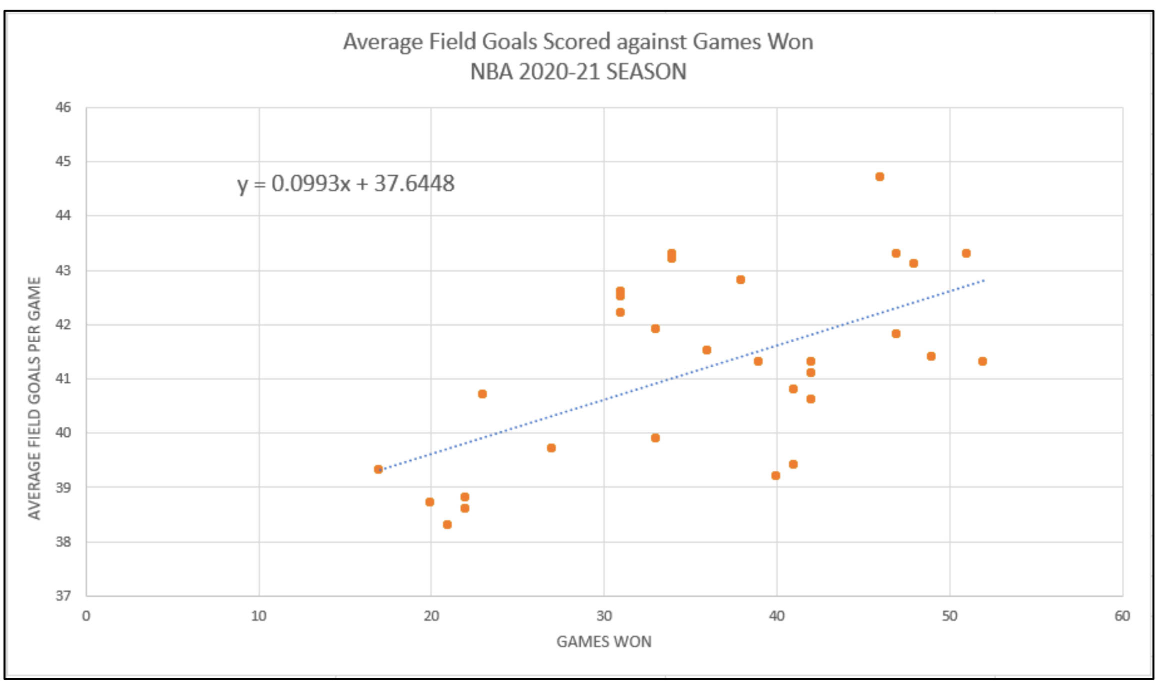


Figure 1 – Results of Linear Regression calculation from GDC

To cross-check and verify the values that have been calculated, I made use of a Graphic Design Calculator (GDC), TI-Nspire CX II to which I got similar values as shown above. Now we plot the linear regression line we have deduced onto the graph.



Graph 2 – Linear Regression line plotted for the average field goals per game against games won

From the graph and the line plotted, we can see that the overall trend of the linear regression calculated is an increasing slope showcasing the relationship between games won and field goals scored, which indicates a positive correlation.

Pearson's Correlation Coefficient

To put the correlation to the test further, we can use Pearson's Correlation Coefficient, which can help back up the claim of a strong correlation with further evidence. The Pearson product-moment correlation coefficient, denoted by r, is a measure of the strength of a linear relationship between two variables.

The coefficient has a value between -1 and +1. A value of 0 implies that no relationship exists between the two variables. A value greater than 0 shows a positive correlation, which suggests that when the value of one variable rises, so will the value of the other. A value less than 0 implies a negative relationship, which means that when the value of one variable rises, the value of the other variable falls.[4]

Relevant mathematics commensurate with the level of the course is used. The mathematics explored is partially correct. Some knowledge and understanding are demonstrated

The formula that will help us calculate this coefficient is[5]:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{(n(\Sigma x^2) - (\Sigma x)^2)(n(\Sigma y^2 - (\Sigma y)^2)}}$$

$$r = \frac{30(44806.4) - (1080)(1236.6)}{\sqrt{(n(41788) - (1080)^2)(30(51056.7) - (1236.6)^2)}} = 0.5841660266 \approx 0.584$$

[4] https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php
[5] https://www.socscistatistics.com/tests/pearson/

To check the accuracy of the equation in calculating the Pearson's Correlation Coefficient, I also calculated it through the use of the GDC and got similar results.

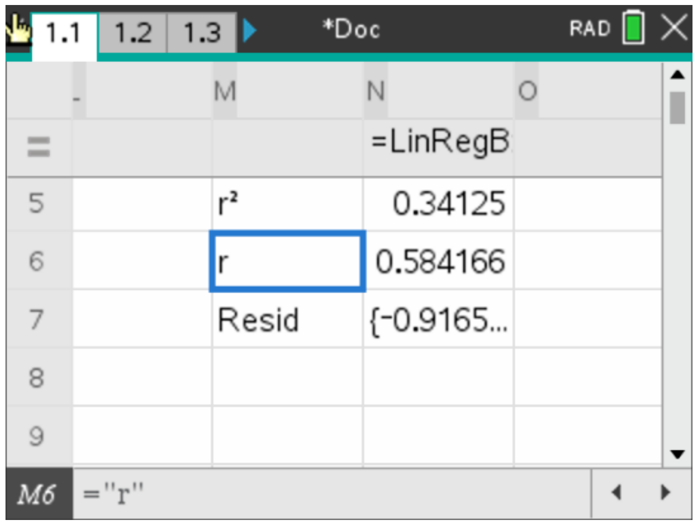


Figure 2 – Results of the Coefficient that was calculated on the GDC

From the calculations made, we were able to get a Pearson's correlation coefficient of 0.584, which is a moderately positive correlation between average field goals scored and games won. Although this is evidence for a strong correlation, we need conclusive evidence that the coefficient is significant enough to represent both the variables, average field goals scored and games won and that they are statistically linked. To aid us in confirming this, we need to calculate the critical correlation coefficient (r_e) which will give conclusive evidence on the relation between the two variables.

Critical Correlation Coefficient (r_e)

For calculating the coefficient, the null hypothesis and alternative hypothesis is stated below:

The null hypothesis (H_0): states that there is no relationship between winning games and scoring field goals, hence the correlation will be 0.

The alternative hypothesis (H_1): states that there is a positive relationship between games won and field goals scored, implying that the more field goals scored each game, the more likely it is that a game will be won.

Now that we have formulated the null and alternative hypotheses, we will be able to find the coefficient. For this, we will be using a One-Tailed Test because we want to determine the correlation in a positive direction. To calculate the degrees of freedom (df), we must subtract the sample size by 2.

$$df = (n - 2) \equiv 30 - 2 = 28$$

For a One-Tailed Test with *'df'* of 28 and a significance level of 0.05 the coefficient is:

$$r_e = 0.306$$

Now that we have the two coefficients, r and r_e

$$r > r_e \equiv 0.584 > 0.306$$

We can see that the Pearson's Correlation Coefficient is **greater** than the critical correlation coefficient. This means that we have enough evidence to reject H_o and accept H_1. From this, we can conclude that statistically, there is a proper linear relationship between the average field goals scored per game and the games won. Now that we have modeled the relationship through the linear regression model, we can proceed to do the same with polynomial regression.

Polynomial (Quadratic) Regression Model

Polynomial Regression is another form of regression analysis that also helps model the relationship between an independent and dependent variable but modeled in the n^{th} degree polynomial. Now using the GDC we can calculate the equation of the polynomial curve and also plot the curve.

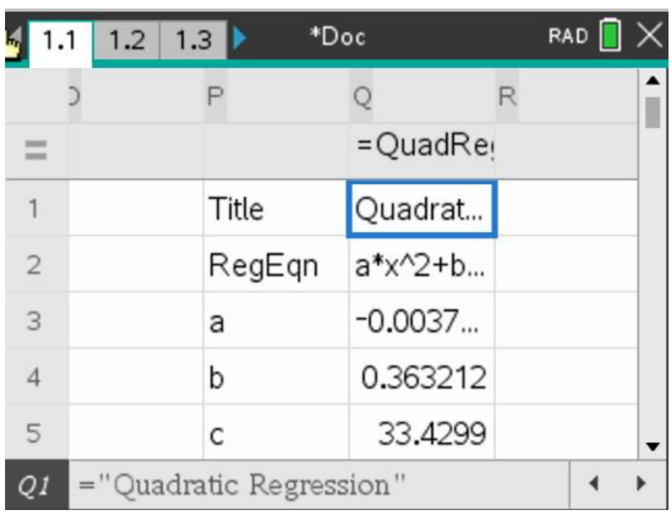
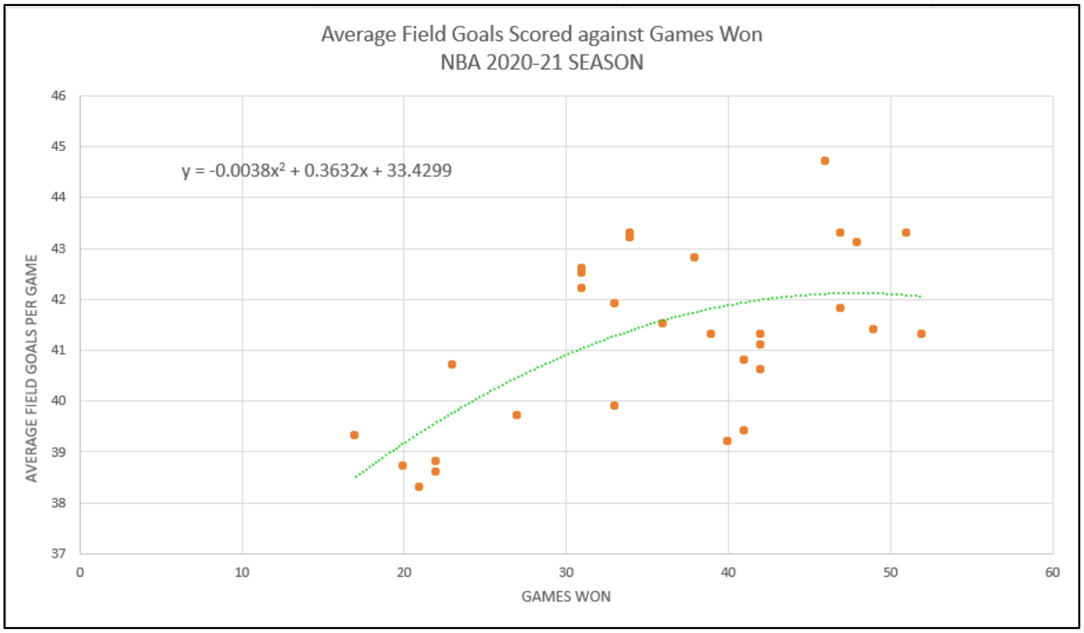


Figure 3 – Results of Polynomial Regression calculation from GDC

From this, we can form the polynomial equation of $y = ax^2 + bx + c$:

$$y_P = -0.0038\mathbf{x}^2 + 0.3632\mathbf{x} + 33.4299$$

Relevant mathematics commensurate with the level of the course is used. The mathematics explored is correct. Thorough knowledge and understanding are demonstrated



Graph 3 – Polynomial Regression curve plotted for the average field goals per game against games won

Calculating Residual Sum of Squares

The residual sum of squares measures the variation in modeling errors. As we have two regression models, linear and polynomial, a residual sum of squares will help us understand which of these two models is the best fit for the dataset used and determines the strength of the relationship between the variables, average field goals made per game and games won. In the residual sum of squares, the smaller the value is, the better the model fits the data, and the bigger the value, the worse the model fits the data. The equation used to calculate the residual sum of squares is[6]:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

In the equation 'n' is the sample size, 'y_i' is the observed values of the dependent variable and '$f(x_i)'$ is the estimated or expected value.

[6] Economopoulos, P., Halsey, T., Doering, S., Ortman, M., Singh, S. N., Gray, P., Harris, D., & Wathall, C. J. (2019). *Oxford IB Diploma Programme IB Mathematics: applications and interpretation, Higher Level, Print and Enhanced Online Course Book Pack*. Oxford University Press.

RSS for Linear Regression

$$y_L = 37.6448 + 0.0993\mathbf{x}$$

Observed (y_i)	Expected ($f(x_i)$)	$y_i - f(x_i)$ – **used later for Paired t-Test**	Residual Square $(y_i - f(x_i))^2$
40.8	41.7161	-0.9161	0.83923921
41.5	41.2196	0.2804	0.07862416
43.1	42.4112	0.6888	0.47444544
39.9	40.9217	-1.0217	1.04387089
42.2	40.7231	1.4769	2.18123361
38.6	39.8294	-1.2294	1.51142436
41.1	41.8154	-0.7154	0.51179716
43.3	42.3119	0.9881	0.97634161
38.7	39.6308	-0.9308	0.86638864
41.3	41.5175	-0.2175	0.04730625
39.3	39.3329	-0.0329	0.00108241
43.3	41.021	2.279	5.193841
41.8	42.3119	-0.5119	0.26204161
40.6	41.8154	-1.2154	1.47719716
42.8	41.4182	1.3818	1.90937124
39.2	41.6168	-2.4168	5.84092224
44.7	42.2126	2.4874	6.18715876
40.7	39.9287	0.7713	0.59490369
42.5	40.7231	1.7769	3.15737361
39.4	41.7161	-2.3161	5.36431921
38.8	39.8294	-1.0294	1.05966436
38.3	39.7301	-1.4301	2.04518601
41.4	42.5105	-1.1105	1.23321025
43.3	42.7091	0.5909	0.34916281
41.3	41.8154	-0.5154	0.26563716
42.6	40.7231	1.8769	3.52275361
41.9	40.9217	0.9783	0.95707089
39.7	40.3259	-0.6259	0.39175081
41.3	42.8084	-1.5084	2.27527056
43.2	41.021	2.179	4.748041

Table 3 – Data Collection for RSS of Linear Regression

To calculate the RSS of linear regression, we need to perform the one-variable statistics of the 'Residual Square' data in the GDC and from that, we can find the value.

$$\text{RSS} = \sum_{i=1}^{30} (y_i - f(x_i))^2 = 55.36662972 \approx \mathbf{55.367}$$

RSS for Polynomial Regression

$$y_P = -0.0038\mathbf{x}^2 + 0.3632\mathbf{x} + 33.4299$$

Observed (y_i)	Expected ($f(x_i)$)	$y_i - f(x_i)$ – **used later for Paired t-Test**	Residual Square $(y_i - f(x_i))^2$
40.8	41.9333	-1.1333	1.28436889
41.5	41.5803	-0.0803	0.00644809
43.1	42.1083	0.9917	0.98346889
39.9	41.2773	-1.3773	1.89695529
42.2	41.0373	1.1627	1.35187129
38.6	39.5811	-0.9811	0.96255721
41.1	41.9811	-0.8811	0.77633721
43.3	42.1061	1.1939	1.42539721
38.7	39.1739	-0.4739	0.22458121
41.3	41.8149	-0.5149	0.26512201
39.3	38.5061	0.7939	0.63027721
43.3	41.3859	1.9141	3.66377881
41.8	42.1061	-0.3061	0.09369721
40.6	41.9811	-1.3811	1.90743721
42.8	41.7443	1.0557	1.11450249
39.2	41.8779	-2.6779	7.17114841
44.7	42.0963	2.6037	6.77925369
40.7	39.7733	0.9267	0.85877289
42.5	41.0373	1.4627	2.13949129
39.4	41.9333	-2.5333	6.41760889
38.8	39.5811	-0.7811	0.61011721
38.3	39.3813	-1.0813	1.16920969
41.4	42.1029	-0.7029	0.49406841
43.3	42.0693	1.2307	1.51462249
41.3	41.9811	-0.6811	0.46389721
42.6	41.0373	1.5627	2.44203129
41.9	41.2773	0.6227	0.38775529
39.7	40.4661	-0.7661	0.58690921
41.3	42.0411	-0.7411	0.54922921
43.2	41.3859	1.8141	3.29095881

Table 4 – Data Collection for RSS of Polynomial Regression

Likewise, for the polynomial regression, we can perform the one-variable statistics on the GDC and find the RSS.

$$\text{RSS} = \sum_{i=1}^{30} (y_i - f(x_i))^2 = 51.46187422 \approx \mathbf{51.462}$$

From the calculations of the two models, we find that the Residual Sum of Squares of the linear regression model is **55.367** and for the polynomial regression model it is **51.462**. Through this, we can see that the polynomial regression is clearly a lower value than the linear regression. This means that the **polynomial model is a better model** that fits this dataset and determines the strength of the relationship between games won and average field goals scored per game.

There is evidence of meaningful reflection

Hypothesis Testing

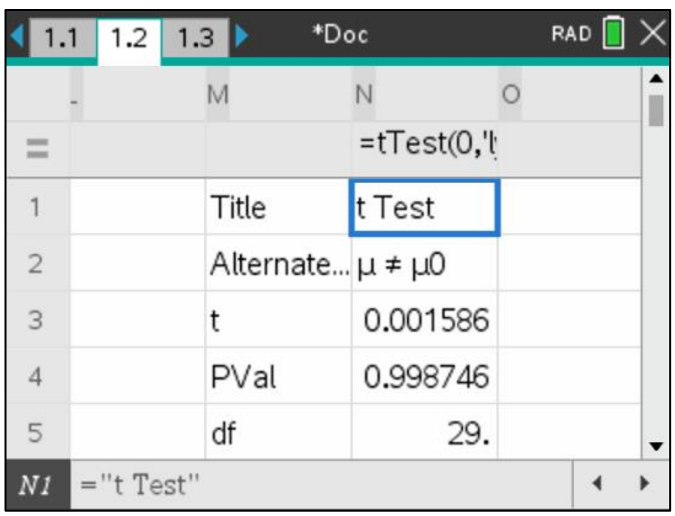
Now that we have found from the investigation that the polynomial model better fits the relationship between the average field goals scored per game and games won, we can use this opportunity to further test the hypothesis through the use of the Paired t-Test. The paired t-test is a method for determining whether or not the mean difference between two measurements (observed and expected) is zero. For each regression model, we will compare the differences between the average number of field goals scored per game and the expected number of field goals to be scored per game. This test will provide us t-statistics and a p-value, which we will use to evaluate whether or not to accept the proposed null hypothesis. A t-statistic is used to determine whether there is a significant difference in the means of two related measurements, observed and expected. The P-value is the possibility of a hypothesis test providing results that are the closest to the observed data, provided that the null hypothesis is valid. However, if the calculated P-value is less than 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence to adopt the alternative hypothesis.

Now to get started with the paired *t*-test, we need to state the null and alternative hypotheses.

Null hypothesis (H_0): There is no difference between the means of the observed average field goals scored per game and the expected number of field goals to be scored per game. This is denoted by $\boldsymbol{H_0 = 0}$.

Alternative hypothesis (H_1): There is a difference between the means of the observed average field goals scored per game and the expected number of field goals to be scored per game. This is denoted by $\boldsymbol{H_1 \neq 0}$.

The test will be done at a significance level of 0.05 and will be a two-tailed test because the mean differences could go both in the positive and negative direction.


1.1 1.2 1.3 *Doc RAD

		M	N	O
=			=tTest(0,'l	
1		Title	t Test	
2		Alternate...	µ ≠ µ0	
3		t	0.001586	
4		PVal	0.998746	
5		df	29.	

N1 ="t Test"

Figure 4 – Paired t-Test Results for Linear Regression

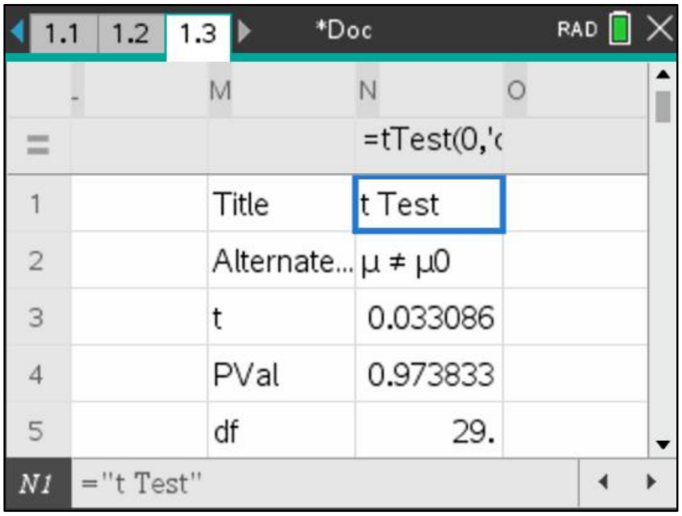


Figure 5 - Paired t-Test Results for Polynomial Regression

	T - Statistics	P - Value
Linear Regression	0.0016	0.9987
Polynomial Regression	0.0331	0.9738

Table 5 – Complied statistics of the Paired t-Test Results

As we can see from the results calculated from the Paired t-Test, the P-value of both the models is much greater than the significance level of 0.05. This means that regardless of the difference between the means of the models are not significant and the null hypothesis still holds. Now we have conclusive evidence to determine that there is a strong relationship between the average field goals scored per game in determining the games won. We can also conclude that the polynomial regression model is the best fit for modeling this due to the results we got from the RSS values (Although the P-value of linear regression is higher, the RSS score provides a higher degree of information to explain the model).

There is substantial evidence of critical reflection.

Discussion and Evaluation

In the NBA, the teams aim to win as many games as possible to qualify for the playoffs and further proceed to the finals. Through this investigation, we were able to model the relationship between field goals and games won through linear and polynomial models and the polynomial model is the best fit. This is something that NBA teams can utilize and better use models to help set targets for their performance in games. From this, we can see that focusing on scoring more field goals might give teams a higher chance of winning a game.

One last test to further check the accuracy of the polynomial model is to calculate the percentage error of the model in predicting the number of field goals that are needed per match to win a certain number of matches. For this percentage error, I will use three examples from the database which are of the Utah Jazz, Boston Celtics, and Houston Rockets from Table 1. This was done because I wanted to average the results of the model with two extreme data and one average data:

- Houston Rockets with the least number of wins (17), placing 30th of the 30 teams.
- Boston Celtics with an average number of wins (36), placing 16th of the 30 teams.
- Utah Jazz with the most number of wins in the season (52), placing 1st of the 30 teams.

Teams	***Games Won (x)***	***Average Field Goals-per game (y)***
Utah Jazz	52	41.3
Boston Celtics	36	41.5
Houston Rockets	17	39.3

Table 6 – Statistics of the three chosen teams displayed from Table 1

Through the use of the polynomial model, we can predict the average number of field goals that need to be scored per match to win the respective games each team has won in one season.

$$y_P = -0.0038\mathbf{x}^2 + 0.3632\mathbf{x} + 33.4299$$

Utah Jazz

$$yP = (-0.0038)(52)^2 + (0.3632)(52) + 33.4299 = 42.0411 \approx 42.04$$

$$\varepsilon \to \left|\frac{42.04 - 41.3}{41.3}\right| \times 100 = 1.7918 \approx 1.79\%$$

Boston Celtics

$$yP = (-0.0038)(36)^2 + (0.3632)(36) + 33.4299 = 41.5803 \approx 41.58$$

$$\varepsilon \to \left|\frac{41.58 - 41.5}{41.5}\right| \times 100 = 0.1928 \approx 0.19\%$$

Houston Rockets

$$yP = (-0.0038)(17)^2 + (0.3632)(17) + 33.4299 = 38.5061 \approx 38.51$$

$$\varepsilon \to \left|\frac{38.51 - 39.3}{39.3}\right| \times 100 = 2.0102 \approx 2.01\%$$

Now we can average the percentage errors:

$$\frac{1.79 + 0.19 + 2.01}{3} = 1.33 \approx 1.3\%$$

From this, we can observe that the percentage error between our polynomial model's prediction and the observed number of fields goals scored per game for a number of matches to be won is very small. This calculation gives conclusive evidence about the model and its ability to predict how many field goals need to be scored per match to win a certain number of games in an NBA season. Through the polynomial model, we can also detect a clear relationship between the number of games won and the number of field goals scored every match, which is not only the best fit model for this relationship but it has also proved a very low percentage error of 1.3% in predictive qualities. But we also get need to realize that the predictions may not be completely accurate due to various other factors. As mentioned before there are other factors of scoring in basketball, such as free-throw points and three-pointers that also factor in a game being won. Apart from the points scored, factors such as a team's morale, other team's performance, home or away court advantage, etc can heavily influence games being won in NBA. Therefore, saying that field goals alone result in winning games is an exaggeration and oversimplified.

There is substantial evidence of critical reflection.

Conclusion

In conclusion, we can determine with conclusive evidence that there is a significantly strong correlation between field goals scored and the number of games a team wins in a season; the greater the number of field goals scored in a match will correspond to a greater number of games being won in a season. Through the use of linear and polynomial models, we were able to model the relationship between the two variables. Additionally, we were able to find that the best regression to model this relationship was found to be the polynomial regression model which was highly accurate for this relationship. Personally, after performing this investigation I am able to understand more about the areas to focus on when playing basketball and that I need to focus on scoring more field goals. After investigating this topic extensively, I come to believe that scoring more field goals in a game, plays an important and defining role in maximizing a team's chance to win more games.

There is evidence of significant personal engagement.

Bibliography

Economopoulos, P., Halsey, T., Doering, S., Ortman, M., Singh, S. N., Gray, P., Harris, D., & Wathall, C. J. (2019). *Oxford IB Diploma Programme IB Mathematics: applications and interpretation, Higher Level, Print and Enhanced Online Course Book Pack*. Oxford University Press.

A. (2021, December 15). *Understanding Polynomial Regression!!! - Analytics Vidhya*. Medium. https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18

Corporate Finance Institute. (2020, May 15). *Sum of Squares*. https://corporatefinanceinstitute.com/resources/knowledge/other/sum-of-squares/

Kibret, H. G. (2019). https://iasks.org/articles/jttm-v01-i2-pp-27-34.pdf. *International Journal of Traffic and Transportation Management*, *01*(2). https://doi.org/10.5383/jttm.01.02.004

Morgan, J. A., & Tatar, J. F. (1972). Calculation of the Residual Sum of Squares for all Possible Regressions. *Technometrics*, *14*(2), 317–325. https://doi.org/10.1080/00401706.1972.10488918

Paired t-Test. (n.d.). Introduction to Statistics | JMP. https://www.jmp.com/en_sg/statistics-knowledge-portal/t-test/paired-t-test.html#:%7E:text=The%20paired%20t%2Dtest%20is,measurements%20is%20zero%20or%20not.

Statistics Solutions. (2021, June 18). *Table of Critical Values: Pearson Correlation*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/table-of-critical-values-pearson-correlation/

Teams Traditional. (n.d.). NBA Stats. https://www.nba.com/stats/teams/traditional/?sort=TEAM_NAME&dir=-1&Season=2020-21&SeasonType=Regular%20Season