# MAT 243 Project Three Summary Report

Nicholas Kreuziger

nicholas.kreuziger@snhu.edu

Southern New Hampshire University

## 1. Introduction

The data being explored is the FiveThirtyEight NBA Elo Dataset hosted by Kaggle. This is a historical dataset of NBA League performance. This data will be used to explore a series of inquiries about the performance of NBA teams. The analyses being performed will be Correlation, Simple Linear Regression and Multiple Linear Regression. These analyses all explore the relationship between different facets of the dataset.

## 2. Data Preparation

The data element avg\_pts\_differential refers to the average difference in points between your team and the opponents team. If your team scores 89 points and the opposing team were to score 70 points, the differential would be 19. If your team scores 60 points and the opposing team scores 70 points, the differential would be -10. The average points differential for a year of your team’s performance indicates on average what the difference in score was between your team and the opposing team.

The data element avg\_elo\_n refers to the average relative skill of your team in a regular season. An Elo Score is a rating that judges the likelihood of a team winning against another team. The average of this score is, on average, what the relative skill was over that period.

## 3. Scatterplot and Correlation for the Total Number of Wins and Average Points Scored

Data visualization techniques are used to study relationship trends between two variables by allowing easier digestion and interpretation of data trends. “Seeing is believing” is a moniker that resonates with visualizations applied to data. Additionally, some relationships are easier to see visually on a graph than they are in rows of data.

Using the Scatterplot generated about the relationship between average points scored and total numbers of wins, you can see a positive correlation.

Chart, scatter chart

Description automatically generated

Correlation between Average Points Scored and the Total Number of Wins

Pearson Correlation Coefficient = 0.4777

P-value = 0.0

A Pearson correlation coefficient is a number that can range from -1 to +1. A -1 indicates a strong negative correlation while a +1 indicates a strong positive correlation. The Pearson Correlation Coefficient for this data is 0.4777, which indicates a moderate positive correlation.

The P-Value for this data is 0.0. This data is used to perform a hypothesis test upon the linear relationship between variables.

H0 (null): ρ = 0 (There is no linear relationship between Number of wins and Average Points scored.)

Ha (alternative): ρ ≠ 0 (There is a linear relationship between Number of wins and Average points scored.)

α (Significance Level) = 0.01

P-Value = 0.00

P-Value 0.00 < Significance Level 0.01; The null hypothesis is rejected in favor of the alternative. There is a linear relationship between number of wins and average points scored.

## 4. Simple Linear Regression: Predicting the Total Number of Wins using Average Points Scored

The Simple Linear Regression model is Y = β0 + β1X. A simple linear regression model functions off two types of variables and regression parameters. The Response variable is the variable of interest you are trying to predict. In this model that Response variable is the total number of wins. The Predictor variable is the variable that you suspect is influencing that response variable. β0 (intercept parameter) and β1 (slope parameter) are regression parameters obtained by sampling the data. In this model the predictor variable is the average points scored. Using this model, you can summarize the relationship between predictor and response variables using an equation.

The equation for this Linear Progression model is Y = -85.5476 + 1.2849x. Y, the response variable, is the games won. X, the predictor variable, is the average points scored. Using these response and predictor variables an F-Test was conducted to determine the relationship between the X and Y variables. Results of Test Statistic and the P-Value of this F-Test are in the table below.

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 182.1  *\*Round off to 2 decimal places.* |
| P-value | 0.0000 (1.52e-36)  *\*Round off to 4 decimal places.* |

H0 (null hypothesis): β1 = 0 (Average points scored are not linearly related to the total number of wins.)

Ha (alternative hypothesis): β1 ≠ 0 (Average points scored is linearly related to the total number of wins.)

α (level of significance) = 0.05

P-Value = 0.000

With a P-Value of 0.000 less than the Level of Significance 0.05 the Null hypothesis is rejected. The Average points scored is linearly related to the total number of wins.  
This hypothesis test determines that there is a correlation between the two values. It does not speak to how strong that correlation is. The results of this overall F-Test indicate a prediction can be made.

If you were to use this model you could begin to make some predictions. For a team averaging 75 points per game, the Linear Progression equation predicts 10 games won. For a team averaging 90 points per game the model predicts 30 games won.

**5. Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill**

Chart, scatter chart

Description automatically generated

Correlation between Average Relative Skill and Total Number of Wins

Pearson Correlation Coefficient = 0.9072

P-value = 0.0

A Pearson correlation coefficient is a number that can range from -1 to +1. A -1 indicates a strong negative correlation while a +1 indicates a strong positive correlation. The Pearson Correlation Coefficient for this data is 0.9072, which indicates a Strong positive correlation.

The P-Value for this data is 0.0. Using the hypothesis test as follows. . .

H0 (null): β1 = 0 (There is no correlation between Number of wins and Average Relative Skill.)

Ha (alternative): β1 ≠ 0 (There is a correlation between total number of wins and Average Relative Skill.)

α (Significance Level) = 0.01

P-Value = 0.00

P-Value 0.00 < Significance Level 0.01; The null hypothesis is rejected in favor of the alternative. There is a correlation between total number of wins and Average Relative Skill.

## 6. Multiple Regression: Predicting the Total Number of Wins using Average Points Scored and Average Relative Skill

A multiple regression model is used to predict a response variable using multiple predictor variables. This model allows for a more dynamic analysis of a response variable and better determination of the impact of different predictor variables. Additionally, while using this model you can test if a predictor variables influence is close enough to zero to warrant removal. A multiple regression model allows for more dynamic analysis.

The equation for this multiple regression model is Y = -152.5736 + 0.3497X1 + 0.1055X2. In this model, X1 is the predictor variable of Average Points, X2 is the Average Relative Skill. Results of Test Statistic and the P-Value of this F-Test are in the table below.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580  *\*Round off to 2 decimal places.* |
| P-value | 0.000 (4.41e-243)  *\*Round off to 4 decimal places.* |

H0 (null hypothesis): β1 = β2 = 0 (The model is not useful, and the slope parameters are all equal to 0.)

Ha (alternative hypothesis): βi ≠ 0 for i = 1, 2 (At least one predictor is useful in predicting Total Wins, not all slope parameters are equal to 0.)

α (level of significance) = 0.05

P-Value = 0.000

P-Value 0.000 < Level of Significance 0.05. The Null hypothesis is rejected in favor of the Alternative Hypothesis. At least one predictor variable in this model is useful in predicting Total Wins.

The individual t-tests of these parameters are 0.000. Testing with a Level of Significance of 0.01 results in both the Average Points Scored and Average Relative Skill being determined as statistically significant.

The Coefficient of Determination is how much variability is accounted for by the model. For this model the Coefficient of Determination is 0.837. This means 83.7% of the variability is accounted for in Total wins by Average Points Scored and Average Relative Skill.

This model predicts that if you averaged 75 Points per game and an average relative skill level of 1350 you would win 16 games that season. If you were to score on average 100 points per game and average a Relative Skill level of 1600 you would win 51 games in a season.

## 7. Multiple Regression: Predicting the Total Number of Wins using Average Points Scored, Average Relative Skill, and Average Points Differential

A multiple regression model is used to predict a response variable using multiple predictor variables. This model allows for a more dynamic analysis of a response variable and better determination of the impact of different predictor variables. Additionally, while using this model you can test if a predictor variables influence is close enough to zero to warrant removal. A multiple regression model allows for more dynamic analysis.

The equation for this multiple regression model is Y = -35.8921 + 0.2406X1 + 0.0348X2 + 1.7621X3. In this model, X1 is the predictor variable of Average Points, X2 is the Average Relative Skill, X3 is the Average Points Differential. Results of Test Statistic and the P-Value of this F-Test are in the table below.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449  *\*Round off to 2 decimal places.* |
| P-value | 0.000 (5.03e-278)  *\*Round off to 4 decimal places.* |

H0 (null hypothesis): β1 = β2 = β3 = 0 (The model is not useful, and the slope parameters are all equal to 0.)

Ha (alternative hypothesis): βi ≠ 0 for i = 1, 2, 3 (At least one predictor is useful in predicting Total Wins, not all slope parameters are equal to 0.)

α (level of significance) = 0.05

P-Value = 0.000

P-Value 0.000 < Level of Significance 0.05. The Null hypothesis is rejected in favor of the Alternative Hypothesis. At least one predictor variable in this model is useful in predicting Total Wins.

The individual t-tests of these parameters are 0.000. Testing with a Level of Significance of 0.01 results in both the Average Points Scored, Average Relative Skill and Average Points Differential being determined as statistically significant.

The Coefficient of Determination is how much variability is accounted for by the model. For this model the Coefficient of Determination is 0.876. This means 87.6% of the variability is accounted for in Total wins by Average Points Scored and Average Relative Skill.

This model predicts that if you averaged 75 Points per game, an average relative skill level of 1350 and an average points differential of + 5 you would win 37 games that season. If you were to score on average 100 points per game, have a Relative Skill level of 1600 and average a points differential of -5 you would win 35 games in a season.

## 8. Conclusion

In conclusion, regression models allow the ability to home in on what predictor variables are most likely to influence the response variables. Based upon our analysis there is a strong correlation between Total Number of Wins, Average Points scored, Average Relative Skill and Average points Differential.

The correlation of Total Wins with Average Points scored was present with a Pearson Correlation Coefficient of 0.4777. A stronger correlation was found between Total Wins and Average Relative Skill, possessing a Pearson Correlation Coefficient of 0.9072. We further assessed this relationship using a Multiple Regression model F-Test assessing the relationship between Total Number of wins, Average Points Scored and Average Relative Skill; The model assessing these variables accounted for 83.7% of the variation. To further analyze this relationship Average Points Differential was added. An F-Test assessing the relationship between Total Number of Wins, Average points Scored, Average Relative Skill and Average points Differential accounted for 87.6% of the variation in Total number of wins.