# MAT 243 Project Three Summary Report

Shannon Haglund

shannon.haglund@snhu.edu

Southern New Hampshire University

## 1. Introduction

We will be examining a large set of historical data from the NBA during the years 1995 and 2015 to analyze performance patterns which will help determine the total wins in a regular season based on variables such as the average points scored, average relative skill, average point differential between the team and their opponents, and average relative skill differential between the team and their opponent. We will use simple linear regression and multiple regression models to make these predictions, based on a variety of the predictor variables. We will visually analyze the data with scatterplots to see the correlation between average relative skill and points won, as well as the relationship between average points scored and points won.

## 2. Data Preparation

The variables that we will use include ‘total\_wins’ which represent the total number of wins in a regular season. The next variable is ‘avg\_pts’ which represents the average points scored in a regular season. The variable ‘avg\_elo\_n’ represents the average relative skill of each team in a regular season. The average point differential between team and their opponents in a regular season is represented by the variable ‘avg\_pts\_differential’. This variable tells us the average difference between the points between each team and their opponents. Finally, the variable ‘avg\_elo\_differential’ represents the average relative skill differential between the team and their opponent in a regular season. This variable tells us the average relative skill difference between each team and their opponents.

## 3. Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill

Data visualization techniques are used to study the relationship trends between two variables by removing the noise from the data and presenting it in a way that is easier to understand. The correlation coefficient is used to determine the strength and direction of the association between two variables. The closer the number is to zero, the weaker the correlation, and the closer to positive or negative one represents a stronger correlation. A positive trend is represented by a positive number and a negative trend is represented by a negative number. Generally, a weak correlation falls between 0.00 and 0.40, a moderate correlation falls between 0.40 and 0.80, and a strong correlation falls between 0.80 and 1.00.

Chart, scatter chart

Description automatically generated

The scatterplot visually represents a positive trend, and based on the closeness of each data point, we can see it may represent a strong correlation. Based on the Pearson Correlation Coefficient, which is 0.9072, we can determine that it does represent a strong, positive trend because it falls between 0.80 and 1.00 (0.80 < 0.9072 < 1.00). The p-value is 0.000. Using a 1% level of significance, or 0.01, the correlation coefficient is statistically significant based on the P-value, because 0.000 < 0.01.

## 4. Simple Linear Regression: Predicting the Total Number of Wins using Average Relative Skill

A simple linear regression model is used to predict the response variable by using the predictor variable. This works by finding a linear equation that describes the correlation of the predictor variable with the dependent variable by finding a line that represents the data using the least squares. We will be using average relative skill to predict total number of wins in this linear regression model. The equation for our model is y = 0.1121(x) - 128.2475. The y variable equals the total number of wins, and the x variable represents the average relative skill. We can use this model to predict the total number of wins based on average relative skill. The null hypothesis is that the slope parameter should equal zero (H0: B1 = 0) and the alternative hypothesis is that the slope coefficient is not equal to zero (H1: B1 ≠ 0). The test statistic is 2865.00 and the p-value is 0.0000. Based on this F-test, we can reject the null hypothesis because 0.0000 < 0.01 or 1%.

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2865.00 |
| P-value | 0.0000 |

Our conclusion based on the F-test means that relative skill can predict the total number of wins in the regular season. Using our model, if a team has an average relative skill of 1550, they will win 45 games (y = 0.1121 \* 1550 - 128.2475 = 45.5075). If we used this model to predict how many games they will win and their average relative skill was 1450, they would win 34 games (y = 0.1121 \* 1450 - 128.2475 = 34.2975).

**5. Scatterplot and Correlation for the Total Number of Wins and Average Points Scored**

*Chart, scatter chart

Description automatically generated*

This scatterplot shows a moderate correlation between average points scored and total number of wins. We can see the graph has a general positive trend and the data points are slightly more scattered around. The Pearson correlation coefficient is 0.4777, which represents a moderate correlation because it falls between 0.40 and 0.80 (0.40 < 0.4777 < 0.80). The p-value is 0.000 which means the correlation coefficient statistically significant because 0.000 is less than the significance level at 1% (0.0000 < 0.01)

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

Multiple linear regression models are used to predict the response variables by using multiple predictor variables, or more than one independent variable when holding the other variables constant. We will be using the average relative skill and the average points scored to make predictions about the total number of wins in this multiple regression model. The equation for this model is y = 0.1055(x1) + 0.3497(x2) – 152.5736. The first x variable represents the average relative skill, and the second x variable represents the average points scored in a regular season. For the F-test, our null hypothesis is that no relationship exists between the predictor variables (H0: B1 = B2 = 0) and the alternative hypothesis is that at least one variable does not equal zero (H1: at least one Bn ≠ 0 for n = 1, 2). The test statistic is 1580.00 and the p-value is 0.0000 which is statistically significant as 0.0000 < 0.01 or 1%

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00 |
| P-value | 0.0000 |

The conclusion for this hypothesis test is that there is a statistically significant linear relationship between total games won based on at least one of the variables average relative skill and average points scored in a season. The results of the individual t-tests are statistically significant using a 1% significance level because both predictor variables have a p-value of 0.0000 which is less than 0.01 (0.0000 < 0.01). The coefficient of determination is 0.837, which tells us that 83.7% of the variance in games won can be explained by the independent variables. Using our model, if a team is averaging 75 points per game and has a relative skill level of 1350, they would win 16 games (y = 0.1055\*(1350) + 0.3497\*(75) – 152.5736 = 16.0789). The predicted number of wins for a team that averages 100 points per game and has an average relative skill level of 1600 would be 51 wins (y = 0.1055\*(1600) + 0.3497\*(100) – 152.5736 = 51.1964).

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

Multiple linear regression models are used to make predictions on the response variable based on more than one predictor variable. We will be using the average relative skill, average points scored, average points differential, and average relative skill differential to make predictions about the total number of wins in this multiple regression model (the variables are in the same order in the equation). The equation for this model is y = -0.0134(x1) + 0.2597 (x2) + 1.6206 (x3) + 0.0525 (x4) + 34.5753. The null hypothesis is that no relationship exists between the predictor variables (H0: B1 = B2 = B3 = B4 = 0). The alternative hypothesis states that there is at least one variable that does not equal zero and that a statistically significant relationship exists. The test statistic is 1102.00 and the p-value is 0.0000, which is statistically significant because 0.0000 is less than the significance level of 1%, or 0.01.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1102.00 |
| P-value | 0.0000 |

Based on the results of the overall F-test, there is at least one predictor that is statistically significant in predicting the number of wins in the season. The conclusion for this hypothesis test is that we can reject the null hypothesis because there is a statistically significant linear relationship that exists between these four variables and the response variable. The individual t-test for the average relative skill variable is not significant at a 1% significance level (0.442 > 0.01). The individual t-test for the total points scored variable is significant at a 1% significance level (0.000 < 0.01). The individual t-test for the average points differential variable is significant at a 1% significance level (0.000 < 0.01). And finally, the individual t-test for the average relative skill differential variable is significant at a 1% significance level (0.004 < 0.01). The coefficient of determination is 0.878 which tells us that 87.8% of the variance in games won can be explained by the four predictor variables. The predicted number of wins in a regular season for a team that is averaging 75 points per game with a relative skill level of 1350, average point differential of -5 and average relative skill differential of -30 is 26 games (y = -0.0134\*(1350) + 0.2597\*(75) + 1.6206\*(-5) + 0.0525\*(-30) + 34.5753 = 26.2848). The predicted total number of wins in a regular season for a team that is averaging 100 points per game with a relative skill level of 1600, average point differential of 5 and average relative skill differential of 95 is 52 games (y = -0.0134\*(1600) + 0.2597\*(100) + 1.6206\*(5) + 0.0525\*(95) + 34.5753 = 52.1958).

**8. Conclusion**

Based on our test results, we can determine that there is a strong, positive correlation between average relative skill and number of wins. This tells us that the higher the average relative skill, the higher the number of games that team will win. Based on these variables, we were able to find an equation that represents the relationship between average relative skill and points won and use the overall F-test to determine that a statistically significant linear relationship exists. We then examined the relationship between average points scored and number of wins, which was determined to be a moderate, positive correlation between these variables. Next, we used the average relative skill and average points scored to create a multiple regression model to predict games won. We were able to determine that there is a statistically significant linear relationship between these variables and total games won based on the overall F-test and individual t-tests. Finally, we used four variables (average points scored, average relative skill, average point differential between the team and their opponents, and average relative skill differential between the team and their opponent) to create a multiple regression model to predict the games won. I would recommend using the multiple regression model that uses the four variables as it passed the F-test and the R-squared value is 0.878, which tells us that 87.8% of the variance in games won can be explained by the four predictor variables and is a higher proportion than the other models. I think this model works well because you would expect that multiple performance metrics will help determine the total number of games won in a regular season.