# MAT 243 Project Three Summary Report

Jon Hoover

Jonathan.Hoover1@snhu.edu

Southern New Hampshire University

## 1. Introduction

In this project, I will use regression models to try and predict how many wins our team will have in a regular game based on the performance metrics that are included in our data set. The data set is the FiveThirtyEight NBA Elo dataset from Kaggle. The results of our analysis will be used to help make key decisions to improve the performance of the team.

## 2. Data Preparation

In the NBA Elo dataset, there are some important variables I use in this project.

* **total\_wins** – Shows the total number of wins in a regular season.
* **avg\_pts** – This is the average number of points scored in a regular season.
* **avg\_elo\_n** – Gives the average relative skill of each team in a regular season.
* **avg\_pts\_differential** – Shows the average point differential between the team and their opponents in a regular season.
* **Avg\_elo\_differential** – This is the average relative skill differential between the team and their opponent in a regular season.

The variable **avg\_elo\_n** gives the average of a team’s relative skill in a regular season, measured using an ELO rating. This measure is calculated using the final score of a game and the game location. It also looks at the outcome of the game relative to the probability of that outcome. If the team scores better in a game than is expected, the ELO rating goes up. And likewise, the ELO rating goes down, if the team fails to meet that expectation in a game.

The variable avg\_pts\_differential compares the average number of points scored in a regular season against the average number of points scored by their opponents in a regular season, and shows the difference of those two outcomes.

## 3. Simple Linear Regression: Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill

Chart, scatter chart

Description automatically generated

The above chart shows total number of wins in relation to the average relative skill of the team. Using a scatterplot, it shows there is a strong positive correlation between these two variables. To further enforce that theory, I calculated a Pearson Correlation Coefficient of 0.9072 which confirms there is a strong positive correlation between number of wins and relative skill of all teams. The p-value result is statistically significant at a 1% level of significance since the p-value is less than 0.01.

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

Simple linear regression uses a linear equation to estimate the total number of wins in a regular season, based on the average relative skill value. This equation is shown below:

The null hypothesis is that a higher average relative skill for a team would mean that the team would have more wins. The alternate hypothesis is that a higher average relative skill for a team means it will not have more wins. With a p-value of less than 0.0000, the results of my test are significant at a 1% level of significance. Therefore, I will accept the null hypothesis that the higher than average relative skill of team is, the more wins they will have in a regular season.

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | -40.731 |
| P-value | 0.0000 |

For a team with an average relative skill of 1550, my model would predict the team to have about 42 wins for a regular season. If the team had an average relative skill of 1450, then the model would predict that team to have about 31 wins for a regular season.

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

Chart, scatter chart

Description automatically generated

For this chart, I plotted the total number of wins in relation to average points scored. The scatter plot shows a weak positive correlation between number of wins and points scored. The Pearson correlation coefficient I calculated is 0.48, which also shows a weak positive correlation. The p-value is less than 0.0000 and concludes that the result is statistically significant within a 1% level of significance.

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

Multiple regression uses multiple predictor variables to predict the response variable. Here is the equation for my multiple regression model:

x1: Average point scored

x2: Average relative skill

y: Total number of wins

The null hypothesis states that if both average skill and number of wins are higher, then the chance of winning is higher. The alternate hypothesis says that if both average skill and number of wins are higher, then the number of wins will not be higher. With a p-value less than 0.0001, our result is statistically significant at a 1% level of significance, and I will accept the null hypothesis.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | -33.90 |
| P-value | 0.0000 |

For a team that is averaging 75 points per game with a relative skill level of 1350, I use my multiple regression equation to predict the team will have 16 total wins in a regular season. For a team that is averaging 100 points per game with an average relative skill level of 1600, I estimate the team will win 51 games in a regular season.

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

In this multiple regression model, we use three predictor variables to predict our response variable. Here is the equation for my multiple regression model:

x1: Average points scored

x2: Average relative skill

x3: Average points differential

x4: Average skill differential

y: Total number of wins

The null hypothesis states that I can use average points scored, average relative skill, average points differential, and average relative skill differential to predict number of wins. The alternative hypothesis says these predictor variables are not able to be used to predict a team’s number of wins. Based on the p-value I received of 0.1820, my result is not statistically significant within a 1% level of significance. Therefore, I will reject the null hypothesis and accept the alternate hypothesis that these predictor values do not help determine number of wins.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1.34 |
| P-value | 0.1820 |

Based on my multiple regression model, since the p-value for total number of wins is greater than 0.01, I am unable to rely on any of the variables to determine number of wins, using this model.

## 8. Conclusion

In conclusion, I can use average points scored by a team and that team’s average relative skill to predict number of wins in a regular season. However, I cannot use average points differential nor average relative skill differential to help with that prediction, based on my multiple regression models.

## 9. Citations

S. (2017, November 29). *p-value and level of significance explained*. Data Science Central. https://www.datasciencecentral.com/p-value-and-level-of-significance-explained/

FiveThirtyEight. (April 26, 2019). FiveThirtyEight NBA Elo dataset. Kaggle. Retrieved from https://www.kaggle.com/fivethirtyeight/fivethirtyeight-nba-elo-dataset/