# MAT 243 Project Three Summary Report

Brandon Petersen

Brandon.petersen@snhu.edu

Southern New Hampshire University

## Introduction

I am a data analyst hired by an NBA team and been provided access to a large set of historical data to analyze performance patterns of different teams. As a result, I have been assigned to predict the total number of wins for a team in the regular season based on key performance metrics. In addition, I will be implementing regression models to help with my prediction and help the management team make important decisions that will impact the team's performance. In these regression models, I will be looking at the total number of wins in a regular season, average points in a regular season, the average Elo score in a regular season, and the average difference between points between the team and their opponents in a regular season. In addition, I will be using various scatterplots and correlation coefficients to express the total wins, average points, and average Elo.

## Data Preparation

For this assignment, I will be using the python programming language to analyze the data and create reports. I will be using two critical variables in my code: **avg\_pts\_differential** and **avg\_elo\_n**. These two variables are essential metrics used in the regression models, scatterplots, and correlation. At the end of a game, the avg\_pts\_differential variable is the point difference between our team and their opponent's team score. For example, if our team scored 90 points and their opponents scored 95 points, the differential would be -5 points. The **avg\_elo\_n** variable is the average relative skill of the team during a regular season. The relative skill points are measured by the final score at the end of a game and where the game was played. If the team wins a match, they will see an increase in their Elo score, and if they lose a game, they will see a decrease in their Elo score.

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

As mentioned previously, I will be utilizing scatterplots and correlation coefficients in this assignment. When comparing multiple items, it is vital to ensure precise data visualization and what it represents. Being able to see the data in forms such as charts and graphs makes it easier to read and understand the data's information. Scatterplots are highly effective when observing the relationship between two variables, precisely what is happening in this assignment. The scatterplot makes it easy to see the trends and whether there is a strong correlation between the two variables being compared. We will be using the correlation coefficient to determine whether the trend observed in the scatterplot is a strong trend or a weak trend. The trend can be positive or negative based on the value that's received for the correlation coefficient. In this assignment, a p-value of 0.00 was calculated at a 1% level of significance. Having a p-value below the level of significance will be statistically significant and means that the null hypothesis will be rejected.

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

“A regression is used to develop a more formal understanding of relationships between variables. In regression, and in statistical modeling in general, we want to model the relationship between an output variable, or a response, and one or more input variables, or factors (JMP).” In general, a simple linear regression is used to show the relationship between two continuous variables. The two variables typically being the response and predictor. Simple linear regression is used to predict the value of the response variable based on the value of the predictor variable. The simple linear regression is used to determine the total number of wins using the average points scored in a regular season. The equation being used for our model is the following:

The null hypothesis is which indicates no correlation between having a high average point per game in a regular season and having a higher win average in a regular season. The alternate hypothesis is which indicates a correlation between having a high average point per game in a regular season and having a higher win average in a regular season. For this hypothesis test, a 5% significance level was used. Using this significance level yielded a test statistic of 182.1 and a p-value of 0.00, as shown below in the table.

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 182.1 |
| P-value | 0.0000 |

As mentioned, having a p-value below the level of significance will be statistically significant. Since the p-value was calculated to be 0.0000 with a significance level at 1%, this indicates we reject the null hypothesis and accept the alternate hypothesis. This concludes that if the team has more wins, they will also have a higher average of scores per game in a regular season. From these results, we can conclude that the f-test is accurate and that the average points scored can be sued to predict the total number of wins in the regular season. For example, the predicted total number of wins in a regular season for a team averaging 75 points per game is 11. The predicted total number of wins in a regular season for a team averaging 90 points per game is 25.

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

Observing the scatterplot below and the Pearson correlation coefficient of 0.4777, we can conclude a positive correlation between the average Elo score and the total number of wins. A positive correlation indicates that when the team sees an increase in their Elo score, there will also be an increase in the number of games they have won. Along with that, having a calculated p-value of 0.00 with a level of significance of 1% indicates that the correlation coefficient that was calculated is statistically significant.

Chart, scatter chart

Description automatically generated

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

Multiple linear regression is similar to simple linear regression. “Multiple linear regression analysis is an extension of simple linear regression analysis, used to assess the association between two or more independent variables and a single continuous dependent variable. Multiple regression analysis is also used to assess whether confounding exists. Since multiple linear regression analysis allows us to estimate the association between a given independent variable and the outcome holding all other variables constant, it provides a way of adjusting for (or accounting for) potentially confounding variables that have been included in the model (BUMC).” In other words, a multiple regression model is used to predict the value of the response variable based on the value of two or more predictor variables. For the multiple regression model used in this assignment, the response variable is the total number of wins, and the predictor variables are the average points scored per game and the average Elo score. The equation being used for our model is the following:

Total Wins = -152.5736 + 0.3497 \* (avg\_pts) + 0.1055 \* (avg\_elo\_n)

The null hypothesis is, where there is no correlation between the two predictor variables, average points scored per game and average Elo score, and the total wins in a regular season. The alternate hypothesis is, where i = 1, 2, 3, infers a correlation between the two predictor variables, average points scored per game and average Elo score, and the total wins in a regular season. For this hypothesis test, a 1% significance level was used. Using this significance level yielded a test statistic of 1580.00 and a p-value of 0.00, as shown below in the table.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00 |
| P-value | 0.00 |

As previously stated, having a p-value below the level of significance indicates statistical significance. Since the p-value was calculated to be 0.0000 with a significance level at 1%, this indicates we reject the null hypothesis and accept the alternate hypothesis. In other words, the average points scored per game and the average Elo score affects the total number of wins in a regular season.

Table

Description automatically generated

The chart above shows that the two predictor variables used in the multiple regression model have p-values of 0.00 while using a significance level of 1%. This indicates that both of the predictor variables are statistically significant to the regression model used. In the same regression model, we calculated a correlation coefficient of 0.837. This correlation coefficient indicates that approximately 83.7% of the data used works with this regression model that was used. This regression model predicts the total number of wins while averaging 75 points with an Elo of 1350 in a regular season is 16. Similarly, this regression model predicts the total number of wins while averaging 100 points with an Elo of 1600 in a regular season of 51.

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

Like a simple linear regression model, a multiple linear regression model is used to estimate the relationship between two or more predictor variables and the response variable. This assignment used the average number of points scored per game, the average Elo score, the average points differential as the predictor variables, and the total number of wins as the response variable. The equation of this multiple linear regression model is the following:

Total Wins = -35.8921 + 0.2406 \* (avg\_pts) + 0.0348 \* (avg\_elo\_n) + 1.7621 \* (avg\_pts\_diff)

The null hypothesis is, which indicates no correlation between the average number of points scored per game, the average Elo score, average points differential in a regular season, and the total number of wins. The alternate hypothesis is, where at least one i = 1, 2, 3 indicates a correlation between the average number of points scored per game, the average Elo score, and average points differential in a regular season and total wins. For this hypothesis test, a 5% significance level was used. Using this significance level yielded a test statistic of 1449.00 and a p-value of 0.00, as shown below in the table.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449.00 |
| P-value | 0.0000 |

As previously stated, having a p-value below the level of significance indicates statistical significance. Since the p-value was calculated to be 0.0000 with a level of significance at 5%, this indicates we reject the null hypothesis and accept the alternate hypothesis. In other words, there is a correlation between the average number of points scored per game, the average Elo score, and the average points differential in a regular season, and the total number of wins.

Text, table, letter

Description automatically generated

The chart above shows that the three predictor variables used in the multiple linear regression model have p-values of 0.00 while using a significance level of 1%. This indicates that the three predictor variables are statistically significant to the regression model used. In the same regression model, we calculated a correlation coefficient of 0.867. This correlation coefficient indicates that approximately 86.7% of the data used works with this regression model that was used. For this model, the predicted total number of wins in a regular season for a team averaging 75 points per game with a relative skill level of 1350 and an average point differential of -5 is 20. Similarly, using this model to predict the total number of wins in a regular season for a team averaging 100 points per game with a relative skill level of 1600 and an average point differential of +5 is 53.

## Conclusion

To come to the point, provided the data using multiple hypothesis tests and regression models, we can conclude that the total number of wins is directed correlated to the average points per game, average Elo score, and average differential score. This was visually shown using scatterplots and correlation coefficients and was shown statistically shown using a simple linear regression model, a multiple regression model, and a multiple linear regression model. This information indicates that if a team wants to increase their total wins in a given regular season, they need to improve their average Elo score directly. To increase their average Elo score, they will need to focus on increasing their average points per game. To have a strong team with many wins, good average points per game, average elo score, and high average points, differential need to be good. All three of these variables indicate how to get the team's total wins higher. While this analysis does not provide a template on how to increase all of these variables, it does show where the management team needs to focus on building a more robust and overall more rounded team.

**References**

* Multiple linear regression analysis. (n.d.). https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_multivariable/bs704\_multivariable7.html.
* *Simple linear regression*. JMP. (n.d.). https://www.jmp.com/en\_us/statistics-knowledge-portal/what-is-regression.html.