# MAT 243 Project Three Summary Report

Andrea Plunkett Jackson

andrea.jackson11@snhu.edu

Southern New Hampshire University

## Introduction

The determination of this summary statement is to analyze and forecast the total number of wins for an NBA team, constructed on crucial performance metrics during the regular season. Team management and the coach have requested the use of regression models in these predictions founded on historical data scrutinizing patterns to support vital decisions for development. The dataset to accomplish this analysis will incorporate the total number of wins, average points scored, average relative skills, and the average point differential concerning the club and their adversaries.

## Data Preparation

This project will employ variables average relative skill differential (avg\_elo\_n) and average point differential (avg\_pts\_differential). The variable average relative skill differential epitomizes the variance in the degree of average skill level comparable between two teams demonstrating which team has greater skills and permits for predictive outcome of a game based on preceding performance. An illustration of this would be if Team A has a higher avg\_elo\_n than Team B, it would signify that Team A executed better than average in previous games consequently is more probable to triumph in impending games against Team B.

Akin to avg\_elo\_n, average point differential takes a proportional assessment into game performance examining final scores, calculating the difference using subtraction and averaging the points differential in the series. Gauging a team’s ascendency in their games by means of statistical analysis offers the average margin in which a team triumphs or fails. Avg\_pts\_differential can be either positive or negative giving indication on the nearness of the outcomes with near zero signifying a tenacious bracket pairing. For instance, if Team A scores 122 points and Team B only accomplishes 100 in a specified game, the points differential for

that specific competition would be 22 points. These results allow for prognostication prospects to be conducted approximating outcome probability.

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

Data visualization allows complex data to be efficiently translated into a manageable and amicable presentation generating a story that brings the data to life while offering acumens through the study of affiliation inclinations among variables. This connection of comparative movement of the variables is a statistical measurement known as the correlation coefficient calculating the strength and direction with parameters ranging from -1.0 to 1.0. A seamless positive correlation indicator of 1.0 signifies that as one variable increases, the other increases. Whereas a -1.0 displays a negative correlation, implying that as one variable increases the other decreases. Conclusively, a correlation coefficient of zero designates no association displaying that deviations in one variable has no predictive impact on the other. This direction of association can straightforwardly be evaluated based on the attribute of the sign being positive or negative. Determination of strength additionally is revealed by the absolute value of the correlation coefficient with calculations near to 1 indicating a strong relationship and conversely those near zero representing a weak affiliation.

A graph of blue dots

Description automatically generated

The scatterplot above visually shows the connection concerning the total number of wins and average relative skill. As the points are moving in an upwards movement and closely organized to a linear pattern it designates a positive correlation. The Pearson correlation coefficient of 0.9072 quantifies the degree of correlation between the total number of wins and average relative skill. Signifying, as the total number of wins rises the average relative skill correspondingly increases therefore as it is close to 1.0 it specifies an extraordinarily strong positive linear correlation. Using a 1% level of significance and calculated p-value equal to 0.0 we can conclude that the correlation is statistically significant showing that the detected correlation did not happened coincidentally.

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

Simple linear regression is a statistical process allowing for summarization and learning the relationships concerning two continuous quantitative variables. One variable is considered the predictor or independent variable with the other observed as the response or dependent. For this analysis, the average relative skill level is the predictor with the total number of wins the response variable. The equation for our model is : Y(total\_wins) = -128.2475 + 0.1121 \* avg\_elo\_n. Evaluating that for every one-unit the avg\_elo\_n (average relative skill) increases, the expected outcome is that there is a 0.1121 comparable rise in the total number of wins.

In providing the results of the overall F-test the assessment of the inclusive efficiency of the model equation, a summary of all critical stages for this hypotheses test have been included. Establishment of the Null Hypothesis , H₀ : β₁ = 0. The null hypothesis or H0 denoted in statistical notation is a fundamental concept in testing. Considered the default postulation that there is zero correlation in the average relative skill variable (β₁) impact on the total number of wins. Testing is conducted with the intention of assessing if there is evidence from the sample data that supports (accepts) or contradicts (rejects). And is considered true until otherwise proven. The alternative hypothesis or Ha in the statistical notation disputes the null hypothesis and epitomizes that there is an influential association. It is compared alongside the null to draw conclusions about the relationship itself. The alternative hypothesis (Ha) stated for our analysis is that the average relative skill level (β₁) does not equal zero, Ha **:** β₁ ≠ 0. The level of significance is a value in which the p-value is equated to evaluate if the null hypothesis is supported or contradicted. The level of significance used in our analysis is 1% ( a = 0.01). As the table below shows the p-value is 0.0 which is less than the level of significance therefore rejecting the null hypothesis that the average relative skill level has zero impact on total number of wins. Demonstrating a statistically significant affiliation on our predictor variable. Based on the results of the overall F- test, it can be determined that average relative skill is a reliable predictor on the total number of wins in a regular season.

Table 1: Hypothesis Test for the Overall F-Test

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

Using the equation for our model, Y(total\_wins) = -128.2475 + 0.1121 \* avg\_elo\_n, we can then predict the total number of wins in a regular reason for a team that has an average relative skill of 1550.

Y(total\_wins) = -128.2475 + 0.1121 \* 1550

Y (total\_wins) = 45 games (rounded down to the nearest integer)

In addition to predicted number of wins in a regular season for a team that has an average relative skill of 1450.

Y(total\_wins) = -128.2475 + 0.1121 \* 1450

Y (total\_wins) = 34 games (rounded down to the nearest integer)

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

The scatterplot below titled “ Total Number of Wins by Average Points Scored” displays an inclination where an increase in average points scored correlates with an increase in total number of wins indicating a positive correlation between the two variables. This can also be collaborated by the Pearson Correlation Coefficient of 0.4777 ,as discussed prior, being as it falls into the range of zero to one it specifies a moderate positive correlation. Nevertheless, although these two variables move in unison it is not an extraordinarily strong correlation . Using a 1% (a = 0.01) level of significance with a P-value of 0.00 it can be concluded that there is a statistical significance limiting the likelihood the observed correlation is merely due to serendipity.

**A graph of blue dots

Description automatically generated**

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

A multiple linear regression model expands upon the concepts of simple linear regression discussed formerly. It is a method that uses more than one variable to predict the outcome of a response variable with the objective of modeling the linear relationship amongst independent variables and the response dependent. The equation for our model is : Y(total\_wins) = -152.5736 + 0.1055 \* (avg\_elo\_n) + 0.3497 \* (avg\_pts).

In delivering the outcomes of the overall F-test the valuation of the comprehensive efficacy of the model equation, a synopsis of all crucial steps for this hypotheses test have been provided. Formation of the Null Hypothesis , H₀ : β₁ = β₂ = 0. Representing that the null hypothesis has no consequence on the dependent variable. The alternative hypothesis or Ha in the statistical notation disputes the null hypothesis and epitomizes that there is an influential association. It is compared alongside the null to draw conclusions about the relationship itself. The alternative hypothesis (Ha) stated for our analysis, Ha **:** at least one i ≠ 0 for i = 1, 2. The level of significance is a value in which the p-value is equated to evaluate if the null hypothesis is supported or contradicted. The level of significance used in our analysis is 1% ( a = 0.01). As the table below shows the p-value is 0.0 which is less than the level of significance therefore rejecting the null hypothesis. Demonstrating a statistically significant affiliation that at minimum one parameter is not equal to zero.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00 |
| P-value | 0.0000 (4.41e-243 essentially zero) |

Based on the results of the overall F-test, we can confirm at least one of the predictors is statistically significant in predicting the total number of wins in the season. The individual t-tests results for the parameters of each predictor variable are as follows:

avg\_elo\_n (average relative skill): t-value = 47.952 and the P-value = 0.000

avg\_pts (average points) = t-value = 7.297 and the P-value = 0.000

Assuming a 1% (a = 0.01) level of significance, mutually both predictor variables are measured statistically significant as equally P-values are less than a (0.01). The coefficient of determination, denoted as R², equals 0.837. Affirming that 83.7% of the variability in the dependent variable total\_wins can be elucidated by the average points scored and average relative skill.

Using the equation for our model, Y (total\_wins) = -152.5736 + 0.1055 \* (avg\_elo\_n) + 0.3497 \* (avg\_pts), we can thenpredict the total number of wins in a regular reason for a team that is averaging 75 points per game with a relative skill level of 1350.

Y (total\_wins) = -152.5736 + 0.1055 \* (1350) + 0.3497 \* (75)

Y (total\_wins) = 16.079 = 16 games

In addition to predicted number of wins in a regular season for a team that is averaging 100 points per game with an average relative skill level of 1600.

Y (total\_wins) = -152.5736 + 0.1055 \* (1600) + 0.3497 \* (100)

Y (total\_wins) = 51.1964 = 51 games

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

A statistical tool in analysis that permits us to recognize the relationship between numerous predictor variables in contradiction of a response dependent variable is called a multiple regression model. Through offering a quantifying way to understand how aspects of the variables average points scored (avg\_pts), average relative skill (avg\_elo\_n), average points differential (avg\_pts\_differential) , and average relative skill differential (avg\_elo\_differential) collectively impact the outcome. The equation for our model is : Y(total\_wins) = 34.5753 - 0.0134 \* (avg\_elo\_n) + 0.2597 \* (avg\_pts) + 1.6206 \* (avg\_pts\_differntial) + 0.0525 \* ( avg\_elo\_differential).

In regard to overall F-test, the null hypothesis H₀ : β₁ = β₂ = β₃ = β₄ = 0, signifying that all predictors are inconsequential with no statistically significant relationship. The alternative hypothesis is that there is a statistically significant association present between at least one of the predictor variables and the response, Ha **:** at least one i ≠ 0 for i = 1, 2, 3, 4. Per the level of significance established at 1% (a = 0.01) being that the overall F-test P-value is **3.07e-278 that it so small a value that it is essentially zero we can conclude that the model is statistically significant as it is less than hence rejects the null hypothesis and the predictor variables at least one is influential in predicting the response variable total\_wins.**

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1102.00 |
| P-value | 0.0000 ( **3.07e-278 essentially zero)** |

The individual t-tests results for the parameters of each predictor variable are as follows:

avg\_elo\_n (average relative skill): t-value = -0.769 and the P-value = 0.442

avg\_pts (average points) = t-value = 6.076 and the P-value = 0.000

avg\_pts\_differential (average point differential between the team and their opponents in a regular season) = t-value = 12.024 and the P-value = 0.000

avg\_elo\_differential (average relative skill differential between the team and their opponent in a regular season) = t-value = 2.915 and the P-value = 0.004

At a 1% (a = 0.01) level of significance, a predictor variable is measured statistically significant if its P=value is less than a (0.01) therefore mutually avg\_pts, avg\_pts\_differential, and avg\_ elo\_differential would be whereas avg\_elo\_n would not. The coefficient of determination, denoted as R², equals 0.878. Affirming that 87.8% of the variability in the dependent variable total\_wins can be elucidated by three of the four predictors. Validating the rejection of our null hypothesis.

Using the equation for our model, Y(total\_wins) = 34.5753 - 0.0134 \* (avg\_elo\_n) + 0.2597 \* (avg\_pts) + 1.6206 \* (avg\_pts\_differntial) + 0.0525 \* ( avg\_elo\_differential), we can thenpredict the total number of wins in a regular reason 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.

Y(total\_wins) = 34.5753 - 0.0134 \* (1350) + 0.2597 \* (75) + 1.6206 \* (-5) + 0.0525 \* (-30).

Y (total\_wins) = 26.2848 = 26 games

In addition to predicted number of wins in a regular season for a team that is averaging 100 points per game with an average relative skill level of 1600, average point differential of +5 and average relative skill differential of +95.

Y(total\_wins) = 34.5753 - 0.0134 \* (1600) + 0.2597 \* (100) + 1.6206 \* (5) + 0.0525 \* (95).

Y (total\_wins) = 52.1958 = 52 games

## Conclusion

In conclusion, execution of statistical analysis to forecast the total number of wins for a team constructed on the factors: the average points scored (avg\_pts), average relative skill (avg\_elo\_n), average points differential (avg\_pts\_differential) , and average relative skill differential (avg\_elo\_differential) provided the following findings. Average points scored was a significant predictor displaying that a team that scores more points on average are inclined to leverage an opportunity. The equivalent for average points differential indicating that teams that scores more than their challenger will gain an edge. Lastly, average relative skill is a worthy predictor as teams that are rated higher than adversaries garner an advantage. Taking that into consideration the research concluded that average relative skill does not meaningfully affect the total number of wins.

The real-world importance of the analyses completed lies in its ability to predict the total number of wins for a team. This can be beneficial for team strategy, focusing on improvement in a plan of action in regard to offense and defense while restraining opponents’ scoring. Finally, an understanding of assets and deficiencies to pilot player recruitment.

## References

*FiveThirtyEight NBA Elo Dataset*. (n.d.). Kaggle.com. <https://www.kaggle.com/fivethirtyeight/fivethirtyeight-nba-elo-dataset/>

‌