# MAT 243 Project Three Summary Report

Sarah Steinbaum

sarah.steinbaum@snhu.edu

Southern New Hampshire University

## 1. Introduction

In this summary report, I will be using performance metrics from a historical data set from the NBA between the years 1995 – 2015. The results of this analysis will be used to create regression models in order to predict the number of wins in a regular game for the coach and management and to understand what metrics the team needs to address to improve performance. I will use the Python programming language to conduct this analysis using variables including total number of wins, average points scored, average relative skill, and the average point differential between the team and their opponents.

## 2. Data Preparation

The variables I will be using to conduct this analysis will assist heavily when making predictions. One variable is *total\_wins* which represents the number of wins in a regular season. The variable *avg\_points* represents the average number of points scored in a regular season. The *variable avg\_elo\_n* is the average relative skill of each team in a regular season. The variable *avg\_points\_differential* is the average point differential between the team and their opponents. The variable *avg\_elo\_differential* is the average relative skill differential between the team and their opponent.

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

Data visualization is the visual representation of data that provides us with an overview of the distribution of data in a graph or chart which makes the data easier to understand and analyze patterns or trends in data. The correlation coefficient is represented by the letter ‘*R*’ and tells us if the correlation is positive or negative as well as the strength of the correlation being made. The correlation coefficient is strong when it is close to 1 or -1, and considered weak if it is closer to zero.

Chart, scatter chart

Description automatically generated

The scatterplot above displays the overall distribution between the total number of wins and average relative skill and allows us to visualize the correlation between the two variables. This graph shows us a positive correlation between the two variables that as the average relative skill increases, the total number of wins also increases. The Pearson correlation coefficient is 0.9072 which signifies a strong correlation as it is very close to 1. The P-value is 0.0 which is less than the level of significance of 0.01 or 1%, indicating that this correlation coefficient is statistically significant.

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

A simple linear regression model is an equation that is used display the relationship between two variables in order to predict a response for a given predictor value. The equation for my model is:

.

The objective for this scenario is to predict the total number of wins based on average relative skill. The null hypothesis is that there is no correlation between the total number of wins and average relative skill. The alternative hypothesis is that there is a correlation between those two variables.

Null Hypothesis : = 0

Alternative Hypothesis : **≠** 0

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2865 |
| P-value | 8.06e-234 or fundamentally 0.0 |

There is sufficient evidence to reject the null hypothesis as the p-value is essentially 0.0 and less than the level of significance. Based on the results of the overall F-test, average relative skill can predict the total number wins in the regular season. A team with an average relative skill of 1550 would be likely to win 45 games:

.

A team that has an average relative skill of 1450 would likely win 34 games:

.

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

Chart, scatter chart

Description automatically generated

This scatterplot above also shows an overall distribution of the total number of wins by average points scored. However, this chart has a larger spread than the previous scatterplot. The Pearson correlation coefficient is 0.4777 which indicates that the strength of this correlation is considered moderate. The p-value is 0.0 and is less than the level of significance of 0.01 or 1% which signifies that the correlation coefficient is statistically significant.

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

A multiple linear regression model is an equation that is used to predict a response for multiple predictor values. The equation for my model is:

.

The null hypothesis in this scenario is that there is no relationship between the response variable or any predictor variable. The alternative hypothesis is that there is a relationship between the response variable and at least one predictor variable. The level of significance is 0.01 or 1%.

Null Hypothesis : = = 0

Alternative Hypothesis : At least one **≠** 0 for i = 1, 2

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580 |
| P-value | 4.41e-243 or fundamentally 0.0 |

The P-value is 0.0 and less than the level of significance of 0.01 or 1% which signifies that there is enough evidence to reject the null hypothesis. This means that at least one of the predictor variables is statistically significant in predicting the total number of wins in a season. The results of individual t-tests for the parameters of each predictor variable are: *avg\_points* = 7.297, and *avg\_elo\_n* = 47.952. Each of the predictor variables have a p-value of 0.0 which means that they are statistically significant as their own p-values are below the level of significance of 0.01 or 1%. The coefficient of determination is 0.837 which signifies that this model has a strong accuracy of roughly 84%. The predicted total number of wins in a regular season for a team that is averaging 75 points per game with a relative skill level of 1350 is 16:

The predicted number of wins for a team averaging 100 points per game with an average relative skill of 1600 is 51:

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

A multiple linear regression model is used to predict the response variable by using any number of predictor variables. The equation for my model is:

+ +

.

The null hypothesis in this scenario is that is no relation between the response variable which is the total number of wins and of the predictor variables. The alternative hypothesis is that there is a correlation between the response variable and at least one predictor variable. There is a significance level of 0.05 or 5%.

Null Hypothesis : = = = = 0

Alternative Hypothesis : **≠** 0 for at least one i = 1, 2, 3, 4

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1102 |
| P-value | 3.07e-278 or fundamentally 0.0 |

There is sufficient evidence to reject the null hypothesis as the p-value is less than the level of significance. This means that at least one of the predictors is statistically significant in predicting the number of wins in the season. The results of the individual t-tests for the parameters of each predictor variable are: *avg\_pts* = 6.070, *avg\_elo\_n* = -0.769, *avg\_pts\_differential* = 12.024, and *avg\_elo\_differential* = 2.915. There are three predictor variables that have a p-value less than the level of significance of 0.01 which indicate that they are statistically significant: *avg\_pts*, *avg\_pts\_differnetial*, and *avg\_elo\_differential*. The predicator variable, *avg\_elo\_n*, has a p-value of 0.442 which is greater than the 0.01 level of significance. The coefficient of determination is 0.878 which means that the accuracy of this model is strong at roughly 88%. 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 would be 26:

+ +

The predicted number of wins for a team that is averaging 100 points per game, with a relative skill level of 1600, average point differential of 5, and an average relative skill differential of 95 would be 52:

+ +

## 8. Conclusion

Overall, the hypothesis tests and regression models used in this analysis show that the total number of wins is impacted by the average points scored, average relative skill, and the average point differential between the team and their opponents in a regular season. In order for the team to increase their total number of wins in future seasons, the coach and management must focus on improving the team’s relative skill, increase points scored per game, and points scored over their opponents. The practical importance of the analyses performed in this report is to assist management and the coach in identifying the areas the team can make improvements in order to win more games.