# MAT 243 Project Three Summary Report

Angel Davila

angel.davila@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## 1. Introduction

For this analysis I am exploring historical NBA basketball data, which is primary focus on the game wins. We are going to study wins to predict how our teams’ wins will look at the end of the season. To run these predictions, we are going to build regression and multiple regressions models.

## 2. Data Preparation

Below is a list of variables which are going to be used during my analysis.

* total\_wins – The total number of wins for a team during the regular season.
* avg\_pts – is the average number of points scored by a team during the regular season.
* avg\_pts\_differential - is the average amount of points scored between 2 opponents. If your average points differential is 10, that means, on average you are at a 10-point difference compared to your opponent.
* avg\_elo\_n – is a score given to each team which represents their skill level during the regular season.

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

Data visualization techniques are used to see the correlation between variables and to be able to break down and study trends or patterns in that data in different ways. The correlation coefficient value is between -1 and +1. Positive and negative values define the direction between 2 variables.

Below is a scatter plot which represents the total number of wins by average points scored.

Chart, scatter chart

Description automatically generated

The scatter plot shows a positive relationship between the number of wins and point scored, that is if the average number of points scored increases, so does the total number of wins.

It is determined that the p-value for this test shows as 0.0 which means that it is less than the 1% significance level. Because of this find, it is safe to reject the null hypothesis and safe to say that this test is statistically significant.

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

A regression model uses a couple factors including a dependent variable, which the factor you are solving for and independent variables which are used for the prediction. A simple regression model takes a line between a dependent and independent variable and shows us the relationship between them.

The equation used for our model is:

The null is a test of the likelihood of the statement presented to be true before we begin our analysis. The statistical notation for this is:

The alternative hypothesis is the idea of determining whether to accept or reject the null hypothesis. The statistical notation for this is:

The level of significance for our analysis is 5%.

For our analysis, the test statistic is F = 182.1 and the P-value is 0

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 182.10 |
| P-value | 0.0000 |

Because the p-value is less than 5%, we will reject the null hypothesis and that the average amount of points scored by our team is statistically significant.

After our analysis we can predict that the number of wins is 46 with an average relative skill of 1550:

After our analysis we can predict that the number of wins is 34 with an average relative skill of 1550:

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

Below is a scatterplot graph showing wins by average skill rating. You can see right away that compared to the graph above which shows total number of wins by average points scored, there is a clear positive relationship between number of wins and average relative skill. The Pearson Correlation Coefficient is 0.4777 and our p-value is 0.0 which is less than 1%, therefore we can say the correlation coefficient is statistically significant.

Chart, scatter chart

Description automatically generated

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

In general, multiple linear regressions is used to predict the dependent variable based on 1 independent variable. To continue with our multiple linear regression model, we will use the following equation:

The null hypothesis will be 0 as in:

The alternate hypothesis is:

F stat and p value are 1580 and 0.0 respectively. We will reject the null hypothesis.

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

* *Based on the results of the overall F-test, is at least one of the predictors statistically significant in predicting the total number of wins in the season?*
* *What are the results of individual t-tests for the parameters of each predictor variable?*

*Is each of the predictor variables statistically significant based on its P-value? Use a 1% level of significance.*

* *Report and interpret the coefficient of determination.*
* *What is 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?*
* *What is the predicted total 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?*

The total wins in the regular season for a team averaging 75 points per game and 1350 elo is 16:

The total wins in the regular season for a team averaging 100 points per game and 1600 elo is 16:

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

Finally, we take a stab at analyzing point differential and how it relates to the wins.

The equation used for our model is:

The null is a test of the likelihood of the statement presented to be true before we begin our analysis. The statistical notation for this is:

The alternative hypothesis is the idea of determining whether to accept or reject the null hypothesis. The statistical notation for this is:

The level of significance for our analysis is 5%.

For our analysis, the test statistic is F = 1449 and the P-value is 0

Our conclusion will be to reject the null hypothesis and accept the alternative. There is sufficient evidence that at least one of the predictors is statistically significant when comparing to number of wins in a season for a team.

Table 3: Hypothesis Test for Overall F-Test

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

R-squared: 0.876

The total wins in the regular season for a team averaging 75 points per game and 1350 elo with the average point differential of -5 and relative skill differential of -30 is 26:

The total wins in the regular season for a team averaging 100 points per game and 1600 elo with the average point differential of +5 and relative skill differential of +95 is 52:

## 8. Conclusion

After reviewing the data, you can immediately see the correlation between high points, elo, etc. with wins. We view the data in detail with scatter plots and see how the data either negatively or positively impacts a win for a basketball team. Once we have our data visualization and know what type of data to analyze, we went ahead and built our predictive models and with our findings, we can see that with a higher point, higher relative skill will lead to a higher number of wins and even ending the season with up to 52 wins!

## 9. Citations

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