# MAT 243 Project Three Summary Report

Bradley Jackson

Bradley.Jackson2@snhu.edu

Southern New Hampshire University

The dataset for this analysis is the *FiveThirtyEight NBA Elo dataset* historical basketball data from Kaggle. I was tasked with analyzing performance patterns by the coach and management of my team. They asked for regression models that predict the number of wins in a regular game based on the performance metrics from the dataset. They will be used to make key decisions to improve the performance of the team. Analyses that I employed in this study were data visualizations (in the form of scatter plots), simple linear regression, and multiple regression models.

Average point differential between the team and their opponents in a regular season is represented by avg\_pts\_differential. In other terms, this is the difference in score between the team and their opponent. A negative value would signify that the opponent scored higher and a positive value indicates that our team scored higher. Average relative skill is represented by avg\_elo\_n, which considers several factors and is a summary of the skill level of a team based on the entire league’s skill.

Data visualization techniques are used to study the relationship trend between two variables by providing a different perspective on the data than raw numbers. Relationships between variables on a scatter plot are more easily interpreted by human eyes than raw data and can simplify explaining complicated abstractions. The absolute value of the correlation coefficient tells us the strength of correlation and the sign denotes the direction. My Pearson correlation coefficient is 0.4777 which is a moderate positive correlation. This means that there is evidence that total wins and average points are tied together. The p-value is zero, which is less than 0.01. If we were to perform a t-test on this data with a significance level of 1%, we would conclude that sufficient evidence that a positive linear relationship exists between total number of wins and the average points scored.

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 182.10  *\*Round off to 2 decimal places.* |
| P-value | 0.0000  *\*Round off to 4 decimal places.* |

Simple linear regression models are used to predict the response variable using a predictor variable by providing an equation based on the relationship between response and predictor variables. The equation for my model is . The null hypothesis for the F-test was that the beta parameter is not significantly different from zero (). The alternative hypothesis is , or that the beta parameter is significantly different from zero. Assuming a significance level of 5%, we can accept the null hypothesis and make the determination that a significant linear relationship exists between avg\_pts and total wins. Average scored points is a viable way of predicting the total number of wins in a regular season. For a team averaging 75 points per game we can expect that their number of wins in a regular season is 10. The predicted number of wins for a team averaging 90 points is 30.

The scatterplot and Pearson correlation coefficient tell us that average relative skill and total number of wins are directly related. Based on a significance level of 0.01 we can determine that there is a positive linear relationship between average relative skill and total number of wins because the p-value is zero.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00  *\*Round off to 2 decimal places.* |
| P-value | 0.0000  *\*Round off to 4 decimal places.* |

A multiple linear regression model is used to predict the response variable using predictor variables by determining the mathematical relationship between them and providing an equation to predict response variable values. The equation for this section is . The null hypothesis is that the slope will equal zero and the alternative is that it will not equal zero . The conclusion of this test based on a 5% level of significance is that we reject the hypothesis for both predictor variables. The individual f-tests for both predictor variables show that both are significant in the model and have a weighted relationship to the response variable. The coefficient of determination is 0.837. This means that 83.7% of the variance in total wins can be explained by the variance in average points and average skill level. 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 rounded down is 16. The predicted total number of wins in a regular season for a team averaging 100 points per game with a relative skill level of 1600 rounded down is 51.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449.00  *\*Round off to 2 decimal places.* |
| P-value | 0.0000  *\*Round off to 4 decimal places.* |

A multiple linear regression model is used to predict the response variable using predictor variables by determining the mathematical relationship between them and providing an equation to predict response variable values. The equation for this section is . The null hypothesis is that the slope will equal zero and the alternative is that it will not equal zero . The conclusion of this test based on a 5% level of significance is that we reject the hypothesis for both predictor variables. The p-values for all three predictor variables are zero. The individual f-tests for both predictor variables show that both are significant in the model and have a weighted relationship to the response variable. The coefficient of determination is 0.876. This means that 87.6% of the variance in total wins can be explained by the variance in average points and average skill level. 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, rounded down, is 20. The predicted 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, rounded down, 37.

Our coach and team management asked for regression models that predict the number of wins in a regular game based on the performance metrics from the dataset. In our analysis above, we created a model that predicted the total number of wins using average points scored, average relative skill, and average points differential. We showed that avg\_pts, avg\_elo\_n, and avg\_pts\_differential are significant variables in the measure of total\_wins. The importance of this model is that it can be used to make key decisions to improve the performance of the team.