# MAT 243 Project Three Summary Report

Joey L. McCrary II

Joey.mccrary@snhu.edu

Southern New Hampshire University

## 1. Introduction

In this study the aggregate data of the entire NBA league over the regular season will be evaluated to investigate correlations between total number of wins and performance metrics. Analyses will be performed via regression models that predict wins with reference to combinations of the performance metrics.

## 2. Data Preparation

Since the data set used here is derived from previous projects, new parameters have also been established and listed below:

total\_wins = Total number of wins in regular season

avg\_pts = Average points scored in regular season

avg\_elo\_n = Average relative skill (ELO) of each team in regular season

Relative skill is measured by the ELO rating, which includes final scores, game location, and game outcome relative to the probability of that outcome.

avg\_pts\_differential = Average point differential between team and opponents in regular season

Point differential is simply the amount of points between winning and losing team, i.e. how much a team wins by.

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

The Step 2 scatterplot helps to visualize the positive (upward) trend. As more points are scored more games are won, which makes sense except when either both teams have really good defense or both have lackluster offense. The correlation coefficient for this test is 0.4777, moderate strength (0.40 < │R│ ≤ 0.80). Strength of correlation is determined from the absolute value of the coefficient against a list of strength thresholds. Since, the coefficient is positive, it confirms the observation of an upward trend. In regression analysis, a positive correlation means that as the predictor variable(s) increases, so does the response variable. The p-value for this test is 0.0 and has statistical significance against most significance levels meaning that the correlation can be accepted as linear.

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

A simple linear regression model can be used to predict response variables using predictor coefficients derived the overall F-test. The derived linear regression equation for this test is Y = -85.5476 + 1.2849avg\_pts. Also from the F-test we get the test statistic and its p-value, see below.

Table 1: Hypothesis Test for the Overall F-Test

H0: β1 = 0 Ha: β1 ≠ 0 α = 0.01 obs = 618 R2 = 0.228

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 182.1 |
| P-value | 0.0 |

Since the p-value is less than the significance level (α), the null hypothesis can be rejected in favor of the alternative. This means that the regression model is linear and predict outcomes. When evaluated at avg\_pts = 75, the expected outcome is ~10 games. At avg\_pts = 90, the expected outcome is ~30 games. For evaluations, the outcomes make sense and correlate with the above scatterplot.

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

The scatterplot for total wins and average relative skill has a positive trend which can be confirmed visually and by the Pearson Correlation Coefficient, 0.9072. Similar to above, the positive trend indicates that’s as average skill increases, so does the total number of wins. The corresponding p-value, 0.0, is less than the level of significance level, α = 0.01, which means that the correlation coefficient is statistically significant.

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

Step 5 uses a multiple regression model, which operates similar to simple linear regression just with more predictor variables. The multiple regression equation for step 5 is Y = -152.5736 + 0.3497X1 + 0.1055X2.

Table 2: Hypothesis Test for the Overall F-Test

H0: β1 = β2 =0 Ha: β*i* ≠ 0 for *i* = 1, 2 α = 0.01 obs = 618 R2 = 0.837

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00 |
| P-value | 0.0 |

Since the p-value is less than the p-value, reject the null hypothesis in favor of the alternative that at least one coefficient is not equal to zero. For the individual t-tests, all p-values are less than the 1% significance level, meaning that all coefficients are statistically significant. The coefficient of determination, R2, meets the criteria for strong correlation (0.80 < │R│ ≤ 1.0) for this model.

Testing the model equation, evaluate at avg\_pts = 75 and avg\_elo\_n = 1350 which yields 16 wins. 16 wins makes sense and correlates with the previous scatterplots. In another test, evaluate at avg\_pts = 100 and avg\_elo\_n = 1600 which yields 203.77 and also visually checks out with the scatterplots.

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

The step 6 multiple regression model predicts the total wins just as the previous model with an additional predictor variable: average points differential. The equation for this regression model is Y = -35.8921 + 0.2406X1 + 0.0348X2 + 1.7621X3.

Table 3: Hypothesis Test for Overall F-Test

H0: β1 = β2 = β3 = 0 Ha: β*i* ≠ 0 for *i* = 1, 2,3 α = 0.01 obs = 618 R2 = 0.837

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449.0 |
| P-value | 0.0 |

The p-value is less than the 1% significance level which signals to reject the null in favor of the alternative, thus at least one coefficient is not equal to zero. In the individual t-tests, all p-values are also less than the 1% significance level, thus all coefficients are statically significant. The model has a strong correlation with the coefficient of determination meeting strong criteria. Testing the model equation, evaluate at avg\_pts = 75, avg\_elo\_n = 1350, and avg\_pts\_differential = -5 yielding 20 games and visually confirms with the scatterplots. In another test, evaluate at avg\_pts = 100, avg\_elo\_n = 1600, and avg\_pts\_differential = +5 yields 52 games which also is confirmed visually.

## 8. Conclusion

The results from this case study shows that the regression model accurately predicts the total wins for the regular season. The practical importance of these results show that as teams’ performance metrics increase total wins also increase resulting in a better season and hopefully play-offs…maybe a championship.