# MAT 243 Project Three Summary Report

Brandon

brandon.hobbs@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this analysis an attempt is made to calculate the number of games a team is expected to win from post-game metrics. The metrics in this report are based on NBA data from the 1995 - 2015 seasons. The main metrics used to compare the teams will be average points scored in a game and the Elo “…a simple measure of strength based on game-by-game results” (FiveThirtyEight, 2022). The higher an Elo rating the better a team has performed. Silver and Fischer-Baum (2015) have created this table, Figure 1 to illustrate the range of Elo values for the NBA.

Table

Description automatically generated

**Figure 1: NBA Elo Ratings Brackets (Silver and Fischer-Baum, 2015)**

Other metrics explored are average points differential (home team minus away team) and average Elo differential (home team Elo minus opponent Elo).

The following comparisons and analyses will be performed:

1. Number of Wins vs. Average Relative Skill (Elo)
   1. Scatterplot and Pearson Coefficient
   2. Simple, Single-Variable Model
2. Number of Wins vs. Average Points Scored
   1. Scatterplot and Pearson Coefficient
3. Number of Wins vs. Average Points Scored and Average Relative Skill
   1. Multi-Variable linear model
4. Number of Wins vs. Average Points Scored, Average Relative Skill, Average Points Differential and Average Relative Skill Differential
   1. Multi-Variable linear model

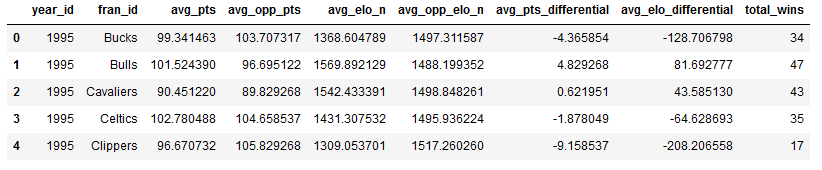
For each of these 4 analyses a linear regression will be computed and then evaluated for its fit and appropriateness.

## 2. Data Preparation

As mentioned above the number of wins is to be predicted from varying post-game metrics. The most obvious metrics explored is average points scored within a season by the home team (**avg\_pts**); after all, you win by scoring the most points. The next metric looks at the home team’s average points vs. the away team’s (**avg\_opp\_pts**) points differential (**avg\_pts\_differential**). This could show how greatly the winning team dominated its opponent – a positive differential means the home team dominated and, conversely, a negative means the opposing team dominated.

The relative score for the home team (**avg\_elo\_n**) was also looked at as a predictor of wins. This seems to be a natural predictor as the Elo score was designed to measure how well a team performs. The differential of home and away teams’ Elo (**avg\_elo\_differential)** was also investigated as it might be expected that the greater the positive differential the more likely a win is to occur.

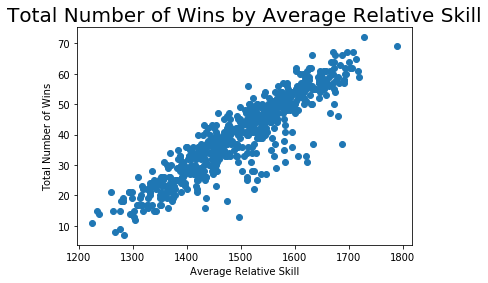
Figure 2 shows the first five lines of the data set and the variables available for the subsequent analyses.



**Figure 2: NBA Data from 1995-2015 and respective Variables used in Analyses**

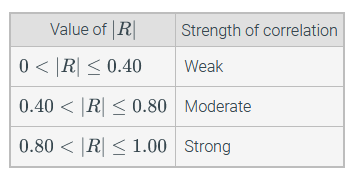
## 3. Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill

The first analysis performed was total wins vs. average relative skill. Plotting the data allows for a visual analysis of any trends, Figure 3 shows the scatterplot. One would instinctively expect that the higher the average Elo the higher the number of wins (this is how the Elo was designed), a positive correlation, and this is what the plot suggest.



**Figure 3: Scatterplot of Average Elo and Total Wins**

To determine how strongly wins is dependent on average Elo the Pearson correlation coefficient (R) was computed. A positive correlation between two variables means that as one variable increases, the other variable increases as well. A negative correlation between two variables means that as one variable increases, the other variable decreases. Figure 4 provides guidance on how to interpret the magnitude of this value.



**Figure 4: Pearson Correlation Coefficient Magnitude and its Interpretation**

The R value calculated for this fit was 0.9072. Confirming the anticipated positive correlation and the strong dependence between average Elo and total wins. Moreover, the P-value of ~0.000 further confirms the statistical relevance of this dependence even to 99% (α = 1%) since 0.000 << 0.01.

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

Figure 3 showed there was a positive correlation between average Elo and total wins and the R value confirmed the strength. Because the trend was statistically relevant, a linear model of the form:

was generated, below:

This model uses the predictor variable avg\_elo\_n to determine the response variable, total wins.

To determine if this model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variable and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = 0*

*Ha: β1 ≠ 0*

The null hypothesis states that *β1* is zero; meaning there is no correlation between avg\_elo\_n and total wins. The alternative states that *β1* isnot zero; meaning there is a correlation between avg\_elo\_n and total wins. This will be evaluated against an α of 1% or 99% confidence interval. Table 1 shows the F-test statistic and its associated P-value:

**Table 1: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2865.00 |
| P-value | 8.06e-234 |

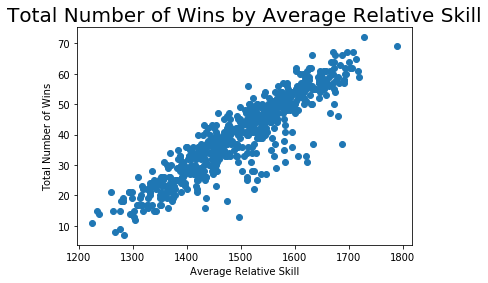
The P-value confirms that the null value may be rejected, 8.06e-234 << 0.01. Moreover, this further confirms that the model shown above is valid at the 99% confidence level.

Furthermore, this model has a coefficient of determination (*R2*) of 0.823. Meaning that 82.3% of total wins is explained by avg\_elo\_n.

This model, having been shown to be relevant, may then be used to predict wins. For example, a team with an average Elo score of 1550 or 1450 are shown below:

* 1550: -128.2475 + 0.1121 \* 1550 = 45 wins
* 1450: -128.2475 + 0.1121 \* 1450 = 34 wins

This is consistent with the scatterplot shown in Figure 3.



**Figure 5: Using Figure 3 to Illustrate the Linear Model Fits the Data**

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

The next analysis performed was total wins vs. average points scored. Average points scored is the indicator most people would assume has the greatest influence on wins – after all, a game is won by scoring the most points. The first analysis step was plotting the data to allow for a visual analysis of trends. Figure 6 shows a scatterplot of the data. One would probably expect that the higher the average score the higher the number of wins, a positive correlation, and this is what the plot suggests, if only loosely.

Chart, scatter chart

Description automatically generated

**Figure 6: Scatterplot of Points Scored and Total Wins**

To determine how strongly wins is dependent on average points scored the Pearson correlation coefficient (R) was computed. A positive correlation between two variables means that as one variable increases, the other variable increases as well. A negative correlation between two variables means that as one variable increases, the other variable decreases. Figure 4, from before, may be used again to provide guidance on how to interpret the magnitude of this value.

The R value calculated for this fit was 0.4777, confirming the anticipated positive correlation but, surprisingly, showing a much weaker dependence than Elo. This weaker correlation is seen in the scatterplot by the disperse nature of the points, i.e., the point cloud is much less linear than seen in Figure 3. However, the P-value of ~0.000 does confirm there is statistical relevance of this dependence even to 99% (α = 1%) since 0.000 << 0.01.

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

Figure 6 showed there was a positive correlation between average score and total wins, and the P-value confirmed the relevance. Figure 3 confirmed that the average Elo score was able to predict totals wins – so what happens if both metrics are used as predictor variables?

To answer this question a multi-variable linear model of the form:

and its regression value was created. The linear model was calculated as:

This model uses the predictor variables avg\_elo\_n and avg\_pts to determine the total wins.

As before, an F-test was run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 = 0*

*Ha: At least one βi ≠ 0 for i = 1 to 2*

The null hypothesis states that *β1* and *β2* are zero; meaning there is no correlation between avg\_elo\_n, avg\_pts and total wins. The alternative states that either, or both, *β1* or *β2* are not zero; meaning there is a correlation between avg\_elo\_n and/or avg\_pts and total wins. This will be evaluated against an α of 1% or a 99% confidence interval. Table 2 shows the F-Test statistic and its associated P-value:

**Table 2: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00 |
| P-value | 4.41e-243 |

The P-value confirms that the null value may be rejected, 4.41e-243 << 0.01; thus, at least one variable is linearly correlated to total wins. Moreover, this further confirms that the model shown above is valid at the 99% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. The t-test will have similar null hypothesis and alternative hypothesis for each predictor variable. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and total wins. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and total wins. Based on these hypotheses the P-values can be used to determine statistical relevance.

**Table 3: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| avg\_pts | 0.000 |
| avg\_elo\_n | 0.000 |

Both P-values are less than the 1% significance level, i.e., 0.000 << 0.01, therefore, both variables are shown to have a statistically relevant relationship.

Moreover, this muti-variable model has a coefficient of determination (*R2*) of 0.837. Meaning that 83.7% of total wins is explained by both avg\_pts and avg\_elo\_n.

This model, having been shown to be relevant, may then be used to predict wins. For example, a team with an average Elo score of 1350 and averaging 75 points or a team with an average Elo score of 1600 and averaging 100 points are shown below:

* 1350|75: -152.5736 + 0.3497 \* 75 + 1350 \* 0.1055 = 16 wins
* 1600|100: -152.5736 + 0.3497 \* 100 + 1600 \* 0.1055 = 51 wins

Again, this model confirms intuition, the more one scores and the higher a team’s average Elo the more likely the team is to win.

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

Lastly, a model including the idea of dominance was built. This dominance model would use the 2-variable model from the last analysis but add average points differential (**avg\_pts\_differential**) between each home team and their opponent and average relative skill differential (**avg\_elo\_differential**) between each home team and their opponent.

With the differential predictor variables added to the previous model it now becomes:

As before, an F-test and individual t-tests were conducted to evaluate the model and the individual predictor variables. The F-test null hypothesis (*H0*) and alternative hypothesis (*Ha*) are shown below:

*H0: β1 = β2 = β3 = β4 =0*

*Ha: At least one βi ≠ 0 for i = 1...4*

The null hypothesis states that *β1* through *β4* are zero; meaning there is no correlation between avg\_elo\_n, avg\_pts, avg\_pts\_differential, avg\_elo\_differential and total wins. The alternative hypothesis states that either, or all, *β1* through *β4* are not zero; meaning there is a correlation between the at least one of the predictor variables and total wins. This will be evaluated against an α of 1% or a 99% confidence interval. Table 4 shows the F-test statistic and its associated P-value:

**Table 4: Hypothesis Test for the Overall F-Test**

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

The P-value confirms that the null value may be rejected, 3.07e-278 << 0.01 – there is dependence between at least one of the predictor variables and total wins. Moreover, this further confirms that the model shown above is valid at the 99% confidence level.

Again, the F-test does not reveal if all predictor variables are relevant just that at least one is relevant. To determine which variables are relevant individual t-Tests are conducted on the variables.

**Table 5: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| avg\_pts | 0.000 |
| avg\_elo\_n | 0.442 |
| avg\_pts\_differential | 0.000 |
| avg\_elo\_differential | 0.004 |

All the P-values do not meet the 1% significance level. The P-value of avg\_elo\_n is much greater than 1%, i.e., 0.442 >> 0.01. This suggests that avg\_elo\_n is not relevant when used with the three other predictor variables.

The remaining P-values are less than the 1% significance level, i.e., 0.000 << 0.01, meaning they are shown to have a statistically relevant relationship.

However, this muti-variable model has a coefficient of determination (*R2*) of 0.878. Meaning that 87.8% of total wins is explained by avg\_pts, avg\_elo\_n, avg\_pts\_differential, and avg\_elo\_differential.

This model, having been shown to be relevant by the F-test, may be used to predict wins. To illustrate its usage 2 teams are considered. One, a team 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. Two, a team that is averaging 100 points per game with a relative skill level of 1600, average point differential of +5 and average relative skill differential of +95?

* Team 1: 34.5753 – 0.0134 \* 1350 + 0.2597 \* 75 + 1.6206 \* (-5) + 0.0525 \* (-30) = 26 wins
* Team 2: 34.5753 – 0.0134 \* 1600 + 0.2597 \* 100 + 1.6206 \* (5) + 0.0525 \* (95) = 52 wins

Even though the model is able to make predictions it is not optimum. As shown avg\_elo\_n is not statistically relevant at the 1% significance level. A new model ignoring this predictor variable may be computed using only avg\_pts, avg\_pts\_differential, and avg\_elo\_differential. This model is:

The F-test, Table 6, and individual t-test, Table 7, results are shown below – the null and alternative hypotheses would be the same as before.

**Table 6: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1470 |
| P-value | 9.89e-280 |

**Table 7: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| avg\_pts | 0.000 |
| avg\_pts\_differential | 0.000 |
| avg\_elo\_differential | 0.000 |

The P-values shown in Table 7, unlike Table 5, are well below the 1% significance level, i.e., 0.000 << 0.01, meaning all are statistically relevant. With this new model the total wins may be predicted, as before:

* Team 1: 14.9483 + 0.2552 \* 75 + 1.6281 \* (-5) + 0.0393 \* (-30) = 24 wins
* Team 2: 14.9483 + 0.2552 \* 100 + 1.6281 \* 5 + 0.0393 \* 95 = 52 wins

This model can explain 87.8% (*R2* = 0.878) of the variability in total wins with the chosen predictor variables. The predicted wins and the *R2* are similar to the previous model, but that is expected since the rejected variable had no statistical relevance.

## 8. Conclusion

Four different analyses were conducted to explore how the coaching staff might predict their season’s win rate. The first two analyses simply plotted the total wins vs. average Elo or average points. Both were found to be positively correlated – with average Elo being a much stronger correlation.

After these simple, one-variable, models the analysis moved to multi-variable linear regression models. The first model used average points scored and average relative skill to predict total wins. Both variables were found to be statistically relevant, and the model had an R2 value of 83.7%.

From there an additional 4-variable model was created using average points scored, average relative skill, average points differential, and average relative skill differential. Average relative skill was found to not be statistically relevant to this model, so it was removed and the model recomputed. This new, 3-variable model was able to explain 87.8% of the variability in total wins. Table 8 shows each model and their R2 values:

**Table 8: R2 values for Each of the Derived Models**

|  |  |
| --- | --- |
| **Model** | **R2** |
| avg\_elo\_n | 0.823 |
| avg\_elo\_n, and avg\_pts | 0.837 |
| avg\_elo\_n, avg\_pts, avg\_pts\_differential, and avg\_elo\_differential | 0.878 |
| avg\_pts, avg\_pts\_differential, and avg\_elo\_differential | 0.878 |

With each refinement of the model, 1-variable to 3-variable, the R2 value moves closer to 1. This is a welcome outcome as each new model is able to explain more variability in total wins.

Now that the optimum model, with available data, is known, the coaching team may predict their total wins based on post-game metrics. They may also use this model as an indicator if they are out- or under-performing previous seasons’ performance.

## 9. Citations

FiveThirtyEight. (2022, June 17). *The complete history of the NBA*. https://projects.fivethirtyeight.com/complete-history-of-the-nba

Hobbs, B. (2022). *MAT 243 project one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 243 project two summary report*. [Unpublished report]. SNHU.

Silver, N. and Fischer-Baum, R. (2015, May 21). *How we calculate NBA Elo ratings*. FiveThirtyEight. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/