MAT 243 Project Three Summary Report

Austin Niwa

Austin.niwa@snhu.edu

Southern New Hampshire University

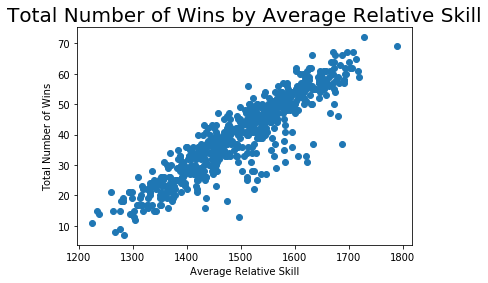
## 1. Introduction

The dataset I will be using in this analysis is the nba\_wins\_df data set as its referred to in the jupyter script or the FiveThirtyEight NBA Elo dataset from Kaggle as its referred to in the rubric. More specifically the average relative skill and total wins on this data set. I will be using this data set to interpret and predict the number of wins for a team in the regular season of the NBA for the coach and management of my team. I will be using this data set to run multiple linear regressions to show the correlations between the Average Relative Skill and the Total Number of Wins of the teams assessed. This will help the coach of my team predict how many games my team might win in a regular season based on metrics such as the average score, average relative skill, average points differential and average relative skill differential between the team and their opponents.(from step 6 of the jupyter script.)

## 2. Data Preparation

* If i had to explain the average points differential variable to someone who doesnt understand the data at all i would tell them it is the average of the difference between our team and the opponents they are facing over all of the wins in that league. For example if an opponent team scored 30 points and our team scored 40 points the differential in that case would be +10.
* What does the variable avg\_elo\_n represent? How would you explain it to someone who does not understand the data? While the average points differential describes the difference in points the average elo n variable shows the skill of our team relative to the skill of other teams in the NBA allowing for the prediction on which team will win based on heir skill. An example for this would be in the first part of the data. In the 5th game in the data there is an elo differential of -208.206558 between the clippers and the average of the opposition. Their point differential was also -9.158537 meaning they were less likely to score than the opposition.

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

* Data visualization techniques are used to find correlations between two variables by showing how they relate to each other. If both variables increase in the chart then they likely have a correlation. This can also identify outliers in the data.
* If one variable increases with the other it is a positive correlation if the opposite is true it is a negative correlation. The strength of a correlation can be found through the correlation coefficient by seeing if its less than or equal to 1 for a strong correlation, less than or equal .80 for a moderate correlation and, .40 for a weak correlation.
* 
* What do the scatterplot and the Pearson correlation coefficient tell you about the association between total number of wins and average relative skill? The scatterplot appears to show a positive correlation between the elo of a team and their number of wins. The correlation coefficient here is 0.9072 indicating a strong correlation between the elo of a team and the wins that a team has gotten.
* The p-value is 0.00 so the correlation coefficient is in fact, statistically significant since it is below the level of significance of 1%

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

* Linear regression models are used to find the response variable by using a regression line.

The equation for my model would likely be -128.2475 + 0.1121elo.

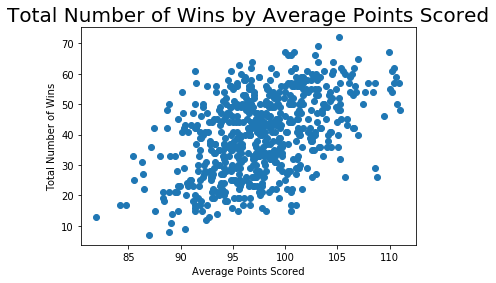
* 1. The null hypothesis here is stating that the average points scored is zero. It would imply there is no relationship between the two variables. HO = SKILL + WINS = 0 should the the notation for the null.
  2. Alternative hypothesis is that the average is /= to zero so a relationship does exist between the two variables. HA = Skills + Wins /= 0 should be the notation for the alternative hypothesis
  3. The level of significance for the linear regression is 0.01 or 1%

Table 1: Hypothesis Test for the Overall F-Test

| Statistic | Value |
| --- | --- |
| Test Statistic | 2865 |
| P-value | 8.06e-234 or 0.0000 |

* The p-value is lower than the level of significance of 0.01 making it statistically significant in predicting the number of wins in a given season. Wins and average relative skill have a strong correlation.
* The predicted number of wins for a team with an elo of 1550 is 45.5075 or 45 games. The predicted number of wins for a team with an elo of 1450 is 34.2975 or 34.

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

* 
* What do the scatterplot and the Pearson correlation coefficient tell you about the association between total number of wins and average points scored? The scatter plot appears to show a moderate correllation as it is going up on both sides but it looks scattered. The coefficient is 0.4777 indicating a moderate positive correlation as well.
* The correlation coefficient is 0.4777 so it is statistically significant as it incicates a moderate correlation and the p-value is 0.00 which is less than the given level of significance at 0.01 or 1%, so the null hypothesis is rejected.

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

* A multiple linear progression model is similar to a single regression model except it uses multiple predictor variables to predict the response variable and an equation for find the sum of the population equation.
* The equation for this model would be Y=(-152.5736) + 0.1055elo + 0.3497pts
* The null hypothesis is that the variables are not statistically significant and that the slope is 0. HO = Wins + Points scored= 0. This would make this model not useful in predicting wins for my team.
  1. The alternative hypothesis is that the variables are statistically significant and the slope is not 0. HO = Wins + Points scored /= 0. This would make this model useful in predicting wins for my team.
  2. The level of significance is 1% or 0.01.

Table 2: Hypothesis Test for the Overall F-Test

| Statistic | Value |
| --- | --- |
| Test Statistic | 1580 |
| P-value | 4.41 × 10-243 or 0.000  \*Round off to 4 decimal places. |

* 1. My conclusion based on the hypothesis test is that the test statistic is significant and the p value indicates that the null hypothesis would be rejected as it is lower than the level of significance of 0.01 or 1% at 0.000.
* Yes, atleast one of the predictors is statistically significant. According to the p-values both of which are 0.000 they are both good predictors for total wins in a given season.
* Both the t-tests for the avg\_elo\_n variable and the avg\_pts variable resulted in a p-value of 0.000 which is less than the level of significance of 1% so the null hypothesisd is rejected. There is a significant relationship between the response variable and the predictor variables.
* The coefficient of determination or r-squared is 0.837 or 83.7% meaning 83.7% of the variance in our data can be explained by the variables avg\_elo\_n and avg\_pts.
* The predicted number of wins for a team that is averaging 75 points per game with an elo of 1350 is about 16.789 or 16 wins. The number of wins for a team that is averaging 100 points with an elo of 1600 is 51.1964 or 51 wins.

## 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 progression model is similar to a single regression model except it uses multiple predictor variables to predict the response variable and an equation for find the sum of the population equation. The equation for my model should be Y = 34.5753 +(-0.0134elo) + (0.2597pts) + (1.6206ptsdiff) + (0.0525elodiff) HO = avg\_elo\_n + avg\_pts + avg\_pts\_differential + avg\_elo\_differential = 0 or that the variables are not statistically significant in regards to this regression model. HA = avg\_elo\_n + avg\_pts + avg\_pts\_differential + avg\_elo\_differential /= 0 or useful to this regression model. The level of significance is 0.01 or 1%

Table 3: Hypothesis Test for Overall F-Test

|  |  |
| --- | --- |
| Statistic | Value |
| Test Statistic | 1102 |
| P-value | 3.07e-278 or 0.0000 |

* Based on the results of the overall F-test, is at least one of the predictors statistically significant in predicting the number of wins in the season? based on the p-value of the f test which was essentially 0.000 at least one of the predictors was significant and if i take the others into account then the avg\_pts, avg\_pts\_differential,and avg\_elo differential have strong correlations with the response variable while avg\_elo\_n has a moderate correlation with it at 0.442.
* The results of the t-test were that avg\_elo\_n has a p-value of 0.182 indicating a strong correlation with the response variable. Avg\_pts had a p-value of 0.442 indicating a moderate correlation with the response variable. While, avg\_pts\_differential and avg\_elo\_differential both had a 0.000 p-value indicating a strong relationship between the two variables and the response variable as they were all under 1.
* My coefficient of determination or r-squared was 0.878 so 87.8% of the variability in my data was a result of the predictor variables.
* 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, average point differential of -5 and average relative skill differential of –30 is Y = 34.5753 +(-0.0134(1350)) + (0.2597(75)) + (1.6206(-5)) + (0.0525(-30)) = 26.2848 or 26
* The predicted total number of wins in a regular season for 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 is Y = 34.5753 +(-0.0134(1600)) + (0.2597(100)) + (1.6206(5)) + (0.0525(95)) = 52.1958 or 52

## 8. Conclusion

Describe the results of the statistical analyses clearly, using proper descriptions of statistical terms and concepts. Fully describe what these results mean for your scenario.

The use of the regression models allowed for the ability to see whether the predictor variables had any effect on the response variable and by how much if they did. The analyses that were conducted showed a significant correlation between the variables avg\_elo\_n, avg\_pts, avg\_pts\_differential, and avg\_elo\_differential. Some of the important data that was found through t he f-tests was how much these variables effect the total number of wins of my team. According to the first f-test the average elo has accounts for 0.823 or 82.3% of the variation of the total wins while the second f-test shows that avg\_elo and avg\_pts account for 0.837 or 83.7% of the variability. Then, taking all of our variables together avg\_elo\_n, avg\_pts, avg\_pts\_differential, and avg\_elo\_differential gives an r-squared of 0.878 or 87.8% of the variation in the total wins of the team.

## 9. Citations

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