# MAT 243 Project Three Summary Report

Mason Cabirac

Mason.cabirac@snhu.edu

Southern New Hampshire University

## 1. Introduction

Using a large set of historical data recorded from the performance of teams in the NBA, including the total number of wins in a regular season, the average number of points scored in a regular season, the average relative skill of each team in a regular season, the average point differential between team and opponents in a regular season and the average relative skill differential between team and opponent in a regular season, I will help predict the number of wins in a regular season.

This will be done using scatterplot graphing and correlation calculation of total wins by average relative skill and the same using simple linear regression. Scatterplot graphing and calculation of correlation will also be done of total wins by average points scored in a regular season. Multiple regression will be used to predict total wins by average points scored and relative skill, and again to predict total wins based on all other aforementioned variables.

These models are intended to help make key decisions regarding the Houston Rockets’ performance, by predicting the total number of wins for a team in a regular season based on performance metrics.

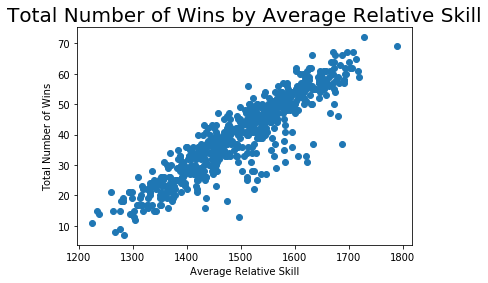
## 2. Data Preparation

The variable name avg\_pts\_differential represents the average point differential between a team and their opponents in a regular season. The average is found for every difference in score for a team and their opponent. The variable name avg\_elo\_n represents the average relative skill of each team in a regular season.

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

Data visualization techniques are used to study relationship trends between two variables by showing how they are related, whether they are positive, negative, or neutral. For instance, by a scatterplot, it may be shown that heavier cars are less fuel efficient, having a negative trend.

If the correlation coefficient is negative, the direction of association between two variables is negative. If the coefficient is positive, this direction is positive. Zero is neutral. The absolute value of the coefficient may not be more than 1. If the absolute value is very close to 1, the correlation is strong, meaning that its implication seldom fails. If it is close to 0 the correlation is weak, meaning there is not a clear implication.



The scatterplot shows a clear positive trend. The Pearson Correlation Coefficient is 0.9072. This shows that the trend is strong. A higher average skill will be very likely to result in more wins. The level of significance is 1% or 0.01. Our P-value is 0. Since the P-value is less than the level of significance, the correlation coefficient is statistically significant.

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

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2865.00  *\*Round off to 2 decimal places.* |
| P-value | 8.0600-234  *\*Round off to 4 decimal places.* |

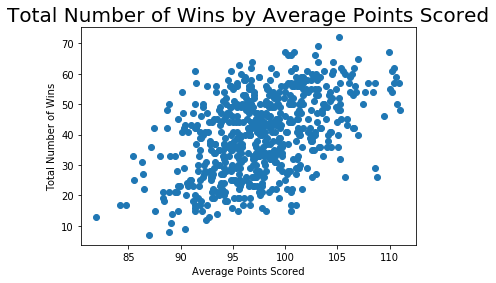
A simple linear regression model is used to predict the response variable using the predictor variable by assigning an intercept to the equation and a coefficient (whose value determines how much its predictor variable affects the response) to that predictor variable, the latter term of which is added to that intercept. The result of substituting numerical values for variables, then reduced, is the predicted response.

The equation for this model is Y = –128.2475 + 0.1121 \* (X1), where Y is the total number of wins in a regular season and X1 is the average relative skill of each team in a regular season.

The null hypothesis for the overall F-test is that the predictor variable is statistically significant, meaning that it does not relate to Y, the total number of wins in a regular season, or in statistical notation, H0: β1 = 0. The alternative hypothesis is that the predictor variable is statistically significant, meaning that it does relate to Y, or in statistical notation, Ha: βi ≠ 0. The level of significance is not requested in my instructions, as far as I can see. However, the P-value at an infinitesimal value, 8.06-234, is less than the smallest significance level I have ever seen, 1% or 0.01, as the result would be the same at any acceptable level. Thus, I can only concede to say that the null hypothesis must be rejected. The predictor variable is statistically significant. The average relative skill can be used to predict the total number of wins in the regular season.

The predicted total number of wins in a regular season for a team that has an average relative skill of 1550 is 45 wins. For an average relative skill of 1450, the prediction is 34. Each prediction has been rounded down to the nearest integer.

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



The Pearson correlation coefficient is 0.4777, which represents a positive trend of a moderate strength. The total number of wins has some basis of prediction in the average number of points scored, but not so much as the average relative skill.

The P-value of this correlation coefficient is zero, so the coefficient is statistically significant based on a 1% or 0.01 level of significance.

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

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00  *\*Round off to 2 decimal places.* |
| P-value | 4.4100-243  *\*Round off to 4 decimal places.* |

In general, a multiple linear regression model is used to predict the response variable using predictor variables by assigning a coefficient to each variable, the terms of which are added together and with an intercept, the result of which is the predicted response.

The equation for this model is Y = -152.5736 + 0.3497 \* (X1) + 0.1055 \* (X2), where Y is the total number of wins, X1 is the average number of points scored in a regular season and X2 is the average relative skill of each team in a regular season.

The null hypothesis for the overall F-test is that no predictor variables are statistically significant, meaning that they do not relate to Y, the total number of wins in a regular season, or in statistical notation, H0: β1 = β2 = ... = βn = 0. The alternative hypothesis is that one or more predictor variables are statistically significant, meaning that they do relate to Y, or in statistical notation, Ha: least one βi ≠ 0 for i = 1, 2, ..., n. Again, I am not given a level of significance to determine the results with and since I have another infinitesimal as a P-value (4.41-243), I will go ahead and assume 1% or 0.01. Based on the results of the overall F-test, at least one of the predictors are statistically significant in predicting the total number of wins in the season.

The individual t-tests show that, based on a 1% or 0.01 significance level that each of the predictor variables are statistically significant, as they all have a P-value of 0.

The coefficient of determination is 0.837, which shows that the model closely predicts the actual number of total wins based on the predictor variables. The model predicts that the total number of wins is 16 for a team that averages 75 points per game with a relative skill level of 1350. For 100 points per game and an average relative skill level of 1600, the total number of wins are predicted to be 51. For both predictions, the total number of wins are rounded down.

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

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1102.00  *\*Round off to 2 decimal places.* |
| P-value | 3.0700-278  *\*Round off to 4 decimal places.* |

The equation for this multiple linear regression model is Y = 34.5753 + 0.2597 \* (X1) + -0.0134 \* (X2) + 1.6206 \* (X3) + 0.0525 \* (X4), where X1 is the average number of points scored, X2 is the average skill level, X3 is the average points differential, and X4 is the average skill level differential.

The null hypothesis for the overall F-test is that no predictor variables are statistically significant, meaning that they do not relate to Y, the total number of wins in a regular season, or in statistical notation, H0: β1 = β2 = ... = βn = 0. The alternative hypothesis is that one or more predictor variables are statistically significant, meaning that they do relate to Y, or in statistical notation, Ha: least one βi ≠ 0 for i = 1, 2, ..., n. Again, having an infinitesimal P-value (3.07-278), and no suggested level of significance, which I will again assume to be 1% or 0.01, the null hypothesis must be rejected and the model is shown to have at least one predictor variable that is statistically significant in predicting the number of wins in the season.

Individual t-tests show, with a 1% level of significance, for the intercept and the average relative skill, statistical insignificance. For the average points, average points differential and skill level differential, statistical significance is shown.

The coefficient of determination is 0.878, which shows the model accurately predicts the total number of wins. 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, the total number of wins in a regular season is predicted to be 26. 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, the total number of wins in a regular season is predicted to be 52.

## 8. Conclusion

* It is clearly shown by the first scatterplot that there is a positive linear correlation, where the total number of wins in a season can be predicted by average relative skill. By the first linear regression model, we can accurately predict the results of this relationship in terms of a single value of average relative skill.
* The second scatterplot shows there is not such a clear linear correlation for the total number of wins and average points scored. However, if both the average points scored and average relative skill are factors in a multiple regression model, there is a more accurate model than the simple regression model of before.
* The second multiple regression model, which factors in all variables to predict the total wins, there is an even closer degree of accuracy, despite some variables proving statistically insignificant. The practical importance of these analyses will be that wins can be predicted if a team’s performance can be held to a standard.