# MAT 243 Project Three Summary Report

Babatunde Ali-Brown
babatunde.ali-brown@snhu.edu
Southern New Hampshire University

## 1. Introduction

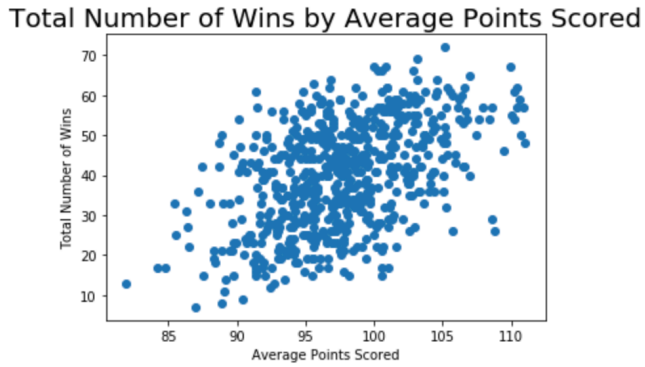
This analysis will be exploring a large set of historical data detailing the performance of teams in a basketball regular season between 1995 and 2015. The result of this analysis will be used to predict the total number of wins a team needs within a regular season based on key performance metrics. This analysis will be conducted using linear regression models.

## 2. Data Preparation

The key performance metric representing the response variable for this analysis is **total_wins**(total wins in a regular season), variables such as **avg_pts**(average points scored in a regular season), **avg_elo_n**(average relative skill of each team in a regular season), and **avg_pts_differential**(average points differential between a team and its opponent in a regular season) are used as predictor variables.

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

In general, data visualization techniques are used to visually perceive the kind of correlation that exists between two variables, also the direction and strength of such correlation. A negative correlation coefficient indicates a negative correlation while a positive correlation coefficient indicates a positive correlation. Correlation coefficients closer to ±1 is considered strong, those closer to 0 is considered weak, and those around ±0.5 is considered moderate.




```
Correlation between Average Points Scored and the Total Number of Wins
Pearson Correlation Coefficient = 0.4777
P-value = 0.0
```

According to the generated scatterplot and the Pearson correlation coefficient, the association between total number of wins and average points scored is positive in direction but only moderate in strength. At a 1% level of significance (alpha=0.01), the P-value of 0.0 being less than 0.01 provides evidence to reject the null hypothesis that no relationship exists between total wins and average points scored. Therefore, the correlation coefficient is statistically significant.

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

In general, a simple linear regression model is used to determine intercept and slope coefficients describing the kind of linear relationship between the response and predictor variable in the model. These coefficients are applied to calculate a value of the response variable based on a hypothesized value of the predictor variable. According to the generated simple regression model, the equation of linear regression between total wins and average points scored is **$Y = -85.5476 + 1.2849X$**.

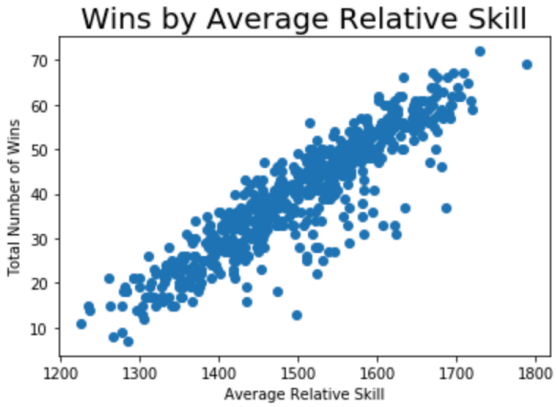
The null hypothesis $H_0 : \beta_1 = 0$ is that no relationship exists between total wins and average points scored, and the alternative hypothesis $H_a : \beta_1 \neq 0$ is that a relationship exists between total wins and average points scored. With a 1% level of significance (alpha=0.01), the overall F-Test returned the following statistics.

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
|---|---|
| Test Statistic | 182.10 |
| P-value | $-1.5200*10^{-36}$ |

With a P-value less than 0.01, there's enough evidence to reject the null hypothesis and accept the alternative hypothesis: therefore, average points scored can be used to predict total number of wins in the regular season. Using the equation of regression, a team averaging 75 points per game can be predicted to win 10 games (**$-85.5476 + 1.2849(75) = 10.8199$**). For a team averaging 90 points, it can be predicted to win 30 games (**$-85.5476 + 1.2849(90) = 30.0934$**).

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




```
Correlation between Average Relative Skill and Total Number of Wins
Pearson Correlation Coefficient = 0.9072
P-value = 0.0
```

The scatterplot and the Pearson correlation coefficient indicate a quite strong positive correlation between total number of wins and average relative skill. With a P-value less than the level of

significance (alpha=0.01), the null hypothesis that no relationship exists between total wins and average relative skill can be sufficiently rejected. Therefore, the generated correlation coefficient is statistically significant.

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

In general, multiple linear regression model considers more than one predictor variable, unlike simple linear regression. The multiple linear request tests linear relationship between the response variable and at least one of the several predictor variables. The equation of multiple regression between total wins (Y) as a response variable and both average points scored ($X_1$) and average relative skill ($X_2$) as predictor variables is $\mathbf{-152.5736 + 0.3497X_1 + 0.1055X_2}$.

The null hypothesis $\boldsymbol{H_0 : \beta_1 = \beta_2 = 0}$ is that no relationship exists between total wins and average points scored or average relative skill; and the alternative hypothesis $\boldsymbol{H_a}$ **: At least one** $\boldsymbol{\beta_i \neq 0}$ **(for** $\boldsymbol{i = 1, 2}$**)** is that a relationship exists between total wins and either average points scored or average relative skill. With a 1% level of significance (alpha=0.01), the overall F-Test returned the following statistics.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
|---|---|
| Test Statistic | 1580.00 |
| P-value | $4.4100*10^{-243}$ |

Based on the result of the overall F-test, a P-value less than 0.01 provides a ground to accept that either average points or average relative skill is statistically significant in predicting the total number of wins in the season.

The individual t-test statistic for average points is 7.297, and that of average relative skill is 47.952. Using a 1% level of significance, both individual t-tests having a P-value of 0.0 (less than 0.01) indicate that both average points and average relative skill are statistically significant in predicting the total number of wins in the season. The coefficient of determination (R-squared) for the multiple regression model is 0.837, meaning that at least 83.7% of the variance in total win is due to the variance in average points and average relative skill.

A team averaging 75 points with an average relative skill of 1350 can be predicted to win 16 games, and one averaging 100 points with an average relative skill of 1600 can be predicted to win 51 games.

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

The equation of multiple regression between total wins (Y) as a response variable and average points scored ($X_1$) and average relative skill ($X_2$), and average point differential ($X_3$) as predictor variables is $\mathbf{-35.8921 + 0.2406X_1 + 0.0348X_2 + 1.7621X_3}$.

The null hypothesis $\boldsymbol{H_0 : \beta_1 = \beta_2 = \beta_3 = 0}$ is that no relationship exists between total wins and all three predictors (average points scored, average relative skill, and average point differential); and the alternative hypothesis $\boldsymbol{H_a}$ **: At least one** $\boldsymbol{\beta_i \neq 0}$ **(for** $\boldsymbol{i = 1, 2, 3}$**)** is that a relationship exists between total wins and either average points scored, average relative skill, or average point differential. With a 1% level of significance (alpha=0.01), the overall F-Test returned the following statistics.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
|---|---|
| Test Statistic | 1449.00 |
| P-value | $5.0300*10^{-278}$ |

Based on the result of the overall F-test, a P-value less than 0.01 provides a ground to accept that either average points, average relative skill, or average point differential is statistically significant in predicting the total number of wins in the season.

The individual t-test statistic for average points is 5.657, that of average relative skill is 6.421, and that of average point differential is 13.928. Using a 1% level of significance, all three individual t-tests having a P-value of 0.0 (less than 0.01) indicate that average points, average relative skill, and average point differential are statistically significant in predicting the total number of wins in the season. The coefficient of determination (R-squared) for the multiple regression model is 0.876, meaning that at least 87.6% of the variance in total win is due to the variance in average points, average relative skill, and average point differential.

A team averaging 75 points with an average relative skill of 1350 and an average point differential of -5 can be predicted to win 20 games, and one averaging 100 points with an average relative skill of 1600 and an average point differential of 5 can be predicted to win 52 games.

## 8. Conclusion

This linear regression analysis helped to determine which of the available performance metrics in the data set can be used as predictor values to predict total numbers of games won. The analysis revealed that a team's average point scored, average relative skill, and average point differential all have a positive correlation with the total number of games won by the team. It can therefore be concluded that for a team to improve its performance in a regular basketball season it needs to increase not only its average point scored, but also its average relative skill, and average point differential.