# Summary Report Simple Linear Regression and Multiple Regression

Ryan Hatch

ryan@rshatch.com

April 20th, 2024

## 1. Introduction

This project consists of an in-depth statistical analysis of the NBA team performance data from 1995 to 2015. The objective is to develop regression models that help to predict total season wins for teams based on several performance indicators like the average points scored, team skill level, and point differentials against other teams and opponents.

The dataset, ‘**nba\_wins\_df**,’ includes essential variables like ‘**total\_wins’** (the total number of wins per season), ‘**avg\_pts’** (the average points scored per game), ‘**avg\_elo\_n’** (the average team skill level), ‘**avg\_pts\_differential’** (the average points difference per game against opponents), and ‘**avg\_elo\_differential’** (the average skill difference with opponents).

Through multiple regression analyses, the objective in this analysis is to identify the predictors that can cause a largest impact on the success of the team, which is mainly defined by their win totals. The discoveries from this summary report are intended to offer valuable insights to coaches and management to make strategic decisions that will improve the team’s performance in upcoming seasons.

## 2. Data Preparation

**Average Points Differential / ‘avg\_pts\_differential’:** This metric captures the average point difference between the team and its opponents throughout a regular season. A positive average points differential can help to prove that the team consistently outscored their opponents, showing a strong offense strategy. None the less, a negative differential point to the team scoring fewer points than its opponents, point out areas where either offensive or defensive strategies or skills needs to be revised and practiced. This variable is necessary to gauge the overall performance and strength of a team’s games during the season.

**Average Elo Rating / ‘avg\_elo\_n’:** The Elo rating system shows the relative skill levels of teams or players in direct competitions. The variable 'avg\_elo\_n' refers to a team's average Elo rating over the season. Teams with higher Elo ratings are typically more dominant and are expected to perform well against their competitors, while lower ratings suggest a lower amount of strength. This system is the primary way of gauging a team's overall skill and performance throughout the season.

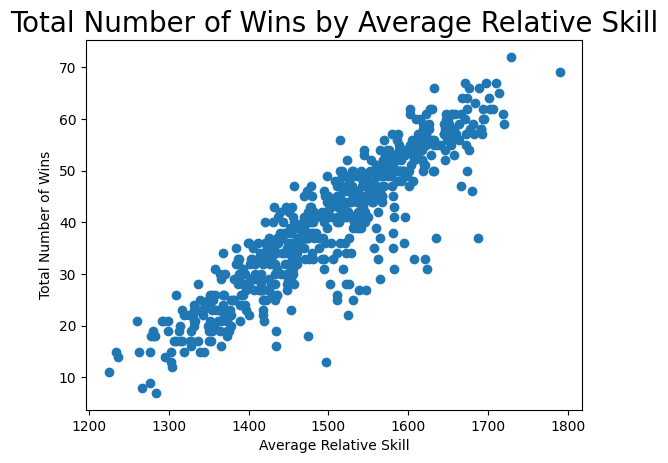
## 3. Simple Linear Regression: Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill

To explore how the total number of wins correlates with the average relative skill of NBA teams, I created a scatterplot. This type of graph is used in data analysis to visually examine the connections between variables. It also helps the analyst in order to spot trends, patterns, and possible correlations at a glance.

I assessed the relationship between team wins and skill levels using the Pearson correlation coefficient. This measure provides a value ranging from -1 to +1, where numbers close to +1 indicate a strong positive correlation. In other words, as the average skill level rises, the number of wins increase also. A result closer to -1 suggests the opposite.

This analysis showed a positive correlation, shown by the upward trend in the scatterplot. The Pearson coefficient was calculated at 0.9072, which shows a strong positive link between the variables.

To ensure this relationship wasn't just a fluke, I tested its statistical significance with a p-value. The p-value helped to determine the likelihood of my results under the assumption that there was no underlying relationship between the variables. In this assessment, a p-value of 0.0, which is significantly below the 1% threshold, confirmed that the observed correlations were indeed significant and not a fluke.



## 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 | 2865.0 |
| P-value | <0.0001 |

I developed a simple linear regression model to predict the total number of wins for NBA teams in a regular season, using average relative skill as the main predictor. This approach helped me to simplify and understanding how variations in skill levels influence the outcome of the games.

The model's equation:

**total\_wins=−128.2475+0.1121×avg\_elo\_ntotal\_wins=−128.2475+0.1121×avg\_elo\_n**

* **Intercept (-128.2475):** This represents the starting point of wins if the average relative skill was zero.
* **Coefficient (0.1121):** This indicates that each unit increase in average relative skill predicts an increase of ~ 0.1121 wins.

When testing the hypothesis, I conducted a F-test to evaluate the model's effectiveness:

* **Null Hypothesis:** No significant relationship existed between average relative skill and total wins.
* **Alternative Hypothesis:** A significant relationship doesnt exist.

With a significance level of 1%, the F-test resulted with:

* **F-statistic:** 2865.0
* **P-value:** <0.0001

The very low p-value caused me to reject the null hypothesis, which proved to me that the average relative skill significantly influenced the number of wins.

**Making Predictions**

I used the regression model in order to make specific predictions for teams based on their skill levels:

* **For a skill level of 1550:** -128.2475 + 0.1121 \* 1550 = **~** **45** wins.
* **For a skill level of 1450:** -128.2475 + 0.1121 \* 1450 = **~ 35** wins.

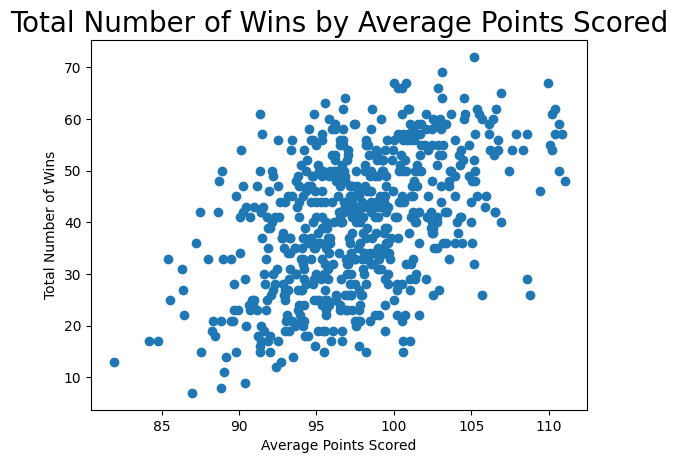
The analysis confirmed that a team's skill level is a strong predictor to their success in terms of wins. This insight is crucial for team management and utilizing different coaching strategies, emphasizing how crucial skill development is in order to achieve better outcomes during the seasons.

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

A scatterplot is a visual representation of the relationship between two variables. It helps to visually display and analyze the nature of the correlation (whether it is positive, negative, or has no correlation) and whether the relationship appears to be linear or non-linear. In this scatterplot you can visually access each team's average points scored against their total number of wins in a season.

The Pearson correlation coefficient indicates the strength and direction of the linear relationship between two variables. In this scenario, the coefficient is 0.4777, which shows a moderate positive correlation. This means that the teams that score more points on average tend to have a higher number of wins, but the relationship is not strong enough to suggest that average points scored are the only factor influencing the total number of wins.

The statistical significance of the correlation is determined by the P-value, which tests the null hypothesis that the true correlation coefficient is zero, or no association. A P-value of 0.0, when compared to the 1% level of significance, shows that the correlation between average points scored and the total number of wins is statistically significant. Therefore, I came to the conclusion that the association is not due to random chance.



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

Multiple linear regression models extend simple linear regression by incorporating more than one predictor variable. This let me see how different factors work together to affect one main result, which is the response variable. In this scenario, the model is used to predict the total number of regular season wins (‘**total\_wins’**) based on a team's average points scored (‘**avg\_pts’**) and the average relative skill (‘**avg\_elo\_n’**).

The regression equation that was derived from the model:

**total\_wins = -152.5736 + 0.1055 \* avg\_elo\_n + 0.3497 \* avg\_pts**

The overall F-test evaluates whether at least one predictor variable in the model has a non-zero coefficient.

* **Null Hypothesis (H0​):** All coefficients are equal to zero with no effect.
* **Alternative Hypothesis (HA​):** At least one coefficient is not equal to zero, where at least one variable has an effect.

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.0 |
| P-value | <0.0001 |

Since the p-value is less than 0.01, I rejected the null hypothesis, which helped to prove that at least one predictor variable significantly contributed to the model.

**Individual t-tests for Predictors**

* **Average Relative Skill (‘avg\_elo\_n’)**: With a p-value of <0.0001, this predictor is statistically significant.
* **Average Points Scored (‘avg\_pts’)**: Also, with a p-value of <0.0001, this predictor is statistically significant.

*Both predictors pass the 1% level of significance, which proves how important they are in predicting the total number of wins.*

The R-squared, or coefficient of determination, is 0.837. This means ~ 83.7% of the changes in the total wins can be explained by the model used. Since the value is pretty high, it shows that the model used fits the data very well.

**Predictions:**

The total number of wins for a team averaging **75** points per game with a relative skill level of **1350**: Predicted Wins = −152.5736 + (0.1055 × 1350) + (0.3497 × 75) = **16.0789**

The total number of wins for a team averaging **100** points per game with an average relative skill level of **1600**: Predicted Wins = −152.5736 + (0.1055 × 1600) + (0.3497 × 100) = **51.1964**

In conclusion, the multiple regression analysis indicated that both the average relative skill and the average points scored are significant predictors of the total number of wins. This model provided itself as a powerful method for the team's management to predict and strategize for the regular season's outcomes.

## 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 regression model predicts a response variable based on multiple predictor variables. It assumes a linear relationship between the predictors and the response. The model can be used for both prediction and to help understand the relationship, as well as how any changes associated between both the response and the predictors affect eachother.

The overall F-test evaluates whether at least one of the predictors in the model is statistically significant in predicting the response variable.

* **Null Hypothesis (H₀):** All the regression coefficients are equal to zero except the intercept. There is no linear relationship between the predictors and the response.
* **Alternative Hypothesis (H₁):** At least one of the regression coefficients is not zero. There is a linear relationship between at least one predictor and the response.
* **Level of Significance:** For this model, I used a 1% level of significance, or noted as the level of significance being at 0.01.
* **The P-value** is returned as < 0.0001 since it is such a small number.  
    
  In conclusion, Since the P-value is less than the level of significance, I rejected the null hypothesis. This helped me to prove that at least one predictor variable is significantly associated with the total number of wins.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1377.00 |
| P-value | <0.0001 |

Based on the results, at least one predictor is statistically significant in predicting the number of wins. Conducting individual t-tests also helped me be able to confirm which predictors were indeed significant.

The coefficient of determination, R-squared, is 0.871. This means that ~ 87.1% of the variability in the total number of wins can be explained by the model's predictors.

**Predicted Wins:**

For a team 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 predicted total number of wins were calculated using this regression equation:

**Predicted Wins = 38.3929 + (0.0010 × 1350) + (1.8045 × (−5)) + (0.0358 × (−30))**

1. Predicted Wins = 38.3929 + (0.0010 × 1350) + (1.8045 × (−5)) + (0.0358 × (−30))
2. Predicted Wins = 38.3929 + 1.35 − 9.0225 − 1.074
3. Predicted Wins = **29.6464**

For a team 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 predicted total number of wins are:

**Predicted Wins = 38.3929 + (0.0010×1600) + (1.8045×5) + (0.0358×95)**

1. Predicted Wins = 38.3929 + (0.0010×1600) + (1.8045×5) + (0.0358×95)
2. Predicted Wins = 38.3929 + 1.6 + 9.0225 + 3.401
3. Predicted Wins = **52.4169**

Therefore, the predicted wins for scenario 1 came out to **~ 29.65** games and the predicted wins for scenario 2 resulted with **~ 52.42** games.

## 8. Conclusion

The analysis that was conducted with the NBA performance data helped to show me some valueble insights into what necessarily can be done to drive a team to success. I used simple and multiple linear regression models to find out that the average points scored and the skill levels both have a lot of value as strong indicators with crucial roles that help to determine how many games a team might win.

I also found a clear link between a team's skill level, measured by the Elo rating system, and their win record. The high R-squared value from the graphs and models help to confirm that a large portion of a team's win rate can be explained by these factors.

The multiple regression models helped to see how complicated winning a game can be and the effort it takes for a team to also win a season. In order to win, teams can not just focus on the basics, like skills, but to look at how many points a team can score, and how as a collective, the team can stack up those scores against their oponents rapidly. These insights showed to me that looking at several factors together is imperative in order to accurately predict and plan for a team’s success.

This information is very valuable for the team managers and coaches. For understanding these connections can drastically help those in charge to make better decisions about trades, training, and game plans and strategies in order to optimize the team performance.   
The coaches could use the information from this analysis report to adjust their training programs and focus on areas that actually will help in winning games, since the information is backed by factual data analysis and not off of a potentially bias or detrimental process based from personal experience as a coach.

These insights are also useful in order to set realistic goals and to make predictions about the teams performance in the future. This summary report can even help in building more sophisticated models that take more variables into account, for example, player injuries, team morale and character, and even economic conditions.

In conclusion, this analysis report absolutely shows how powerful statistical tools can be in turning raw data into useful strategies and information. With the right data, the coaches and leaders can all individually make informed decisions for their team in order develop a higher win rate; leading to more scores per game and winning championships each season.