# MAT 243 Project Three Summary Report

Carlos Miguel Bracho Ortega

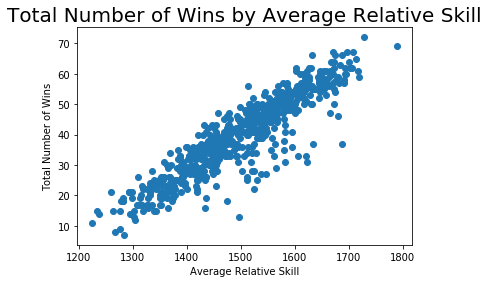
[carlos.bracho@snhu.edu](mailto:carlos.bracho@snhu.edu)

Southern New Hampshire University

The primary objective of this report is to develop a regression model that can predict the number of wins the Golden State Warriors or any other team in the NBA will achieve in a regular season based on the performance metrics included in a dataset. The dataset I will be exploring contains historical data with various performance metrics for the Warriors and their opponents in the NBA. The results of this report will help the team make strategic decisions. The coach and management team can use these predictions to create training programs, adjust game strategies, and optimize player rotations. Additionally, they can use this information to identify key indicators that can help them win more games. Finally, I will run a series of statistical analyses focusing on both simple and multiple regression models to analyze the total number of wins using the average relative skill, average points scored, average points differential, and average relative skill differential. These analyses will help the coach and management team understand the impact of various metrics on the Warriors' performance.

In this report, we'll be exploring several important variables. One of these is **avg\_pts\_differential**, which represents the average difference in points scored between the Golden State Warriors and their opponents during a regular season. In other words, it's simply a measure of how much, on average, the Warriors outscored their opponents. Another important variable is **avg\_elo\_n**. This metric represents the average relative skill level of the Golden State Warriors throughout a regular season. It's estimated based on the Warriors' performance in past games. In simple terms, we can say that the average relative skill is how good a team is in comparison to other teams.

In general, data visualization techniques, like scatterplots, are used to visually represent the relationship between two variables. By examining the patterns on these graphs, we can determine the nature of the relationship between two or more variables. For example, we can see if there is a positive relationship between variables by looking at the graph's direction. Moreover, the correlation coefficient (**R**) tells us the strength and direction of the relationship between two variables. We can also determine the direction of the relationship by looking at the correlation coefficient value. A positive value indicates a positive relationship, and vice versa. Additionally, this value tells us the strength of the relationship. A value close to 1 or -1 indicates a strong relationship, while a value close to 0 indicates a weak or no relationship between the variables.



This Scatterplot shows a strong relationship between the total number of wins and the average relative skill. As the average relative skill increases, the total number of wins also tends to increase. In addition, I calculated the Pearson correlation coefficient (**R**) to be 0.9072 (**R = 0.9072**). This value is close to 1, meaning that there is a very strong positive relationship between the variables total number of wins and average relative skill. Furthermore, to determine if the correlation coefficient is statistically significant, I calculated the p-value (**p**) and used a significance level (**α**) of 1% (**α = 0.01**). The p-value (**p**) was calculated as 0.0 (**p = 0.0**). In this case, the null hypothesis () states that there is no correlation between the total number of wins and average relative skill (), while the alternative hypothesis () suggests there is a correlation between these two variables (). Since the p-value is less than the level of significance (**p < α**), we can reject the null hypothesis. Therefore, we conclude that the correlation coefficient is statistically significant.

In general, a simple linear regression model predicts the response variable based on a single predictor variable. The equation for the linear equation is represented like this: , where **Y** is the response variable, is the intercept, and is the slope coefficient of the predictor variable **X**. This model can help make predictions of the response variable. In my case, I used this to create a model that can help predict the total number of wins using the average relative skill as the predictor variable. In Python, I created an OLS Regression Results table that displays the intercept and the coefficient of the predictor variable. After examining the table, I found the value for the intercept , which is **-128.2475,** and the slope coefficient , which is **0.1121**. After substituting these values in the linear equation, the model looks like this:

In this case, we performed an overall f-test, the null hypothesis () is that the coefficient of the average relative skill (**avg\_elo\_n**) is equal to zero. In other words, the average relative skill does not predict the total number of wins in the regular season. This can be represented as . The alternative hypothesis () states that the coefficient of the predictor variable is not equal to zero, meaning that the average relative skill does predict the total number of wins in the regular season. This can be represented as . The level of significance (**α**) used for this part of the report is 1% (**α = 0.01**).

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2865 |
| P-value | 0.0 |

From the table and the information above, we see that the p-value (**p**) is less than the level of significance (**α**). This can be represented as **p < α** or **0.0 < 0.01**. Therefore, we can reject the null hypothesis () because there is enough evidence to do so. This indicates that the average relative skill (**avg\_elo\_n**) is a significant predictor of the total number of wins in the regular season, meaning that a low p-value (**p**) indicates that the predictor variable is statistically significant.

If we want to predict the total number of wins for a team that has an average relative skill of 1550 (**avg\_elo\_n = 1550**), we just need to substitute this value in the equation. It would look like this:

**The result of this equation is 45 (rounded down) or total\_wins = 45**

Similarly, to predict the total number of wins for a team with an average relative skill of 1450 (**avg\_elo\_n = 1450**), we just need to follow the same procedure:

**The result of this equation is 34 (rounded down) or total\_wins = 34**

I also constructed a scatterplot for the total number of wins and average points scored. I did this because I am adding another predictor variable to the model. See the scatterplot below.

A blue dots on a white background

Description automatically generated

The scatterplot shows a positive association between the total number of wins and the average points scored. It has an upward trend from left to right. This indicates that as the average points scored increases, the total number of wins also tends to increase. The Pearson correlation coefficient (**R**) was calculated at 0.4777. The value suggests a moderate positive relationship between the average points scored and the total number of wins; however, we still need to verify if this correlation coefficient is statistically significant, using the p-value (**p**), which was calculated at 0.0.

The null hypothesis for this () is that the is no relationship between the average points scored and the total number of wins (). On the other hand, the alternative hypothesis () is that there is a relationship between the two variables (). The level of significance (**α**) used is 1% (0.01). Given that the p-value (**p = 0.0**) is less than the level of significance (**α = 0.01**), we can reject the null hypothesis and say that there is a linear relationship between the average points scored and the total number of wins. In simple words, the correlation coefficient (**R**) is statistically significant.

In general, a multiple linear regression model is used to predict the response variable by establishing a relationship between the response variable and several predictor variables. The equation is similar to the simple linear regression model, with the only difference being that there is more than one predictor variable: . **Y** represents the response variable, the **X** values are the predictor variables, is the intercept, and the rest of the ***β*** values are the coefficients that represent the relationship between the predictors and the response variable.

I translated this information to our case where we want to create a multiple regression model taking the total number of wins as the response variable, with average points scored and average relative skill as predictor variables. I created an OLS regression model using these variables to help identify the coefficient (***β***) of each predictor variable and the intercept (. Substituting those values into the multiple linear regression equation we have:

I then performed an overall F-test, where the null hypothesis () is that none of the predictor variables (**avg\_elo\_n** and **avg\_pts**) have any effect on the total number of wins. It is mathematically represented as . The alternative hypothesis () is that at least one of the predictor variables influences the total number of wins. This is represented as . The level of significance used in this report is 1% (**α = 0.01**).

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580 |
| P-value | 0.0 |

Given that the P-value (**p = 0.0**) is much smaller than the level of significance (**α = 0.01**), we reject the null hypothesis (). This can be represented as (**p < α**). This tells us that the model is significant. In other words, at least one of the predictor variables coefficients is significant when predicting the total number of wins in the season.

Since the overall F-test is significant, we can investigate each predictor variable parameter and perform a t-test to verify if all the variables are significant using a 1% level of significance (**α = 0.01**). For the average relative skill (**avg\_elo\_n**), the coefficient () is 0.1055 or , and its p-value (**p**) is 0.0 or **p = 0.0**. The null hypothesis () is that the parameter is not significant ()**,** and the alternativehypothesis () will be that this parameter is significant (). Since the p-value (**p**) is less than the level of significance (**α = 0.01**), we can reject the null hypothesis and say that the coefficient is significant in the model. Similarly, the average points scored (**avg\_points**) has a coefficient () of 0.3497 or , and the p-value (**p**) for this variable is also 0.0 or **p = 0.0**.The null and alternative hypotheses for this case are the same where I am evaluating if the average points score parameter is significant or not. Since the p-value (**p**) is also less than the level of significance (**α = 0.01**), we can also reject the null hypothesis. Therefore, this predictor variable coefficient is statistically significant as well.

The coefficient of determination () according to the OLS regression model is 0.837 or . This tells us that approximately **83.7%** of the variance in the total number of wins can be explained by the multiple linear regression equation provided above. This is a high value, which means that this model is strong.

After verifying that the equation and its variables are significant, I tested it by predicting the total of wins in a regular season for a team that is averaging 75 points per game and has a relative skill level of 1350. To do this, I just substituted (**avg\_pts**) and (**avg\_elo\_n**) with the corresponding values:

**Original equation:**

**After substitution:**

**Result:**

I used this equation again to calculate the total wins of a team averaging 100 points per game with an average relative skill of 1600:

**Original equation:**

**After substitution:**

**Result:**

After this, I created another multiple regression model for the total number of wins; however, this time, I used two more predictor variables. As mentioned before in this report, a multiple linear regression model is used to predict the response variable by establishing a relationship between the response variable and several predictor variables. If we use the same equation I provided before, we have that multiple regression model equation looks like this: . I am going to keep the same variables total wins(**total\_wins**) as theresponse variable, average points (**avg\_pts**) and average relative skill (**avg\_elo\_n**) as predictor variables, but I am now introducing the average points differential (**avg\_pts\_differential**), and average relative skill differential (**avg\_elo\_differential**). Using the OLS regression results I got using the Python programming language, we have the following coefficients: ,,,and. Once I substituted these values and variables into the multiple regression model equation, I ended up with the following equation:

To verify that this module is correct, I performed an overall F-test. For this F-test, the null hypothesis () is that none of the predictor variables have a significant relationship with the dependent variable **total\_wins**. This is represented as . The alternative value () is that at least one of the predictor variables has a significant relationship with the dependent variable **total\_wins**. This is represented as . The level of significance (**α**) used in this part is 1% (**α = 0.01**).

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1102 |
| P-value | 0.0 |

As we can see from the table, the p-value (**p**) is 0.0 which is less than the level of significance (**α = 0.01**). We see this mathematically as **p < α**.This meansthat we can reject the null hypothesis (), implying that at least one of the predictors is statistically significant in predicting the number of wins in a season.

From the OLS regression results tables, I got all the p-values for each predictor variable coefficient. This can help us run a t-test on each variable and identify if they are all statistically significant in the model using a 1% level of significance (**α = 0.01**). The null hypothesis () will be that each predictor variable is not significant to the model (), and the alternative hypothesis () will be that each predictor variable is significant to the model (). First, the average relative skill (**avg\_elo\_n**) has a p-value of **p** = **0.442**. The average points (**avg\_pts**) and the average points differential(**avg\_pts\_differential**) both have a value of **p = 0.0**. Lastly, the average relative skill differential (**avg\_elo\_differential**) has a p-value of **p = 0.004**. If we compare all these p-values (**p**) to the level of significance of **α = 0.01**, we see that all the variables except the average relative skill (**avg\_elo\_n**) have a **p** less than **α**.This means that all of the variables are statistically significant except for the average relative skill.

The coefficient of determination () is located in the OSL regression results table, which is or 87.8%. This means that 87.8% of the variance in the total number of wins can be explained by the variables in the model.

I then predicted the 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, an average point differential of -5, and an average relative skill differential of -30. As we did in the previous multiple linear regression equation, we just need to substitute these values in our equation:

**Original equation:**

**Substituting values:**

**Result:**

I also predicted the total wins 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:

**Original equation:**

**Substituting values:**

**Result:**

In this report, I found that these regression models can help predict the Golden State Warriors' total wins in a season based on various performance metrics. A simple linear regression model using average relative skill demonstrated a strong positive correlation with total wins, indicating that higher skill levels lead to more wins. Adding average points scored in a multiple regression model further refined the predictions, showing a moderate positive relationship with wins. Extending the model to include average points differential and average relative skill differential provided a more accurate prediction of wins. For example, teams with higher skill levels, scoring more points, and having favorable differentials were predicted to win significantly more games.

Lastly, the analyses enable the Warriors to predict the team’s performance and make informed adjustments to training, strategies, and player rotations. By identifying key performance indicators such as average relative skill and points scored, the team can focus on improving these areas to increase their win rate.