# MAT 243 Project Three Summary Report

[Luna Saccoia]

[Luna.Saccoia@snhu.edu]

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## 1. Introduction

*Discuss the statement of the problem in terms of the statistical analyses that are being performed. Be sure to address the following:*

* *The data set I am exploring is NBA wins with the average points scored, average relative skill and average point differential between the team and their opponents in the regular season.*
* *My results will be used to predict the total number of wins for a team in the regular season based on based on past performance metrics.*
* *What type of analyses will you be running in this project? I will be running a multiple linear regression analysis to calculate the requested information the coach has requested. This being the predicted total number of wins of a team for the regular season.*

## 2. Data Preparation

*There are some important variables that are used in this project. Identify and explain these variables.*

*See the introductory section and Step 1 of the Python script to address the following:*

* *The variable* ***avg\_pts\_differential*** *represents the average point differential between the team and their opponents in a regular season. This number represents how much higher or lower the teams average points are than their opponents average points. For example if the team’s average points are calculated as 104 and their opponents average points are calculated as 106, the average points differential would be -2. It is negative because their average points calculation is 2 less than their opponents average points calculation.*
* *The variable* ***avg\_elo\_n*** *represents the average relative skill of a team. This is calculated based on the final score of a game, the game location, and the outcome of the game relative to the probability of that outcome. The higher the number, the higher the relative skill of a team.*

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

*You constructed a scatterplot of the total number of wins and the average points scored to study their correlation. You also calculated the Pearson correlation coefficient along with its P-value.*

*See Step 2 in the Python script to address the following items:*

* *In general, data visualization techniques are used to study relationship trends between two variables by showing if there is a correlation between the two variables. For example, if one variable increases and we see the other increase along with it, this can show a positive correlation.*
* *The correlation coefficient can be used to find the strength and direction of the association between two variables by looking at where it falls. If it is less than or equal to .40 it has a weak correlation, if it is greater than .40 and less than or equal to .80 it has a moderate correlation, if it is greater than .80 and less than or equal to 1.0 it is a strong correlation. If one variable increases and the other increases as well, the correlation is positive. If one variable increases while the other decreases, the correlation is negative.*
* *The scatter plot and the correlation coefficient show a moderately positive correlation between total number of wins and average points scored. This is due to the correlation coefficient being 0.4777 which is greater than 0.40 but less than 0.80*
* *The correlation coefficient is statistically significant, based on the p-value of 0.0 being less than the 1% level of significance.*

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

*You created a simple linear regression model for the total number of wins in a regular season using the average points scored as the predictor variable.*

*See Step 3 in the Python script to address the following items:*

* *In general, how is a simple linear regression model used to predict the response variable using the predictor variable? A simple linear regression model can be used to predict the response variable using the predictor variable by using a line drawn between the points called a regression line.*
* *The equation for this model would be where*
* *What are the results of the overall F-test? Summarize all important steps of this hypothesis test. This includes:*
  1. *The null hypothesis states that the value is equal to 0. This would mean that no statistically significant relationship exists between X and Y.*
  2. *The alternative hypothesis states that the value is not equal to 0. This would mean that a statistically significant relationship exists between X and Y.*
  3. *Level of Significance is standard at 5%*
  4. *Report the test statistic and the P-value in a formatted table as shown below:*

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 182.10 |
| P-value | 0.000 |

* 1. *Conclusion of the hypothesis test and its interpretation based on the P-value The conclusion of the hypothesis test and its interpretation based on the p-value would be that the null hypothesis be rejected in favor of the alternative hypothesis. This is due to the p-value of 0.0000 being less than our level of significance of 0.05.This implies that a statistically significant relationship exists between X and Y.*
* *Yes. A significant relationship exists between X and Y as shown by our hypothesis testing rejecting our null hypothesis in favor of our alternative hypothesis which states that a relationship exists.*
* *The predicted number of games a team would win in the regular season, with an average points value of 75 would be 10.8208 or 11 games, rounding to the nearest whole number.*
* *If the team has an average of 90 points per game, the predicted number of games they would win in the regular season is 30.0943 or 30 games, rounding to the nearest whole number.*

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

*You constructed a scatterplot of total number of wins and average relative skill. You also calculated the Pearson correlation coefficient along with its P-value.*

*See Step 4 in the Python script to answer the following questions:*

* *The scatter plot shows a positive correlation and correlation coefficient of 0.9072 shows that the correlation is strong. So the association between total number of wins and average relative skill is that there is a strong positive correlation between these values.*
* *The correlation coefficient is statistically significant based on the p-value of 0.0. This is less than the 1% level of significance. Using our null hypothesis that states and no statistically significant relationship exists between X and Y, we would reject this based on the p-value of 0.0. So we would reject it in favor of our alternative hypothesis that states and that a statistically significant relationship does exist between X and Y.*

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

*You created a multiple regression model with the total number of wins as the response variable, with average points scored and average relative skill as predictor variables.*

*See Step 5 in the Python script to answer the following questions:*

* *A multiple linear regression model is used to predict a response variable using multiple predictor variables, by using the equation for the sum of the population regression function and the regression error term.*
* *The equation for the model is for my model it is*

* *What are the results of the overall F-test? Summarize all important steps of this hypothesis test. This includes:*
  1. *The null hypothesis is that the model is not useful and all slope parameters are equal to zero. The statistical notation for this is*
  2. *The alternative hypothesis is that at least one parameter is not equal to zero, that is*
  3. *The level of significance is standard error, or 5%.*
  4. *Report the test statistic and the P-value in a formatted table as shown below:*

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00*.* |
| P-value | 0.0000 |

* 1. *My conclusion of the hypothesis test and its interpretation based on the P-value is that the null hypothesis should be rejected based on the p-value of 0.0000 being less than our level of significance of 0.05. This implies that at least one parameter is not equal to zero and instead has a correlation to the response variable.*
* *Based on the results of the overall F-test, yes, at least one of the predictors is statistically significant in predicting the total number of wins in the season.*
* *Performing a t-test for each parameter are as follows: t-statistic is for average points with a p-value of 0.0000 , and t-statistic is for average relative skill with a p-value of 0.0000.This is less than our 1% level of significance and therefore rejects our null hypothesis in favor of our alternative hypothesis. This means that a statistically significant relationship exists between our response variable and at least one of our predictor variables; in this case it shows that a relationship exists between both of the predictor variables individually with our response variable.*
* *The coefficient of determination is, this states that 83.7% of the variance in total wins can be explained by the variance in average points scored and average relative skill.*
* *Predicted .The predicted total number of wins in the regular season for a team that is averaging 75 points per game and has a relative skill of 1350 is 16 games*
* *. The predicted total number of wins in the regular season for a team that is averaging 100 points per game and has a relative skill of 1600 is 51 games.*

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

*You created a multiple regression model with the total number of wins as the response variable, with average points scored, average relative skill, and average points differential as predictor variables.*

*See Step 6 in the Python script to answer the following questions:*

* *A multiple linear regression model is used to predict a response variable using multiple predictor variables, by using the equation for the sum of the population regression function and the regression error term.*
* *The equation for this model is . For my model it is*
* *What are the results of the overall F-test? Summarize all important steps of this hypothesis test. This includes:*
  1. *The null hypothesis is that the model is useless, all slope parameters are equal to zero and no statistically significant relationship exists between any of our predictor variables and our response variable;*
  2. *The alternative hypothesis is that the model is useful, at least one slope parameter is not equal to zero and a statistically significant relationship exists between at least one of our predictor variables and our response variable;*
  3. *Level of Significance is standard at 5%. Even using the lower level of significance referenced in later in this section of 1%, the p-value is still less than it and rejects the null hypothesis.*
  4. *Report the test statistic and the P-value in a formatted table as shown below:*

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449.00 |
| P-value | *0.0000* |

* 1. *My conclusion is that the null hypothesis should be rejected in favor of our alternative hypothesis. The p-value of 0.0000 is less than our level of significance and therefore rejects our null hypothesis. This implies that the model is useful and that a statistically significant relationship exists between at least one of our predictor variables and our response variable.*
* *Based on the results of the overall F-test, yes, at least one of the predictors is statistically significant in predicting the number of wins in the season.*
* *Performing a t-test for each parameter is as follows: t-statistic is for average points, t-statistic is for average relative skill, and t-statistic is for average points differential with a p-value of 0.0000 for each one. This is less than our level of significance of 1% for every one of these variables, as the p-value is 0.0000 for each which is less than 0.01. This means we should reject our null hypothesis in favor of our alternative hypothesis. This shows that a statistically significant relationship exists between our response variable and at least one of our predictor variables; in this case it shows that a statistically significant relationship exists between our response variable and every one of our predictor variables.*
* *The coefficient of determination is . This states that 87.6% of the variance in total wins can be explained by the variance in average points scored, average relative skill, and average points differential.*
* *Predicted wins. 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 and average point differential of -5 is 20.3224 or 20 games.*
* *Predicted wins. 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 and average point differential of +5 is 52.6584 or 53 games.*

## 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 result of all these analyses is that there is a statistically significant relationship between a team’s history and future predictions. Teams that have higher average points per game, higher relative skill levels and higher average points differentials are likely to win more games in future regular seasons than teams with lower values in these areas. There is a strong positive correlation between these variables and the prediction of how many games they will likely win in the upcoming regular season.*
* *The practical importance of the analyses that were performed is that coaches can use these predictions to see where they need to improve their team’s performance in order to give them a higher prediction for number of wins in the upcoming regular season.*

## 9. Citations

Zybooks.(2019). Zybooks: Applied Statistics I for Science, Technology, Engineering, and Math