MAT 243 Project Three Summary Report

Nicholas Cleveland

nicholas.cleveland1@snhu.edu

Southern New Hampshire University

1. Introduction

The data set that I will be exploring in this report is the correlation between predictor and response variables used in given basketball game statistics from NBA teams. I will be looking at the correlation specifically between the average number of points to the number of wins, the average relative skill to the number of wins, and both predictors, the average number of points and skill level to the number of wins. I will finally add a third predictor variable of the average points differential along with the first two predictor variables compared to the response variable of the number if game wins. The results will be used to conclude connections between predictor and response variables. The types of analyses that will be conducted will be simple linear regression tests, multiple linear regression tests, and hypotheses tests.

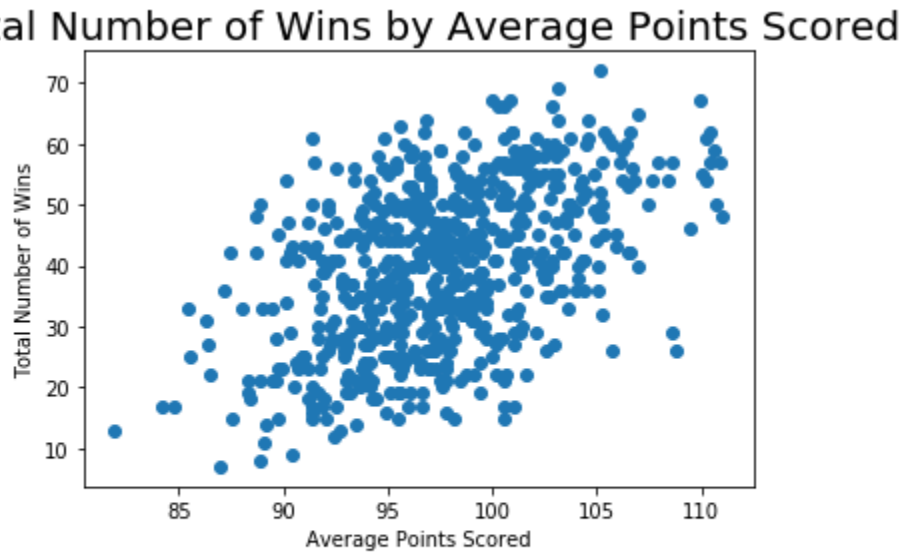
2. Data Preparation

The average points differential represents the average difference of points between a team and their opponents during the season. The average relative skill level variable represents the relative skill of each team in a regular season.

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

In general, data visualization techniques can be used to show relationships between two variables more easily than another representation such as a table or similar. The data visualization manifested as a scatter plot can show a clear relationship between a predictor variable and a response variable to the point where the trend can be identified. Most of the time trends will show representations such as a positive or negative trend. The correlation coefficient is used to describe the relationship between two variables. There are variables levels of strength

that range from weak, to moderate, to strong. Weak is a coefficient between 0 and 0.40, moderate is a coefficient between 0.40 and 0.80 and strong is a coefficient between 0.80 and 1.00.




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

The scatterplot shows a positive correlation between the predictor and response variables here, being the average points scored during a season and the total wins. If we were to plot a line to show this, it would show a line with a positive slope. The Pearson Correlation Coefficient here was 0.4777, which a **moderate** strength of correlation due to it being between 0.40 and 0.80.

Since we have a P-value of approximately 0.00, and it is less than the level of significance of 0.01, we can reject the null hypothesis of the correlation being not statistically significant and favor the alternative hypothesis to which we say that this 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 predict response variable via an equation where it takes a predictor variable and will output the response variable, Y. The equation for this model is: Y = -85.5476 + 1.2849 * X1 Or, **total wins = -85.5476 + (1.2849 * avg_points)**. The null hypothesis is that average points does not predict the number of wins during a season. The alternative hypothesis is that the averages points does predict a high number of wins during the basketball season. The level of significance by default is 0.05 or 95%.

Statistic	**Value**
Test Statistic	182.10
P-value	1.5200×10^{-38}

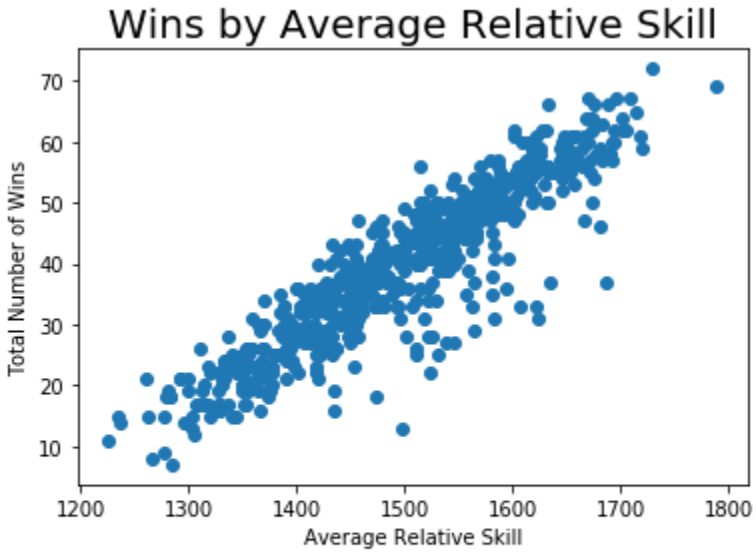
Table 1: Hypothesis Test for the Overall F-Test

Based on the results of the overall F-test, the F-statistic that was generated was 182.10 with a p-value of 1.5200×10^{-38}. If we were to use the hypothesis test here, we can see that the p-value is much less than the usual level of significance of 0.05, which means we would reject the given null hypothesis and favor the alternative hypothesis. The null hypothesis in this situation was that there was not a correlation between the average number of points scored with the number of wins, while the alternative hypothesis was that as the number of average points increases, the high number of wins are received. Since we favored the alternative hypothesis

based on our test, we can confidently say that there is a significant correlation between the number of average points scored and the number of games won.

In a situation where a team is scoring 75 points per game, we can use our generated model equation of **total wins = -85.5476 + (1.2849 * avg_points)**, where we plug 75 for the avg_points value to get a **total wins** of **10 (rounded) games**. In a situation where a team is scoring 90 points per game, we can again use our generated model equation of **total wins = -85.5476 + (1.2849 * avg_points)**, where we plug 90 for the avg_points value to get a **total wins** of **30 (rounded) games**.

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 itself shows a very clear positive linear relationship between the predictor variable, average relative skill, and the response variable; the total number of wins. We

generated a Pearson coefficient correlation of 0.9072, which is very high. It scales well above the 0.80 region of the strength scale and would be considered a **strong** correlation coefficient.

Using a 1% or 0.01 level of significance to compare to our p-value, we generated a p-value of 0.000 which is below the alpha of 0.01. If we were to take our hypothesis test for this situation, where the null hypothesis is that the average relative skill does not impact the number of wins and the alternative hypothesis being that the higher the skill level is, the more games will be won, we would favor the alternative hypothesis here. In which, we can confidently say that there is a significant correlation between the predictor variable and the response variable.

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

In general, a multiple linear regression model is used to predict response variable via an equation where it takes two or more predictor variables and will output the response variable, Y. The equation for the model is Y = -152.5736 + (0.3497 * X1) + (0.1055 * X2), or more specifically, total_wins = -152.5736 + (0.3497 * avg_pts) + (0.1055 * avg_elo_n), where total_wins is the total games won, avg_pts is the average number of points, and avg_elo_n is the average relative skill level.

The null hypothesis in this case is that none of the predictor variables have a correlation with the response variable. In this case, it would be that neither the average number of points nor the average relative skill level affects the total number of points. The alternative hypothesis in this case would be that one or more of the predictor variables has a correlation with the response

variable. This means that although we have more than one predictor variable, only at least one must be correlated to create a significant correlation. The level of significance in this case is the default of 0.05 or 95%.

Statistic	**Value**
Test Statistic	1580.00
P-value	4.4100×10^{-245}

Table 2: Hypothesis Test for the Overall F-Test

Since we got a p-value of 4.4100×10^{-245}, this tells us that at least one of the predictor variables is statistically significant in predicting the total number of wins in the season. Based on the results of the t-test, both predictor variables, average points and average skill level are statistically significant variables in predicting the total number of wins in the season. Both factors have a p-value of 0.000, which is less than the given alpha of 0.01.

The coefficient of determination was calculated to be 0.837. This is a significantly high value for our coefficient of determination, or R^2. If we wanted to look at the Pearson coefficient of this all we would have to do is find R which is the square root of R^2, by doing this, we get a Pearson coefficient of 0.91, which falls into the **strong** correlation range. Given our equation for our model, total_wins = -152.5736 + (0.3497 * avg_pts) + (0.1055 * avg_elo_n), to evaluate the predicted total number of wins given the average points of 75 per game and a relative skill level of 1350, we simply plug these values in to the equation and receive the output of the total wins.

By doing so, we receive a value of **total_wins = 16.0789** or rounded down to 16. Again, given our equation for our model, total_wins = -152.5736 + (0.3497 * avg_pts) + (0.1055 * avg_elo_n), to evaluate the predicted total number of wins given the average points of 100 per game and a relative skill level of 1600, we simply plug these values in to the equation and receive the output of the total wins. By doing so, we receive a value of **total_wins = 51.1964** or rounded down to 51.

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

In general, a multiple linear regression model is used to predict response variable via an equation where it takes two or more predictor variables and will output the response variable, Y. The equation for the model is Y = -35.8921 + (0.2406 * X1) + (0.0348 * X2) + (1.7621 * X3), or more specifically, total_wins = -35.8921 + (0.2406 * avg_pts) + (0.0348 * avg_elo_n) + (1.7621 * avg_pts_differential), where total_wins is the total games won, avg_pts is the average number of points, avg_elo_n is the average relative skill level, and avg_pts_differential is the average points difference between each team and their opponents in the season.

The null hypothesis in this case would be that the predictor variables of neither the average points, nor the average relative skill, nor the average points differential would have a correlation with the response variable. The alternative hypothesis in this case would be that at least one of the predictor variables has a significant correlation with the response variable. The level of significance in the generated table by default is 0.05 or 95%.

Statistic	Value
Test Statistic	1449.00
P-value	5.0300×10^{-280}

Table 3: Hypothesis Test for Overall F-Test

Based on the overall results of the overall F-test, we have a test statistic of 1449.00 and a p-value of 5.0300×10^{-280}, which is less than the default alpha of 0.05, which favors the alternative hypothesis. For each individual predictor variable, a t-test was done in which a t-statistic and a p-value was calculated. For each predictor, average points, the average relative skill level, and the average points differential, all have a p-value of 0.0000. If we were to take a hypothesis test where the null hypothesis is that there is no correlation between each predictor variable and the response variable versus an alternative hypothesis where there is a correlation of a predictor variable with the response variable, we would simply have to compare each p-value to the given alpha value of 0.01. Since all 0.000 values are less than 0.01, we can say there is significant evidence to deny the null hypothesis and to favor the alternative hypothesis, meaning, there is significant evidence that there is a correlation between all the predictor variables and the response variable.

The coefficient of determination, R^2, was calculated to be 0.876. If we were to look at the Pearson coefficient, R, which measures weak, moderate, or strong correlations, we would get a value of 0.94, which is categorized as a **strong** correlation since it is above 0.80 and below 1.00. Given the equation for the model, total_wins = -35.8921 + (0.2406 * avg_pts) + (0.0348 *

avg_elo_n) + (1.7621 * avg_pts_differential), we would simply need to plug in each corresponding predictor value to get the response variable, num_wins. By doing so, we generated: total_wins = -35.8921 + (0.2406 * 75) + (0.0348 * 1350) + (1.7621 * -5) = 20.3224 or rounded down to 20. Again, given the equation for the model, total_wins = -35.8921 + (0.2406 * avg_pts) + (0.0348 * avg_elo_n) + (1.7621 * avg_pts_differential), we would simply need to plug in each corresponding predictor value to get the response variable, num_wins. By doing so, we generated: total_wins = -35.8921 + (0.2406 * 100) + (0.0348 * 1600) + (1.7621 * 5) = 52.6584 or rounded down to 52.

8. Conclusion

In conclusion, after performing various statistical tests on all of the predictor variables with the dependent response variable, all predictor variables showed a significant statistical correlation between themselves and the response variable. The tests that were performed with multiple regression and multiple predictor variables also showed a significant correlation between them and the response variable. All these conclusions matched up with the intuitive analysis of looking at the trend of each graph as well. The predictor variables of average score, average relative skill level, and average points difference all proved to be significant indicators that directly affect the response variable, the number of games won.