# MAT 243 Project Three Summary Report

Alexander Varljen

Alexander.varljen@snhu.edu

Southern New Hampshire University

## 1. Introduction

I am a data analyst given the task to perform statistical analysis by an NBA basketball team’s coach and team management for the purpose of finding the general patterns to success for the team in the league. I’ll be doing this by finding the correlation between the total number of games won by a team and average points scored by a team via a scatterplot, by using simple linear regression to predict total number of wins using the average points scored by a team, by finding the correlation between the total number of wins and the average relative skill level of a team, by predicting a teams total number of wins using average points scored and average relative skill level via multiple regression, and by predicting the total number of wins for a team using average points scored, average relative skill, and the average points differential. The data set that I am using includes the average points scored in a regular season, the average points scored in a regular season, the average relative skill rating, the average opponents regular skill rating, the average points differential, the average relative skill level differential, and the total number of wins in a regular season for each NBA basketball team from years 1995 to 2015.

## 2. Data Preparation

The average points differential, which is denoted by the variable “avg\_pts\_differential”, is the difference between the average points scored by a team in a regular season and their opponents average points scored. For example, if one team scores an average of 90 points per game in a season, and their opponents score an average of 95 points, the average points differential would be -5. The average relative skill, which is denoted by the variable “avg\_elo\_n”, is the average of a team’s relative skill in a season. The relative skill of a team is based off the final score for a given game, the location of the game, and the outcome of a game relative to its expected result. In simpler terms, the higher the number, the higher the relative skill for a team.

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

Below is the scatterplot and correlation for the total number of wins and average points scored in a regular season for all NBA teams:

Chart, scatter chart

Description automatically generated

Correlation between Average Points Scored and the Total Number of Wins

Pearson Correlation Coefficient = 0.4777

P-value = 0.0

The scatterplot above shows the relationship between the total number of wins by all teams and the average points scored by a team in a regular season by NBA teams from years 1995 to 2015. As we can see from the graph, in general, as the average points scored by a team increase, so does the total number of wins. Below the graph, we have the correlation coefficient presented, which is about 0.478, or R = 0.478. This number gives us an accurate depiction of the strength of the correlation between two variables. In this case, since R = 0.478 and is between the values of 0.8 and 0.4, we know that there is a moderate negative correlation between the total number of wins in a regular season and the average points scored per game in a regular season. We can also see that the p-value is 0.000, which is less than the level of significance, α = 0.01. This means that the correlation between these two variables is statistically significant. This means that the team’s coach is right to expect that the team is more likely to win if it scores a higher average number of points per game.

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

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1.82  *\*Round off to 2 decimal places.* |
| P-value | 0.0000  *\*Round off to 4 decimal places.* |

A simple linear regression model uses the basic format of a linear equation, which is y = m\*x + b, to predict the response variable, y, by replacing the variable m with the predictor variables coefficient, replacing b with the y-intercept, and plugging in a value for x. In this example, the equation is total\_wins = 1.285\*(avg\_pts) – 85.548, where total\_wins is the total number of wins of a game in a regular season and avg\_pts is the average points scored per game by a team in a regular season. The null hypothesis, which is based on the coach’s suggestion, is H0 : B1 > 0, where B1 is the predictor variable, average points. In other words, the null hypothesis is that the average points scored per game will have a positive effect on the total number of wins. The alternative hypothesis is for this simple regression model is Ha : B1 = 0, where B1 is the predictor variable, average points. In other words, the alternative hypothesis is that the average points scored per game in a regular season has no effect on the total number of wins for a team. The level of significance is 1%, or α = 0.01. As we can see from the table above, the overall F-test has found that the test statistic is about 1.82 and the p-value is about 0.0000. Based on this, we can conclude that a statistically significant linear relationship exists as we fail to reject the null hypothesis, as the p-value is below the level of significance. This means that this linear model can be used to predict the total number of wins in a regular season.

If we plug in an average of 75 points per game into our linear equation, or total\_wins = 1.285(75) - 85.548, we obtain an answer of 10.827, which means that a team that averages 75 points per game is likely to win about 10 games per season. If we plug in an average of 90 points per game into our linear equation, or total\_wins = 1.285(90) – 85.548, we obtain an answer of 30.102, which means that a team that averages 90 points per game is likely to win about 30 games per season.

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

Below is the scatterplot, correlation coefficient and p-value for the total number of wins per season compared to the average relative skill of a team:

Chart, scatter chart

Description automatically generated

Correlation between Average Relative Skill and Total Number of Wins

Pearson Correlation Coefficient = 0.9072

P-value = 0.0

As we can see from the graph above, there seems to be a positive correlation between a team’s average relative skill and the total number of games won by a team in a regular season. In other words, it appears that as the average relative skill of a team increases, so does its total number of wins, on average. The correlation coefficient, which is about R = 0.907, suggests that this correlation is a strong positive correlation, as the value is between 1.0 and 0.8. Since the p-value, which is about 0.0, is less than the level of significance, α = 0.01, we can say that the correlation coefficient is statistically significant.

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

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580.00  *\*Round off to 2 decimal places.* |
| P-value | 0.0000  *\*Round off to 4 decimal places.* |

A multiple linear regression model uses a format like that of a linear equation with one more m value, or y = β1\*x + β2\*x + b, to predict the response variable, y, by replacing the variable β1 with the first predictor variables coefficient, replacing β2 with the second predictor variables coefficient, replacing b with the y-intercept, and plugging values for the x variables. In this example, the equation is total\_wins = 0.350\*(avg\_pts) + 0.106(avg\_elo\_n) – 152.574, where total\_wins is the total number of wins of a game in a regular season, avg\_pts is the average points scored per game by a team, and avg\_elo\_n is the average relative skill level of a team in a regular season. The null hypothesis, which is based on the coach’s suggestion, is H0 : βn > 0, where βn is one of the two predictor variables. In other words, the null hypothesis is that either, the average points scored per game, or the average ELO rating will have a positive effect on the total number of wins. The alternative hypothesis is for this simple regression model is Ha : β1 = β2 = 0. In other words, the alternative hypothesis is that both the average points scored per game in a regular season and the average ELO rating has no effect on the total number of wins for a team. The level of significance is 1%, or α = 0.01. As we can see from the table above, the overall F-test has found that the test statistic is about 1580.00 and the p-value is about 0.0000. Based on this, we can conclude that a statistically significant linear relationship exists as we fail to reject the null hypothesis, as the p-value is below the level of significance. This means that this linear model can be used to predict the total number of wins in a regular season, as at least one of the predictor variables is not equal to zero.

The individual t-test for the first parameter, average points scored, shows a test statistic of 7.297 and a p-value of 0.000. This means that the effect of the parameter average points scored has a statistically significant effect on the predictive model, as the p-value is less than the level of significance α = 0.01. The individual t-test for the second parameter, average relative skill level, shows a test statistic of 47.952 and a p-value of 0.000. This means that the effect of the parameter average relative skill level also has a statistically significant effect on the predictive model, as the p-value is less than the level of significance, α = 0.01. The coefficient of determination is R^2 = 0.837, which means that 83.7% of the variance in the model can be explained by the two parameters and that the other 16.3% is explained by other factors or is possibly due to error.

If we plug in an average of 75 points per game and an average ELO rating of 1350 into our linear equation, or total\_wins = 0.350(75) + 0.106(1350) – 152.574, we obtain an answer of 11.776, which means that a team that averages 75 points per game and has an average ELO rating of 1350 is likely to win about 12 games per season. If we plug in an average of 100 points per game and an average ELO rating of 1600 into our linear equation, or total\_wins = 0.350(100) + 0.106(1600) – 152.574, we obtain an answer of 52.026, which means that a team that averages 90 points per game and has an average ELO rating of 1600 is likely to win about 52 games per season.

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

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449.00  *\*Round off to 2 decimal places.* |
| P-value | 0.0000  *\*Round off to 4 decimal places.* |

A multiple linear regression model uses a format like that of a linear equation with one more m value, or y = β1\*x + β2\*x + βn + b, to predict the response variable, y, by replacing the variable β1 with the first predictor variables coefficient, replacing β2 with the second predictor variables coefficient, replacing βn with any additional predictor variable coefficients, replacing b with the y-intercept, and plugging values for the x variables. In this example, the equation is total\_wins = 0.241\*(avg\_pts) + 0.035(avg\_elo\_n) + 1.762(avg\_pts\_differential) – 35.892, where total\_wins is the total number of wins of a game in a regular season, avg\_pts is the average points scored per game by a team, avg\_elo\_n is the average relative skill level of a team, and avg\_pts\_differential is the average points scored differential in a regular season. The null hypothesis, which is based on the coach’s suggestion, is H0 : βn > 0, where βn is one of the three predictor variables. In other words, the null hypothesis is that either, the average points scored per game, or the average ELO rating or the average points scored differential will have a positive effect on the total number of wins. The alternative hypothesis is for this simple regression model is Ha : β1 = β2 = β3 = 0. In other words, the alternative hypothesis is that the average points scored per game in a regular season, the average ELO rating, and the average points differential has no effect on the total number of wins for a team. The level of significance is 1%, or α = 0.01. As we can see from the table above, the overall F-test has found that the test statistic is about 1449.00 and the p-value is about 0.0000. Based on this, we can conclude that a statistically significant linear relationship exists as we fail to reject the null hypothesis, as the p-value is below the level of significance. This means that this linear model can be used to predict the total number of wins in a regular season, as at least one of the predictor variables is not equal to zero.

The individual t-test for the first parameter, average points scored, shows a test statistic of 5.657 and a p-value of 0.000. This means that the effect of the parameter average points scored has a statistically significant effect on the predictive model, as the p-value is less than the level of significance α = 0.01. The individual t-test for the second parameter, average relative skill level, shows a test statistic of 6.421 and a p-value of 0.000. This means that the effect of the parameter average relative skill level also has a statistically significant effect on the predictive model, as the p-value is less than the level of significance, α = 0.01. The individual t-test for the third parameter, the average points scored differential, shows a test-statistic of 13.928 and a p-value of 0.000. This means that the effect of the parameter average points scored differential has a statistically significant effect on the model, as the p-value is less than the level of significance α = 0.01. The coefficient of determination is R^2 = 0.876, which means that 87.6% of the variance in the model can be explained by the three parameters and that the other 12.4% is explained by other factors or is possibly due to error.

If we plug in an average of 75 points per game, an average ELO rating of 1350, and an average points differential of -5 into our linear equation, or total\_wins = 0.241(75) + 0.035(1350) + 1.762(-5) – 35.892, we obtain an answer of 20.623, which means that a team that averages 75 points per game, has an average ELO rating of 1350, and has an average points differential of -5 is likely to win about 21 games per season. If we plug in an average of 100 points per game, an average ELO rating of 1600, and an average points differential of 5 into our linear equation, or total\_wins = 0.241(100) + 0.035(1600) + 1.762(5) – 35.892, we obtain an answer of 53.018, which means that a team that averages 90 points per game, has an average ELO rating of 1600, and has an average points differential of 5 is likely to win about 53 games per season.

## 8. Conclusion

After doing the analysis, I have found that there is a moderate positive correlation between the average points scored per game in a regular season and the total number of games won by a team, a strong positive correlation between a team’s average relative skill and the number of games won by a team in a regular season, and that the average points scored per game, the average ELO rating and the average points differential all have statistically significant predictive power, either on their own or together, for a model that predicts the total number of games won by a team. In other words, a team who scores higher per game is likely to have more total wins per season, a team who has an average relative skill level is likely to have more total wins, and a model can be made whereby the team’s coach and management can predict their total number of wins in future seasons with a fair bit of accuracy if the team’s average points scored per game, average ELO rating, and the average points differential for a regular season are known. In summation, this analysis not only shows that the coach’s predictions are correct and that all three factors are predictors of a team’s success, but as these scores, ratings and differentials continue to be gathered, the team can reasonably predict what the number of wins from season to season will be.