# MAT 243 Project Three Summary Report

Devon McCain

[Devon.mccain@snhu.edu](mailto:Devon.mccain@snhu.edu)

Southern New Hampshire University

The data that I have been set to explore is from the NBA to analyze performance patterns. I have been tasked to come up with regression models that can predict the total number of wins for a team within their regular season based on their prior performances. These results will be used to help made decisions that can improve the performance of the team. The analyses will consist of scatterplots to detect correlations and linear regressions to predict the number of wins.

There are a few different variables that may be difficult to understand without prior knowledge of the data. When looking at the different variables, there will be avg\_pts\_differential. What this variable represents is the average point differential between two different teams within their regular season. In simpler terms, avg\_pts\_differential is essentially comparing the two teams. One of the other variables, avg\_elo\_n, represents the average points scored in a regular season. To break that down, this variable is taking all the points the team has scored and finding the average among that data.

When looking at data, it can be difficult to understand when you are simply looking at numbers. Using a data visualization technique aids the reader in a way where they can quickly see what the data is doing and whether there are any patterns or correlations. When studying the relationship trends between two variables, it helps to find the correlations that may be present in the data. The correlation coefficient is used to find the strength and direction when pertaining to the association between two different variables. So, there can be a Positive, negative, and a Zero correlation. When the correlation coefficient is near 1 or -1, this means that linear relationship is strong, however if it is near zero, this means that the relationship is weak. Typically, the direction is based on whether the correlation coefficient is positive (to the right) or negative (to the left). From what the scatterplot is showing, it appears that the higher number of wins a team has, the more points that the team scored. The Pearson correlation coefficient is showing the same things, but it appears to be on the weak side of the correlation spectrum. When looking at the P-Value, it is at a 0.0. If this is less than the significance level of 0.01, the null hypothesis will be rejected. This will mean that there will be no correlation between the average points scored and the number of wins. However, if the P-value is greater, then the null hypothesis is not rejected. In this case, it appears that the null hypothesis will be rejected due to the P-Value.

Chart, scatter chart

Description automatically generated

Table 1: Hypothesis Test for the Overall F-Test

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

The simple linear regression model used to predict the response variable using the predictor variable works by predicting or estimating the equation for the simple linear progression based on the intercepts. The equation used for this model is . While the equation that I used based on this is . When pertaining to the overall F-Test, the null hypothesis is and the alternative hypothesis is . The level of significance used here is 95% or 0.05. With the P-Value being 0, we should reject the null hypothesis due to it being less than the alpha value. With this, there is enough evidence to show that the overall regression model can somewhat predict a win based upon a score. Based on this information, a team that scores an average 75 points a game will have roughly 14 wins. As for 90 points, this will be about 35.

With the scatterplot and Pearson correlation coefficient, both show that there is a strong correlation between the total number of wins and the average relative skill. With the P-Value being 0, the correlation coefficient is not statistically significant based on the P-Value. This is because the P-Value is less than the 1% or 0.01 for the level of significance.

Chart, scatter chart

Description automatically generated

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449 |
| P-value | 0.000 |

Multiple linear regression is a technique that can use several variables that can predict the outcome of a response variable. The purpose of multiple linear regression is to model the linear relationship between the independent and dependent variables. These work by using different formulas to get the calculations. The equation for the model is . When pertaining to the overall F-Test, the null hypothesis is for the model with no independent variables (). As for the alternative hypothesis, it will fit better than an intercept model (). The level of significance used here is 5% or 0.05. With the P-Value being less than the level of significance, the null hypothesis is rejected. This means that at least one predictor variable is statistically significant when predicting the number of wins. Even with a 0.01 level of significance, there is still only one variable that is statistically significant based on the 0 P-Value. The coefficient of determination is 0.876 or 87.6%. This means that there is an 87.6% chance of variability when considering he number of wins. The predicted total number of wins for a team in regular season that is averaging 75 points per game with a relative skill level of 1350 is roughly 29 games. However, the predicted total number of wins in a regular season for a team that is averaging 100 points per game with an average relative skill level of 1600 is about 44 games.

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1449.00 |
| P-value | *0.000* |

Multiple linear regression is used to predict the outcome of a response variable by using several predictor variables. This works by using different formulas and or equations to get the calculations to interpret what the results could be. The general equation for the model is . As for the null hypothesis, it is predicting that there is no linear relationship between the number of wins in the season (). As for the alternative hypothesis, it is predicting that there is at least one of the predictors that are statistically significant when predicting the number of wins in the regular season (). The level of significance used in this model is 5% or 0.05. With the P-Value being 0, this means that the null hypothesis is rejected, and the alternative is accepted. This means that at least one variable is statistically significant in predicting the number of wins. This is also true if the level of significance changed to 1% or 0.01. The coefficient of determination is 0.876 or 87.6%. When wondering what the odds of a team in regular season with an average of 75 points per game and a relative skill level of 1350 and average point differential of -5 winning is about 20 games in the season. If this were to change to an average of 100 points per game with a relative skill level of 1600 and an average point differential of +5, then they would win about 52 games in the season.

The results show that there is at least one variable that is tied to predicting the results of the teams winning a certain number of games within a regular season. The importance of this information is for the teams, as they can take it and see that they may need improvement and to work from there. This is also good information if someone was to gamble and wanted to find the best odds of a specific team winning. Overall, there are multiple uses that someone may use with the analyses that were performed.