MAT 243 Project Three Summary Report

Walker Martin

walker.martin@snhu.edu

Southern New Hampshire University

1. Introduction

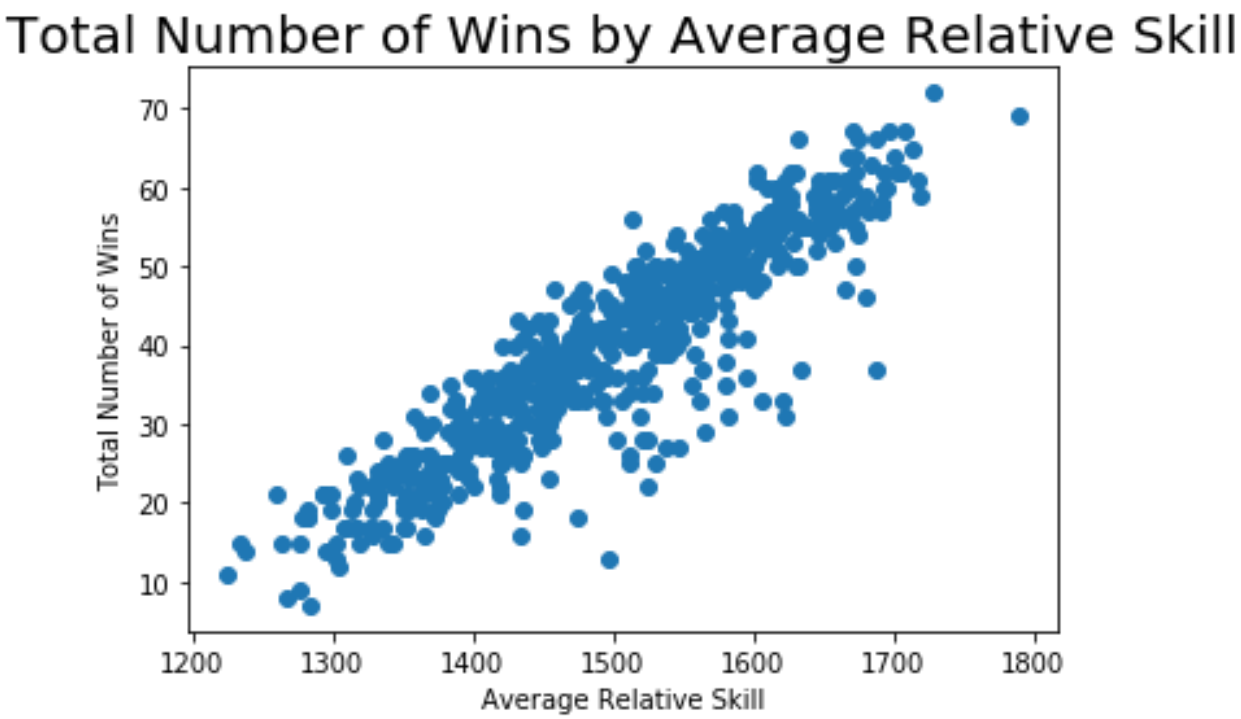
Throughout this project the data being used for statistical analysis is a large csv containing multiple variables such as total wins, relative skill, average points per game. All thirty teams spanning from 1995-2015 are included in the dataset. We will be creating scatter plots and running regression models on the different variables within the dataset to explore their correlation then come to a conclusion regarding various hypotheses.

2. Data Preparation

There are five variables within the script; avg_pts_differential is the average points scored difference between a given team and their opponent during a regular season game. While ave_elo_n is the average relative skill of a given team during the regular season. All of the data and variables are focused on regular season games assumingly to keep the sample size per team even since regular season is the only time where all thirty teams are active.

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

Data visualizations are essential for interpreting mass amounts of data points in a short period of time or when presenting analytics to others. Scatter plots specifically provide insight into the relationship between two variables by comparing the two directly on either axis. The coefficient of correlation is used to determine the strength of the variables relationship to each other, the higher the coefficient the stronger the correlation is. Viewing the scatter plot total number of wins by average relative skill it's clear that here is a positive correlation between the two. The result is a close grouping with a linear slope and slightly more outliers below the trend line than above. The P value between the variables is 0. The correlation coefficient being strong and the p value being under the level of significance alpha 0.01 confirms that there is a strong linear relationship between the total wins and average relative skill.



4. Simple Linear Regression: Predicting the Total Number of Wins using Average Relative Skill

Simple linear regression models are used to depict the relationship between two variables. They use a line of best fit or otherwise called a regression line; using a regression line we can determine average points scored and drawing from the regression line is how we can determine difference from average points.

In this scenario the equation model is $y = -128.2475 + 0.1121b_1$.

The Null Hypothesis is H_0 : $B_1=0$; average relative skill is not significantly relevant to total wins.

The Alternative Hypothesis is: H_a: $B_1 \neq 0$, the relationship between average relative skill and total wins is significantly relevant.

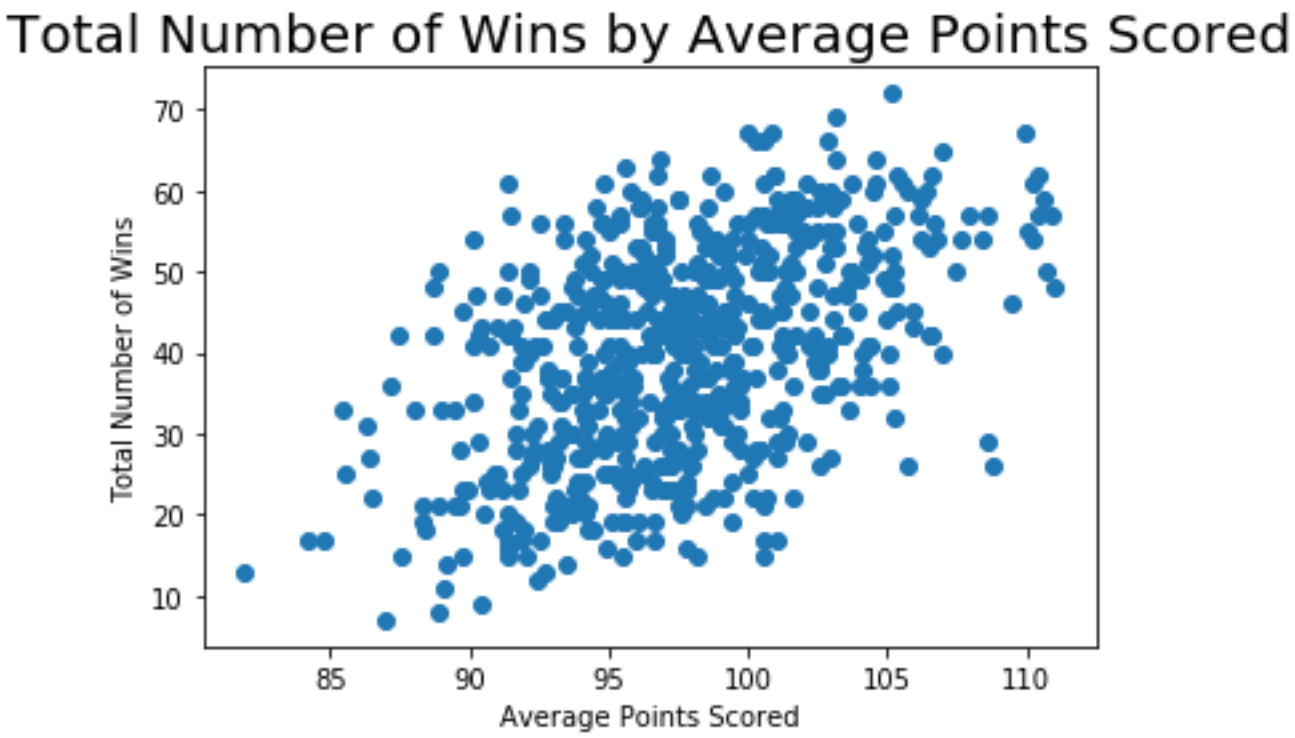
Table 1: Hypothesis Test for the Overall F-Test

Statistic	**Value**
Test Statistic	2865.00
P-value	8.06e-234

The coaches hypothesis is that keeping a high relative skill through the regular season will result in a higher amount of wins. Since the P value is well below the level of significance we reject the null hypothesis and accept the alternative. Therefore the coach is correct and a significant relationship exists between relative skill and total wins. Since the F statistic is high and the P value is low we can use relative skill to predict total wins. Plugging in a relative skill of 1550 into the equation model the number of wins is 45. When we plug 1450 into the model the predicted number of wins is 34.

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

Based on viewing the scatter plot there is a weak linear relationship between the two variables. The coefficient of correlation is 0.4777 which proves that there is a weak relationship between the variables. Additionally the P value is less than the significance level, therefore, we can reject the null hypothesis of no significant relationship between the two variables.



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

Multiple regression models are used to determine if a relationship between the response and a predictor variable exists. Using a regression model provides the values for the regression formula that can be used to input predictors and output the response variables.

The equation for the given model is: $y = -152.5736 + 0.3497b_1 + 0.1055b_2$

The Null Hypothesis is $H_0: B_1 = B_2 = ... = B_n = 0$; There is no significant linear relationship that exists between either predictor variable and the response.

The Alternative Hypothesis is: H_a: At least one $B_i \neq 0$ for i = 1, 2, …, n; At least one significant relationship between a predictor and the response variable exists.

Table 2: Hypothesis Test for the Overall F-Test

Statistic	**Value**
Test Statistic	1580.00
P-value	4.41e-243

Because of the low P value we can reject the null hypothesis. Therefore we know at least one predictor variable is linearly significant. Looking at the individual t test p values we see that both are under the level of significance therefore both predictors are significant. The coefficient of determination is 0.837 which shows that a strong correlation does exist. Plugging in 75 points per game and 1350 relative skill into the equation the predicted number of wins are 16. Plugging in 100 points and 1600 relative skill the predicted win amount is 51.

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

This multiple regression model allows us to predict the response variable wins by using four predictor variables. This arguably provides a more accurate representation of the response variable than using fewer predictors as long as the predictor variables are all significant to the response variable individually.

The equation for the given model is: $y = 34.5753 + 0.2597b_1 - 0.0134b_2 + 1.6206b_3 + 0.0525b_4$

The Null Hypothesis is $H_0 : B_1 = B_2 = ... = B_n = 0$; There is no significant linear relationship that exists between any of the predictor variables and the response.

The Alternative Hypothesis is: H_a: At least one $B_i \neq 0$ for i = 1, 2, …, n; At least one significant relationship between a predictor and the response variable exists.

Table 3: Hypothesis Test for Overall F-Test

Statistic	**Value**
Test Statistic	1102
P-value	3.07e-278

The P value indicates a significant linear relationship with at least one of the predictor variables. Therefore we can reject the null hypothesis. Based on the results from the individual T tests we can determine that avg_pts, avg_pts_differential, and avg_elo_differential, while avg_elo_n does not have an effect on the response variable due to its P value being greater than alpha 0.01. The coefficient of determination or R^2 is 0.878, showing strong correlation. Given the inputs of 75 points per game, relative skill of 1350, differential in points of -5 and an average skill differential of -30 the predicted total wins per regular season is 26. With inputs of 100 points a game on average, 1600 relative skill, +5 differential in points and +95 differential in skill the predicted response variable is 52.

8. Conclusion

Throughout the analysis we interpreted different comparisons and regressions how the response variable wins. In step two we determined that there is a strong linear correlation between the total number of wins and the predictor variable average relative skill. What's interesting about this finding is that in step six the individual T test shows that the average relative skill becomes insignificant when paired with the other three predictor variables in a multiple regression model. Therefore just because a predictor is significant to a response variable in a closed context, when the model becomes more complicated the significance level may change.

Step three confirms the result that relative skill is linearly significant to total wins by providing a coefficient of determination and an individual t test. The scatter plot of wins by average points scored also is linearly correlated however is much weaker than wins by average relative skill. The coefficient of correlation is around half as much, however the P value is constant. In step five we combined the average points and average relative skill predators to create a regression model that predicted the total wins given various inputs. Thep six built upon the previous regression model however added the differential variables for both points and relative skill. By doing so the coefficient of determination slightly increased from step five. However as previously mentioned in the complex regression models average relative skill is not significant. Possibly leading to why the regression with the four predictors have a lower overall f test statistic than the model with two predictor variables.