# MAT 243 Project Three Summary Report

Kristina Mohundro

Kristina.mohundro@snhu.edu

Southern New Hampshire University

## 1. Introduction

*The problem I have been asked to solve is to come up with a regression model that help predict the total number of wins for a regular game season based on performance metrics. The data set I will be using is the average relative points in a season, the average relative skill level of each team in a regular season, the average differential points between the team and their opponents in a regular season, and the average differential relative skill level between my team and their opponents in a regular season. The results will be used by the coach and the management team to help determine how many average wins they can expect this upcoming season. The type of analyses I will be running on this project will be simple linear regressions and multiple regression models for practical use.*

## 

## 2. Data Preparation

The data set that I am using will be the important variables that are being used for this project. The average differential points or, avg\_pts\_differential, compares the scores of each team and finds the highest and lowest value. The average relative skill level for each team or, avg\_elo\_n, is an average combination of teams’ performance and wins against each other.

## 3. Simple Linear Regression: Scatterplot and Correlation for the Total Number of Wins and Average Relative Skill

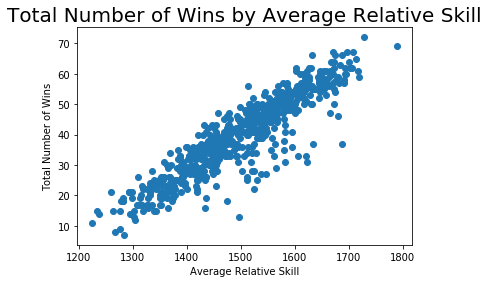
*In general, the easiest way to study relationship trends between two variables is a scatterplot. This data visualization technique is best because you can visualize the trends and correlations. A correlation coefficient that is greater than zero signifies a positive and stronger linear relationship, a correlation coefficient that is less than zero signifies a negative and weak linear relationship. Whether the coefficient is positive or negative indicates the direction and how strong and weak the relationship is.*

*Correlation between Average Relative Skill and the Total Number of Wins*

*Pearson Correlation Coefficient = 0.9072*

*P-value = 0.0*

*The p-value is 0 and the Pearson correlation coefficient is 0.09072, since the correlation coefficient is greater than zero, this indicates a strong positive relationship between the total number of wins and the average relative skill level. The correlation coefficient is statistically significant based on the p-value. Even though the correlation is 0.9072, the p-value is 0, and is less than the level of significance which is 0.01.*



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

OLS Regression Results

==============================================================================

Dep. Variable: total\_wins R-squared: 0.823

Model: OLS Adj. R-squared: 0.823

Method: Least Squares F-statistic: 2865.

Date: Wed, 21 Aug 2024 Prob (F-statistic): 8.06e-234

Time: 16:57:04 Log-Likelihood: -1930.3

No. Observations: 618 AIC: 3865.

Df Residuals: 616 BIC: 3873.

Df Model: 1

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept -128.2475 3.149 -40.731 0.000 -134.431 -122.064

avg\_elo\_n 0.1121 0.002 53.523 0.000 0.108 0.116

==============================================================================

Omnibus: 152.822 Durbin-Watson: 1.098

Prob(Omnibus): 0.000 Jarque-Bera (JB): 393.223

Skew: -1.247 Prob(JB): 4.10e-86

Kurtosis: 6.009 Cond. No. 2.14e+04

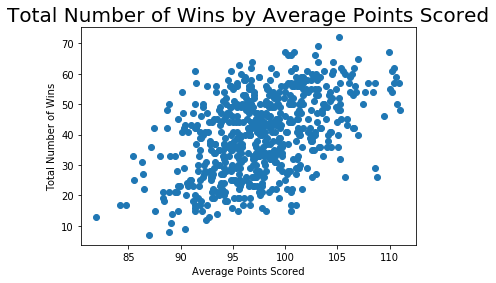
The equation for my model is Y=0.1121X-128.2475 or total\_wins=β0+β1X , where our y-intercept is -128.2475 and our slope is 0.1121. For the results of our overall f-test our null hypothesis is that a relationship does not exist between our variables, the total number of wins and the average relative skill. H0 : β1 = 0 The alternative hypothesis is that a relationship does exist between the two variables, the total number of wins and the average relative skill. H1 : β1 ≠ 0 The level of significance is 1% or 0.01. The f-statistic is 2865 and the p-value is 8.06e-234. Since the p-value is less than the level of significance there is sufficient evidence to reject the null hypothesis. The results are in favor of the alternative hypothesis that a relationship does exist between the total number of wins and the average points scored. Based on the results of the overall F-test we can predict the total number of wins in a regular season, response variable, by using the average relative skill level, predictor variable. The predicted total number of wins in a regular season for a team that has an average relative skill level of 1550 is 46 games. Y=0.1121(1550)-128.2475. The average number of wins in a regular season for a team that has a relative skill level of 1450 is 34 games. Y=0.1121(1450)-128.2475

Table 1: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2865 |
| P-value | 0.000 |

**5. Multiple Regression: Scatterplot and Correlation for the Total Number of Wins and Average Points Scored**

Looking at the scatterplot we can see that the Pearson Correlation Coefficient is 0.477 which indicates a moderate positive strength of correlation. Having a positive correlation indicates that when the average number of points increase so does the total number of wins, however with the Pearson Correlation Coefficient being lower than I thought it would, this is not always the case with the strength only being moderate and closer to the weak side. Since the p-value is 0 it is less than the 1% level of significance and therefore shows that the correlation coefficient is statistically significant.



*Correlation between Average Points Scored and the Total Number of Wins*

*Pearson Correlation Coefficient = 0.4777*

*P-value = 0.0*

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

*A multiple linear regression model is used to predict the response variable by using more than one predictor variable. For this model I will use the total number of wins as the response variable and the average of points scored and the average relative skill level for the predictor variables. The equation for my model is Y = β0 + β1X + β2X , total\_wins=0.3497X+0.1055X-152.5736. The results of the over F-test for the null hypothesis is that the dependent variable, the total number of wins, has no relationship with the predictor variables, the average relative skill level and the average of points scored. H0 : β1 =β2 =0. The alternative hypothesis is that a relationship exists between the dependent variable and at least one of the predictor variables. H1 : at least one of β1 or β2 ≠0. The level of significance is 0.01, or 1%. The test statistic is 1580 and the corresponding p-value is 4.41e-243. In conclusion of the hypothesis test we can reject the null hypothesis and say that the predictor variable all has an influence in the total number of wins. Based on the results all the variables are statistically significant in predicting the total number of wins in a regular season because each one has a p-value of 0 and is less than the level of significance which is 1%. The coefficient of determination, or R-squared is 0.837 or 83.7%, which means that 83.7% of the response variable (total number of wins) can be explained accurately by the predictor variables. The predicted total number of wins in regular season for a team that is averaging 75 points per game with a relative skill level of 1350 is 16 games. Y=0.3497(75)+0.1055(1350)-152.5736 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 51 games. Y=0.3497(100)+0.1055(1600)-152.5736*

OLS Regression Results

==============================================================================

Dep. Variable: total\_wins R-squared: 0.837

Model: OLS Adj. R-squared: 0.837

Method: Least Squares F-statistic: 1580.

Date: Wed, 21 Aug 2024 Prob (F-statistic): 4.41e-243

Time: 17:03:13 Log-Likelihood: -1904.6

No. Observations: 618 AIC: 3815.

Df Residuals: 615 BIC: 3829.

Df Model: 2

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept -152.5736 4.500 -33.903 0.000 -161.411 -143.736

avg\_elo\_n 0.1055 0.002 47.952 0.000 0.101 0.110

avg\_pts 0.3497 0.048 7.297 0.000 0.256 0.444

==============================================================================

Omnibus: 89.087 Durbin-Watson: 1.203

Prob(Omnibus): 0.000 Jarque-Bera (JB): 160.540

Skew: -0.869 Prob(JB): 1.38e-35

Kurtosis: 4.793 Cond. No. 3.19e+04

==============================================================================

Table 2: Hypothesis Test for the Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1580 |
| P-value | 0.0000 |

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

*A multiple linear regression model is used to predict the response variable by using more than one predictor variable. For this model I will use the total number of wins as my response variable, and I will use average points scored, average relative skill level, the average point differential between the teams and their opponents, and the average relative sill level differential between the teams and their opponents all as the predictor variables. The equation for my model is*

*Y=β0 + β1X + β2X + β3X + β4X, total\_wins = 34.5753-0.0134X+0.2597X+1.6206X+0.0525X. The results for the F-test for the null hypothesis is the response variable, the total number of wins, does not have a relationship with the predictor variables, the average points scored, that average relative skill level, the average point differential between the teams and their opponents, and the average relative skill level differential between the teams and their opponents. H0 : β1 = β2 = β3 = β4 = 0. The alternative hypothesis is that the response variable has a relationship with any of the predictor variables. H1 : at least one of β1  or β2 or β3 or β4  ≠ 0. The level of significance is .01, or 1%. The test statistic is 1102 and the p-value is 3.07e-278. Since the p-value is less than the level of significance we will reject the null hypothesis and be in favor of the alternate hypothesis. The results of the individual t-tests are interesting based on their p-values and the 1% of level of significance. All but one of the variables is statistically significant in being able to predict the number of wins in a regular season. The only variable whose p -value that is not less than the level of significance is the average relative skill level. The coefficient of determination, or R-squared is 0.878 or 87.8%, which means that 87.8% of the response variable (total number of wins) can be explained accurately by all the predictor variables. 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, average differential of -5, and the average relative skill differential of -30 is 26 games in a regular season. Total wins (26) = 34.5453-0.0134(1350)+0.2597(75)+1.6206(-5)+0.0525(-30). The predicted total number of wins for a team that is averaging 100 points per game with a relative skill level of 1600, an average point differential of +5 and an average relative skill level differential of +95 is 52 games in a regular season. Total wins (52) = 34.5453-0.0134(1600)+0.2597(100)+1.6206(5)+0.0525(95).*

Table 3: Hypothesis Test for Overall F-Test

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 1102 |
| P-value | 0.0000 |

OLS Regression Results

==============================================================================

Dep. Variable: total\_wins R-squared: 0.878

Model: OLS Adj. R-squared: 0.877

Method: Least Squares F-statistic: 1102.

Date: Sun, 25 Aug 2024 Prob (F-statistic): 3.07e-278

Time: 12:42:34 Log-Likelihood: -1815.5

No. Observations: 618 AIC: 3641.

Df Residuals: 613 BIC: 3663.

Df Model: 4

Covariance Type: nonrobust

========================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------

Intercept 34.5753 25.867 1.337 0.182 -16.223 85.373

avg\_elo\_n -0.0134 0.017 -0.769 0.442 -0.048 0.021

avg\_pts 0.2597 0.043 6.070 0.000 0.176 0.344

avg\_pts\_differential 1.6206 0.135 12.024 0.000 1.356 1.885

avg\_elo\_differential 0.0525 0.018 2.915 0.004 0.017 0.088

==============================================================================

Omnibus: 193.608 Durbin-Watson: 0.979

Prob(Omnibus): 0.000 Jarque-Bera (JB): 598.416

Skew: -1.503 Prob(JB): 1.14e-130

Kurtosis: 6.769 Cond. No. 2.11e+05

==============================================================================

## 8. Conclusion

*In conclusion, I used simple linear regression models and multiple regression models to predict the total number of wins for any team based on their statistics**. With this being said I can confidently predict that the higher the average skill level of the team will determine how many average points they have per game, what their differential skill level is, and what their average differential points are. All these variables play a role in the total number of wins, whether they be higher or lower will determine this total.*