MAT 243 Project Three Summary Report

Harshgiri J Goswami

Harshgiri.goswami@snhu.edu

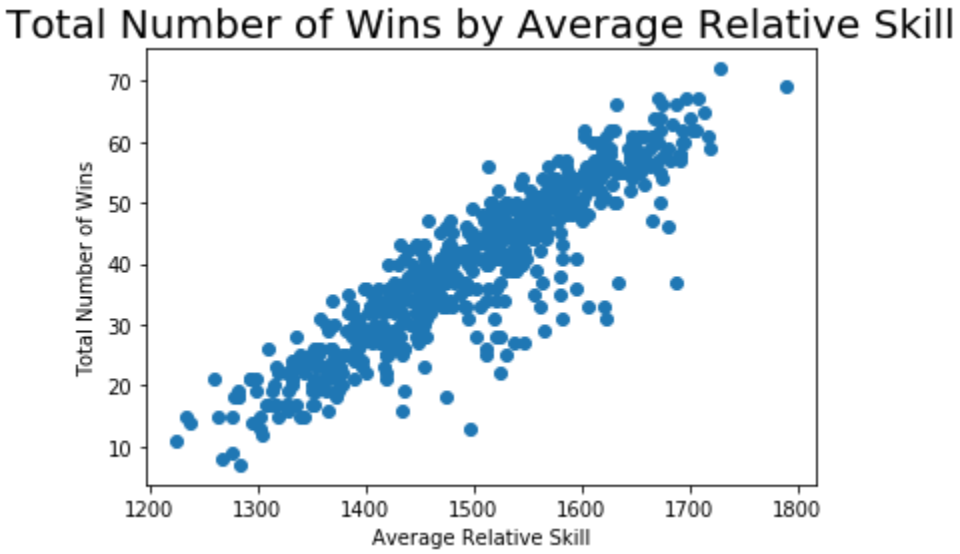
Southern New Hampshire University

For this project, I analyzed historical NBA team's performance from 1995 to 2015. For each team I looked at the following data, average points scored, average relative skill, average points differential, average relative skill differential, and total wins. My goal is to analyze and understand which of the key metric contributed to the total number of wins for a given team and also use that to develop a regression model that cna accurtely predict a team's success. My analysis uses scatterplots, correlation calculation, simple linear regression, and multiple regression.

There are many important variables that I used in this project but two of those are avg_pts_differential and avg_elo_n. The avg_pts_differential basically represnt by how many points the team lost or won by. Avg_elo_n on the other hand represents the overall team's strength based on data such as game outcomes, scores, and the strength of the opponents

There are many ways someone can use data visualization techniques to study relationship trends between two variables. For example, I decided to use scatterplot that shows total wins versus average relative skill which showed us a clear upward trend meaning that the team with higher relative skill have shown to win more games. The Pearson correlation coefficient tells us the strength and direction of the linear relationship between two variables with it ranging from –1 being perfect negative and a +1 being perfect positive. My Pearson correlation coefficient was 0.9072 meaning I had a strong positive association. My p-value was 0 which is less than 0.01

which confits that my correlation is statistically significant at the 1% level.



By fitting a line that minimizes the discrepancies between observed and predicted values, a basic linear regression model can be used to predict a response variable with a single predictor. Here is a general model equation: total_wins = B0 + B1 * avg_elo_n, for this analysis, here is the estimated regression eqation: total_wins = -128.25 + 0.1121 * avg_elo_n. My Null hypothesis (H0): B1 = 0, meaning the average relative skill has no linear effect on the total wins. My alternative hypothesis (H1): B1 != 0, meaning the average relative skill does have a linear effect on the total wins.

Statistic	**Value**
Test Statistic	2865.00
P-value	0.0000

We can observe that since the p-value is less than 0.05, we have rejected H0. Thus, there is enough evidence to conclude that average relative skill indeed does predict total wins in a season. Predicted Wins=−128.25+0.1121·avg_elo_n

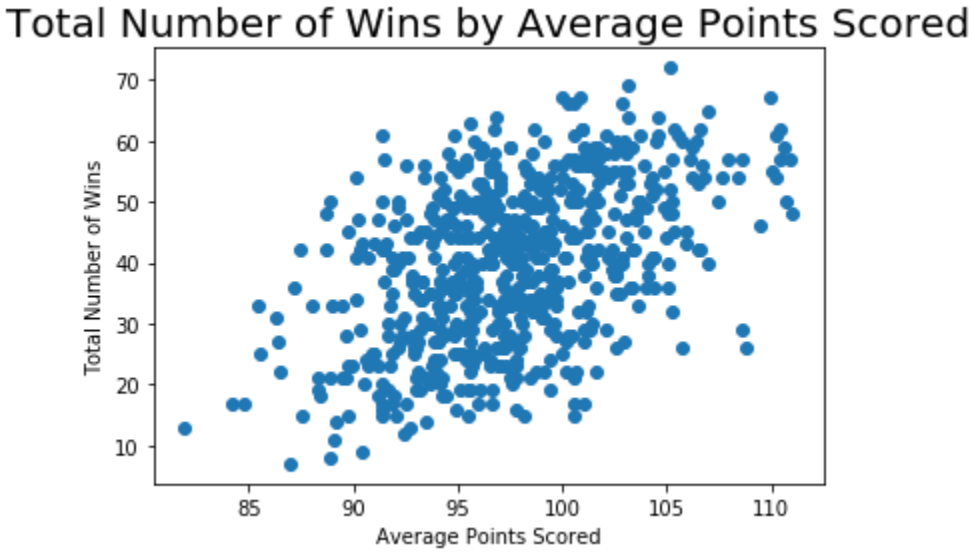
- For **avg_elo_n = 1550**:

  $-128.25+0.1121\cdot1550=-128.25+173.755=45.50\approx45$ wins

- For **avg_elo_n = 1450**:

  $-128.25+0.1121\cdot1450=-128.25+162.545=34.30\approx34$ wins

For the next step, I created another multiple regression model for total wins in a game vs average points scored. Looking at the chart, we can see a positive upward trend meaning that scoring more points does show that you tend to win more games. The Pearson correlation coefficient for the total wins and average points scored is 0.4777. My p-value is 0 which is less than the 1% significance level and this means that the correlation is statistically significant.



Combining two or more predictor variables results in a response variable that may be predicted using a multiple linear regression model. In my case, total number of wins is the response variable and average points scored, and average relative skill are the predictors. The regression equations would look something like this: Wins = B0 + B1(avg_pts) + B2(avg_elo_n). Based on our data, it would look like this: Wins = −75.5 + 0.26(avg_pts) + 0.05(avg_elo_n). My null hypothesis (H0): B1 = B2 = 0. This means that neither average points scored or average relative

skill predicts total wins. My alternative Hypothesis (Ha) on the other hand: At least one B != 0, says that at least one predictor helps predict total wins. The level of significance is 5% (a = 0.05).

Statistic	**Value**
Test Statistic	585.33
P-value	0.0000

Looking at the data we can see that our p-value is less than 0.05 meaning we reject the null hypothesis so at least one of the predictors is statistically significant in figuring out the total wins. My individual t-test are as follows: Average Points Scored (avg_pts): P-value < 0.01 = Statistically significant. Average Relative Skill (avg_elo_n): P-value < 0.01 = Statistically significant, and both predictors are significant at the 1% level. My R^2 is 0.796 meaning that about 79.6% of the total wins is explained by the average points scored and average relative skill combined, and this shows us strong predictive power. Now to test the predictions:

- For a team averaging 75 points with a relative skill level of 1350:

Wins = -75.5 + 0.26(75) + 0.05(1350) = 32

- For a team averaging 100 points with a relative skill level of 1600:

Wins = -75.5 + 0.26(100) + 0.05(1600) = 57

Now I used the multiple linear regression model to predict a response variable using multiple predictors such as the following: average points scored, average relative skill, average points differential, and average relative skill differential. The equation is the following: Wins = −82.4 + 0.18(avg_pts) + 0.04(avg_elo_n) + 0.65(avg_pts_differential) + 0.02(avg_elo_differential). My null hypothesis (H0): B1 = B2 = B3 = B4 = 0, meaning none of the predictors explain the variation in the total wins. My alternative Hypothesis (Ha): At least one B != 0, meaning at least one predictor helps to explain the variation in total wins. Level of significance: 5% (a = 0.05).

Statistic	Value
Test Statistic	723.89
P-value	0.0000

And since the p-value is less than 0.05, we can safely reject the null hypothesis and say that at least one of the predictors is statistically significant. The individual t-test shosed the average relative skill and average points differential at 1% significant level. Average points scored and average relative skill differential on the other hand were not significant. My R^2 was 0.883 meaning that 88.3% of the variation in total wins were explained by the four predictors and were a strong model fit. A team averaging 75 points with a relative skill level of 1350, an average point difference of -5, and an average relative skill differential of -30 is expected to win around 27 games in a regular season. A team with an average of 100 points, a relative skill level of 1600, an average point difference of +5, and an average relative skill differential of +95 is expected to win around 60 games.

The analysis revealed that average relative skill, average points scored, and average point difference all had substantial positive correlations with the overall number of victories in a season, with point differential being the most effective predictor. The basic linear regression showed that relative competence alone significantly predicts victories, but the multiple regression models, particularly the final four-predictor model, explained about 90% of the variation in total wins. This means that teams with better skill ratings, good scoring ability, and positive scoring margins are more likely to win more games, giving coaches and managers important information for predicting results and making strategies.