# MAT 243 Project Three Summary Report

Benjamin Leanna

Benjamin.leanna@snhu.edu

Southern New Hampshire University

In this last report I am once again taking on the role of a data analyst for a basketball team to analyze performance patterns. I will be presenting data visualization in the form of scatter plots, a t-test (looking at correlations between one variable), and a few overall F-tests (looking at correlations from multiple variables). The data used here is the same as it’s been for the last two reports.

The different data variables that I’m using in this report include ‘avg\_pts\_differential” and “avg\_elo\_differential”. I’ll break down each one of them for a clearer understanding. First the “avg\_pts\_differential”, this is the average point differential between the team and their opponents in a regular season. What this means is I am taking the ending scores between teams and their opponents for the entirety of the season and then getting the average of that number. The “avg\_elo\_differential” is the same except I am taking the difference in average skill level between all teams throughout the season.

![Chart, scatter chart

Description automatically generated]() For my data visualization technique, I’ll be utilizing a scatter plot. The way it is designed is to naturally take the dependent variable along the x-axis and the response variable along the y-axis. Using this feature allows someone to take a quick look at the plotting to see if they can observe an obvious trend happening between these variables. Along with the scatter plot, the Pearson correlation coefficient was also calculated. This common formula is used in linear regression to determine if there is a relationship between the variables. With ranges between -1 and 1, where -1 is a strong negative, 0 is no relationship, and 1 is a strong positive, our coefficient is 0.9072; This shows a strong positive correlation.

With a strong positive correlation, this tells us that the relationship between a team having a high average relative skill level would normally also have a high number of wins. Looking at this along with a p-value of 0.0 in relation to a 1% significance level, this is showing a very strong linear relationship. This means that we can trust the coefficient with very high confidence.

With this same information, I also created a summary of a simple linear regression model (SLR) in an overall F-test. The SLR is a type of model used to describe the relationship between 2 variables, a dependent and independent, by fitting a line to the observed data. Using this in an overall F-test, it allows us to assess multiple coefficients simultaneously. Here is the summary:

![Table

Description automatically generated]()

I used this information to base against the null and alternative hypothesis of the coach. He expects a team to have more wins in a season if it maintains a high average relative skill during that season. The null hypothesis being that there is no relationship between the average skill level and number of wins and the alternate being there is.

Using this data with the equation , we’ll fill it in. ‘Y’ is considered the outcome, or the dependent, while is the intercept, or the independent, is ‘avg\_elo\_n’, and is the multiplying variable of our choice. With all this known beside our chosen variable, the equation becomes . Our p-value here is also 0.000. When this is put against a normal 5%, or even 1% significance level, we are still seeing evidence to reject our null hypothesis and see there in fact is a strong positive linear relationship between our two variables.

Using this model, which is 82.3% positive, we can predict a total number of wins in a regular season. If a team has an average skill level of 1550, then the number of wins becomes around 45.5. If a team has an average skill level of 1450, then that number lowers to 34.3. We can look at how accurate this equation is be simply putting in actual information and comparing it to our equation model. For example, in 1995, the Chicago Bulls had an average skill level of 1569.89. Plugging this into our equation, the number of wins they should have had is 47.74. In reality, the team did have 47 wins that season. This tells us that we should have confidence in the model.

The next section printed is another scatter plot. This time I used the average points that the team scored (instead of average skill) against the total number of wins. The plot looks like this:

![Chart, scatter chart

Description automatically generated]()

Right away, you can tell by looking at this plot compared to the one based with average skill, this graph does not have as tight of groupings but is still showing a positive correlation. With the Pearson’s R calculated at 0.4777, it indeed is a positive relationship, albeit moderate instead of strong. Still, putting this against a 1% level of significance, we can still have confidence in the linear correlation since our p-value is below it at 0.0.

![A close-up of a document

Description automatically generated with medium confidence]() In the next section here, I am calculating a multiple linear regression model. This means that instead of using just the one predictor variable, I will be using multiple. A multiple variable model is used to see if we can get a better read on the overall correlation from the prediction to the response.

From this summary, we can rework our original equation to since we now have a second predictor variable, denoted as , and our input to that as . Our new equation becomes . First, I want to look at each individual predictor to see if their p-value is below our normal 5% level of significance. Since both of theirs are 0.000, I will include them in this model. Next, I’ll be taking the overall F-test’s p-value (located after the ‘Prob (F-Statistic)’) as . This along with the (otherwise known as the coefficient of determination) being 83.7% positive, I would confidently conclude that this model using both average points and average skill level would be great to use. To test this, we can use a few inputted values in points and skill. If a team is averaging 75 points a game with a relative skill of 1350, the number of wins in a season should be 16 and a team averaging 100 points per game with an average skill of 1600 should win 51.

This next multiple regression model is one where I now took all four calculated stats, average points, average relative skill, average points differential, and average relative skill differential. The summary on this is just a bit different. Here it is so I can explain:

![A close-up of a document

Description automatically generated with medium confidence]()

Looking at this summary now, the individual predictors are now throwing off the model. We can see specifically that the p-value of the average relative skill in this model is 0.442. This isn’t bad. Our null hypothesis would be that and the alternate being that at least one of the . We can see that at least one of our p-values (average points, points differential, or average skill differential) are lower than our significance level of 1%. This means that we would reject the null hypothesis and state that in fact there is a linear relationship between all four variables and the response outcome. This makes the final equation . We can use this now to calculate a team that scores an average 75 points, with a relative skill of 1350, average point differential of -5, and an average skill differential of -30 to have approximately 26 wins in a season. A team averaging 100 points, with a average skill of 1600, point differential of 5 and average skill differential of 95 to win approximately 52 games.

In conclusion, this was a very good chance to look over the differences in results of performing individual t-tests with a simple linear regression model compared to doing an overall F-test using a multiple regression model. One can see that each time we used a multiple regression, the determination coefficient () was greater than the simple linear models. While thinking about why anyone would use a simple model over a multiple model, perhaps the data collected only includes the two different variables or the similarity between more than the two variables makes the outcome skewed. In our last test, the p-value of the average skill was quite high (0.442). This could indicate multicollinearity between two or more of the variables and show numerical problems. Overall, being able to see different ways that variables can be statistically calculated means we can choose the best one for our specific problem we’re working out.

## 9. Citations

Bevans, R. (2022a, June 1). *Simple Linear Regression | An Easy Introduction & Examples*. Scribbr. https://www.scribbr.com/statistics/simple-linear-regression/

Frost, J. (2022, July 22). *How to Interpret the F-test of Overall Significance in Regression Analysis*. Statistics By Jim. https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/