**Estimating the Run Values of Different Batting Outcomes Using the Cobb-Douglas Production Function**

**The Background**

When evaluating a player’s offensive value for trades, there are many new metrics that can be used to gauge a player’s value. The most common of which are offensive WAR, OPS, and WRC+. All of these are useful when comparing players to one another, however, they do not give much insight into how they may help a specific team. There are a variety of different ways players can achieve high statistics in these categories, whether they are a contact hitter, a slugger, or someone who controls the zone and walks at an elite rate.

What this does not address, is that an offense needs a diverse set of hitting styles. A solo home run has much less impact on a game than a multi-run homerun. A multi-homerun not only has the potential to put the game out of reach score-wise but is also extremely demoralizing. As Earl Weaver said, “The key to winning baseball games is pitching, fundamentals, and three-run home runs”. This means that a team cannot solely consist of sluggers who may slug over .450, however, have an OBP of less than .300. On the other side of the coin, teams need extra base hits to score runs consistently. Research has proven that home runs lead to more wins. With that being said, it is important to find a balance. Singles make extra-base hits more productive, and vis-versa.

The question then becomes, what does the team need more of? Teams will get varying degrees of productivity from the different hits based on how the team is currently built. It is possible to figure out what type of batter is needed, however when comparing batters within the same category or batters who may be in-between categories it would be useful to quantify who would bring more run production to the team. This is where the Cobb-Douglas production function comes in. The function in an economic sense is as follows:

Y is the economic output, measured in the real-value of goods. A is the total factor productivity (TFP), which includes the assumptions on technology and other factors that affect the productivity of labor and capital. The TFP scales labor, L, and capital, K. Beta and alpha are the output elasticities of their respective inputs. If an output elasticity is less than 1, this means that there is a decreasing marginal return on that input. An output elasticity of 1 is a linear return, and greater than 1 is an increasing marginal return on the input.

What makes the Cobb-Douglas function useful is that it is an interaction term of multiple inputs, and more than two inputs can be used. This means that the inputs are complementary, so the marginal productivity is dependent on the value of other inputs. This has been used in business cases for problems such as finding the optimal amount managers and associates at a retail store. Managers bring more value, however, are more costly and their added value is needed less the more managers that are present. For this reason, the most profitable combination of associates and managers varies based off of productivity estimates. It is for these reasons that the Cobb-Douglas function could be modified to be used for expected runs over the course of a season based on the production of outcomes from an at-bat.

**The Baseball Application**

In order to apply this function to baseball, some modifications need to be made. In order to do this, I propose two possible equations:

**Function 1:**

**R:** The predicted amount of runs over the course of a season

**E:** The run-scoring environment

**S:** Total Singles

**X:** Total Doubles and Triples

**H:** Total Home Runs

**B:** Total Base on balls and HBP

**Β, α, λ, δ:** run-production elasticities

**Function 2:**

**R:** The predicted amount of runs over the course of a season

**E:** The run-scoring environment

**S:** Total singles

**D:** Total doubles

**T:** Total triples

**H:** Total home

**B:** total base on balls and HBP

**β, α, γ, λ, δ:** run-production elasticities

Both of these equations estimate the number of runs scored over the course of a season based on the different outcomes from an at-bat. They use the same principles of the Cobb-Douglas function that each outcome affects one another, and uses exponents to modify how productive an additional occurrence of an outcome is. The difference is in the number of outcomes being used to create the estimate. Both of these equations have their merits and should be tested. It is unclear yet the size of the data set, and by bucketing similar outcomes it may lead to a more accurate prediction as rarer outcomes such as triples may be weighed incorrectly on their own due to their rarity. The second equation, however, is more precise as it evaluates each outcome individually.

**Reasoning and Theory**

Through prospective research, I have found similar studies using similar functions to predict wins in baseball, however, I have not seen the use of this style of function for run prediction. I believe that this is grounded in sound reasoning as the economic-baseball parrels are logical. Runs are equivalent to real value in economics as they are a way to measure the offensive outcome of a team. The run scoring environment is akin to the TPF as it is a constant that has an effect on the productivity of outcomes. In a lower run-scoring environment, the productivity of extra-base hits, singles, and base on balls is going to be lower as the likely hood of batting someone in or being batted in is lower. The run-production elasticities are good parallels for output elasticities as they estimate the amount of additional runs created through an additional outcome, and in economics, the output elasticities estimate an percent change in real value based on an additional unit of input. The background presented the argument as to why an interaction function is useful in this situation, as each outcome has a different productivity based on the number of other outcomes.

**Data and Empirical Design**

The data that will be used is team batting statistics by season from fangraphs. This data was chosen as it is easy to export and included all necessary variables including the batting outcomes, and runs scored by a team in that season. This data also should not need to be cleaned with the exception of creating a variable called “Free Bases” which is the sum of BB, IBB, and HBP. The year 2020 will also not be used as that was the covid year, and may introduce bias into the model.

In order to assess the hypothesis that the Cobb-Douglas function can be applied to baseball in order to predict runs scored per season, I will be using linear regression to predict the output elasticities for each batting outcome. In order to do this, it is required to first take the natural log of runs scored and all batting outcomes. This will change the interpretation of the regression to one in which a one percent change in one of the independent variables is correlated with a percent change in runs scored that is equivalent to the coefficient. This is the definition of output elasticity, meaning this coefficient will in turn be used to predict the runs scored in a season.

After estimating the output elasticities, it is possible to predict the runs scored in a season using the aforementioned equation.

This can be done by plugging in the estimated elasticities into their respective superscript, then using actual batting outcomes for each team to predict their runs scored.

**Measures of Accuracy**

In order to know whether or not this method is successful it is important to create a few measures of success. The methods used will include statistical significance of the model and each independent variable, along with measures of accuracy of the predictive model.

This first will be statistical measures. These will include p-values, t-statistics, and the F-statistic of the regression. The p-value and t-statistic will be used to measure whether or not it is possible to reject the null hypothesis that each batting outcome does not have an impact on runs scored. In order to reject this null hypothesis and say that the batting outcomes do have an effect runs scored in a season, the p-value will have to be less than .05. This is a commonly used number that means that it can be said with 95% certainty that the coefficient is statistically different from 0. The t-statistic is the coefficient divided by the standard error. The t-statistics will need to be 2 or greater. This means that the effect will be statistically significant from 0 as there is no chance that when including the standard error the effect is 0. The F-Value is a way to evaluate the statistical significance of the model as a whole. If the f-value is greater than the critical value, then the model is jointly statistically significant. The critical value for this regression is 3.174. This is found using an f-distribution table and using the regression and residual degrees of freedom. For this regression, the regression degrees of freedom is 5 and the residual degrees of freedom is 174.

Besides statistical significance, it is important to measure how well the model works in the real world. In order to measure this, backtesting will be used. Important metrics include the average error, the average error percentage, the min and max of the errors, and the percentage of predictions that are within certain error ranges.

The backtesting will be done on the sample done, along with data from 2012-2016. The years 2012-2016 were chosen as it provides well over 100 data points to backtest on, and is close enough in time to how modern baseball is played. By this time sabermetrics had started to be used, the idea of “money ball” had been adopted by many teams, and the way of thinking about the game is similar to how it is in the current age. The average error will be calculated by taking the average of the absolute value of the predicted runs scored – the actual runs scored. Absolute value is used for two reasons. The first is the errors can be either positive or negative, the second is that linear regressions by design have an average error of 0 using R2.

**Results**:

The results of the regression are promising after running both regressions. The regression I feel is best is the first regression where doubles and triples are summed. This is because the p-value on triples was not significant at the 95% CI. This was most likely due to the lack of variation in amount of triples, and the luck factors that go into triples. The expected run values of triples are also fairly similar to that of doubles, so combining seems like a reasonable action to take.

The f-statistic of the second regression was 350, well above the critical value of 3.174. This means that this regression is jointly significant. All of the p-values were less than .01, meaning that they were all significant at the 99% confidence interval. All t-stats were greater than 2.0, reinforcing that all variables were individually, statistically significant. The regression also had an R2 of .904. This means that 90% of the variation in the runs scored by a team is explained by this model.

After backtesting, there is also reason to believe this model can be used for estimating runs scored over the course of a season. The averge error was 3.33%, and the average residual was at 22.32 runs. The median of these numbers were 2.35% and 15.56 runs respectively. This leads me to believe that the errors are biased by extremes. The max residual and error was 84.42 runs, or 12.04%. Another way of analyzing this is by looking at the averages and medians of difference in runs between teams. This includes both league wide differences, and differences within the top 10 teams.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Run Differential | | | | | |
| All Data | | | Not Including 2015 | | |
| League Wide | **0.50** | **5.10%** | 9.4 | **0.9** | **9.57%** |
| Top 10 | **-0.68** | **-5.19%** | 11.1 | **1.3** | **11.56%** |
| Top 15 | **-0.25** | **-2.32%** | 9.7 | **0.7** | **7.33%** |
| **Runs Scored** | | | | | |
| All Data | | | Not Including 2015 | | |
| League Wide | **58.00** | **8.52%** | 678.3 | **60.5** | **8.91%** |
| Top 10 | **70.80** | **9.47%** | 748.0 | **70.8** | **9.47%** |
| Top 15 | **67.20** | **9.23%** | 727.3 | **67.9** | **9.33%** |

The graph above depicts the difference in the run scoring environment between thet time period that the model was trained on, and the back testing time period. As you can see the period that the model was trained on had a fairly significant increase in the run scoring environment compared to the time period that was used for back testing. The average amount of runs scored increased by almost 9%, the average difference between teams increased by 5% at the league wide level when including 2015, and almost 10% not including 2015. The reason I chose to also look at the back testing period without 2015 is due to a significant outlier. The Blue Jays scored 127 more runs than the team with the second most runs. The next largest difference was in the 40s. This had a significant effect on the run differential averages, and therefore shoulld not be used.

These results mean that the backtesting period may not have been as great of a choice as I had previously

thought. This will still be used to assess the accuracy of the model, however, I will talk about other methods I would like to use later in the “Next Steps” portion of the paper.

|  |  |
| --- | --- |
| Period Average | |
| Run Differential | |
| League Wide | 10.3 |
| Top 10 | 12.4 |
| Top 15 | 10.4 |
| Runs Scored | |
| League Wide | 738.8 |
| Top 10 | 818.8 |
| Top 15 | 795.2 |

|  |  |
| --- | --- |
| Period Averages | |
| League Wide | 9.8 |
| Top 10 | 13.1 |
| Top 15 | 10.6 |
| Runs Scored | |
| League Wide | 680.8 |
| Top 10 | 748.0 |
| Top 15 | 728.0 |

This table depicts the run scoring environments