|  |  |
| --- | --- |
| Final Report: An Analysis on Baseball | Nour Saidane: 20037807 Vratislav Havlik: 20038753 Michael Krakovsky: 10134030 Busra Papila: 20039559 Joshua Low: 20042150 Zhang Feiran: 20042032 |

Table of Contents

# Introduction………………………………………………………….….2

# Modelling Methods- Logistical Models………………………………...2

### 2.1 Interpreting Coefficients………………………………………………………...3

# Reading Baseball Statistics……………………………………………..4

# Acquiring the Data……...………………………………………………5

# Predicting a Winning Team…………………………………………….7

# Impact of Payroll and Fan Attendance………………………………….8

# Impact a ‘Slugging’ Strategy……………………………………………10

# 7.1 Classifying the Teams……………………………………………………………10

## 7.2 Strategy Dependent on Ball Parks………………………………………………..12

## 7.3 The Final Model…………………………………………………………………..13

# 8. Utilizing the Information…………………………………………………14

# 9. Conclusion………………………………………………………………..15

# 10. References………………………………………………………………16

1. Introduction

Baseball is a sport played by two types of individuals: the players and the statisticians. After the publication of Michael Lewis’s famous book *Moneyball*,the sport has evolved into a game oriented in statistical analysis and evidence-based decision making. Players are no longer selected based on “gut feeling” but mainly on the stats that describe their performance. Individuals who were overlooked by the conventional methods of scouting can no longer be overlooked through statistical analysis because numbers possess no bias. Nevertheless, the type of players within baseball has not changed within this transformation; however, the way the players are selected and the type of strategies they use has changed dramatically.

Our paper will look to examine the basic statistical measures that dictate the outcome of a baseball game. Furthermore, we will analyse other indicators that may not be as obvious to the reader. The report will be split into three main sections. First, we will create a baseline model out of the fundamentals statistics of baseball that will determine the winning team. Second, we will attempt to show the impact of the team payroll and fan attendance on winning percentage. Finally, we will attempt to identify the impact of ‘Slugging’ strategies on your team’s ability to win games.

The knowledge acquired from the models will give us valuable information regarding the likelihood of whether a team will win or lose. With this information, we can perform two actions. First, we can act as consultants to the management of baseball teams to explain the impact of these strategies on their team’s performance. Otherwise, we can use the information to create a detailed model that predicts a winning team allowing us to bet against the odds set forth by bookmakers.

1. Modelling Methods- Logistical Model

When defining a logit model, we will first explain the linear probability model. A linear probability regression model is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  |  |

is a constant variable, is a vector of observed explaining the variables and is a vector of unknown regression parameters. When deriving the values of estimators of explaining variables in the interval <0,1>, we need to include a chance of the situation, where the outcome will be true. This is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Next, we transform the model above using a logarithm to obtain a logit model. When interpreting the results, we invert this relationship back to get real values. When we estimate a logit model, we need to use a logit transformation of a chance:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where is a vector of explaining variables and is a vector of unknown parameters. A mean value of an estimated variable is defined as a non-linear function of explaining variables. The previous transformation is used here again:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |
|  |  | (5) |
|  |  |  |
|  |  |  |

Due to this transformation, a constraint for estimated explaining variables from an interval <0,1> is now satisfied. For estimating such a regression, we use the maximum likelihood method.

## 2.1 Interpreting Coefficients

By using the maximum likelihood method, we obtain a vector , where is a level constant and vectors stand for coefficients connected to the variables in a model.

Parameters in a logit model are not so simple to explain because of a logit transformation. Logit represents a logarithm of a chance that the explained variable (here winning a game) will be true. A value of *β*0 gives an information of the mean value under a condition that all other explaining variables have a 0 value. If *β*0 > 0, a chance of (Y = 1) is greater than 1, thus *π* > 0.5. For *β*0 < 0 explains a chance lower than 1 and *π* < 0.5. For *β*0 = 0 is a chance equal to 1, thus *π* = 0.5. Therefore, when we one variable increases its value by one the chance of the response variable will increase *e βj* times. Furthermore, if *β*k > 0 for variable *k*, the probability of winning will increase because of the *k* variable, otherwise it will decrease.

To evaluate the model, we use two levels of evaluation. First, we look at variables and decide whether they are significant or not. Second, we evaluate the model in its entirety. The method for testing a significance coefficient is like the simple regression model. The null hypothesis is that a parameter is equal to 0 at a 5% level of significance. According to the p-value we reject or fail to reject our hypothesis about non-significance.

When evaluating the entire model, we use the McFadden Pseudo R2 and the Chi-square test for testing insignificance of the model. Pseudo R2 is analogous to R2 in OLS regression, the only difference here is that it does not use the sum of the squared residuals, but uses instead the log-likelihood statistics. The result means that we have xx % fits between our prediction and an actual data.

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Finally, we will use a percentage correctly predicted cases which tells the accuracy of our model.

1. Reading Baseball Statistics

There are hundreds of measures to describe the events in baseball; however, we will only focus on the basics since they create a very accurate model while more advanced statistics are extremely difficult to acquire. Here is a comprehensive list of the statistical measures we will utilize[[1]](#footnote-1):

**At Bat (AB):** Signifies a batter’s turn against a pitcher. A batter will not receive the credit for an At Bat in the event the batter receives a sacrifice fly, a sacrifice hit, a hit by pitch, or a walk.

**Caught Stealing (CS):** When a base runner attempts to advance in the absence of batted ball and is tagged out by a defender before he reaches another base.

**Grounded into Double Play (GIDP):** When multiple outs occur on the bases after a hit ball. For the report, we simplified the statistic and calculated GIDP to include the event which two outs occurred in a single AB.

**Hit by Pitch (HBP):** When a batter is hit with a baseball and is automatically allowed to advance to first base.

**Sacrifice Fly / Sacrifice Hit (SF / SH):** When a batter hits a fly ball into the outfield that allows a runner to score. A sacrifice fly does not count towards an eligible at bat; nonetheless, the run still counts.

**Single (1B):** When a batter hits a ball and successfully reaches first base.

**Double (2B):** When a batter hits a ball and successfully reaches second base.

**Triple (3B):** When a batter hits a ball and successfully reaches third base.

**Home Run (HR):** When a batter hits a ball and successfully scores on the same play.

**Strikeout (K):** When a pitcher throws three strikes to a batter resulting in an out.

**Walks (BB):** When a pitcher throws four pitches outside of the strikeout allowing the player to freely reach first base.

**Stolen Base (SB):** When a baserunner advances in the absence of a batted ball and successfully advances to the next base.

**Ballpark Factor (BF):** A unique baseball statistic that indicates the effect of a ballpark on an offensive player’s batting statistics.

1. Acquiring the Data:

The data was obtained from the website retrosheet.org, which is an open source data collection website that offers its information without charge. The initial data file contained over 190’000 data entries each with 36 columns. Each separate data entry was designed to describe every event within a baseball game throughout the entire 2016 baseball season. For instance, the first column provides a unique game code where the event took place. Other information within the file informs the user about the players involved within the event, the type of play that occurred, and even the location of where the ball was hit. The following images shows the process we went through to acquire the data. The first image is the raw data obtained from retrosheet.org:

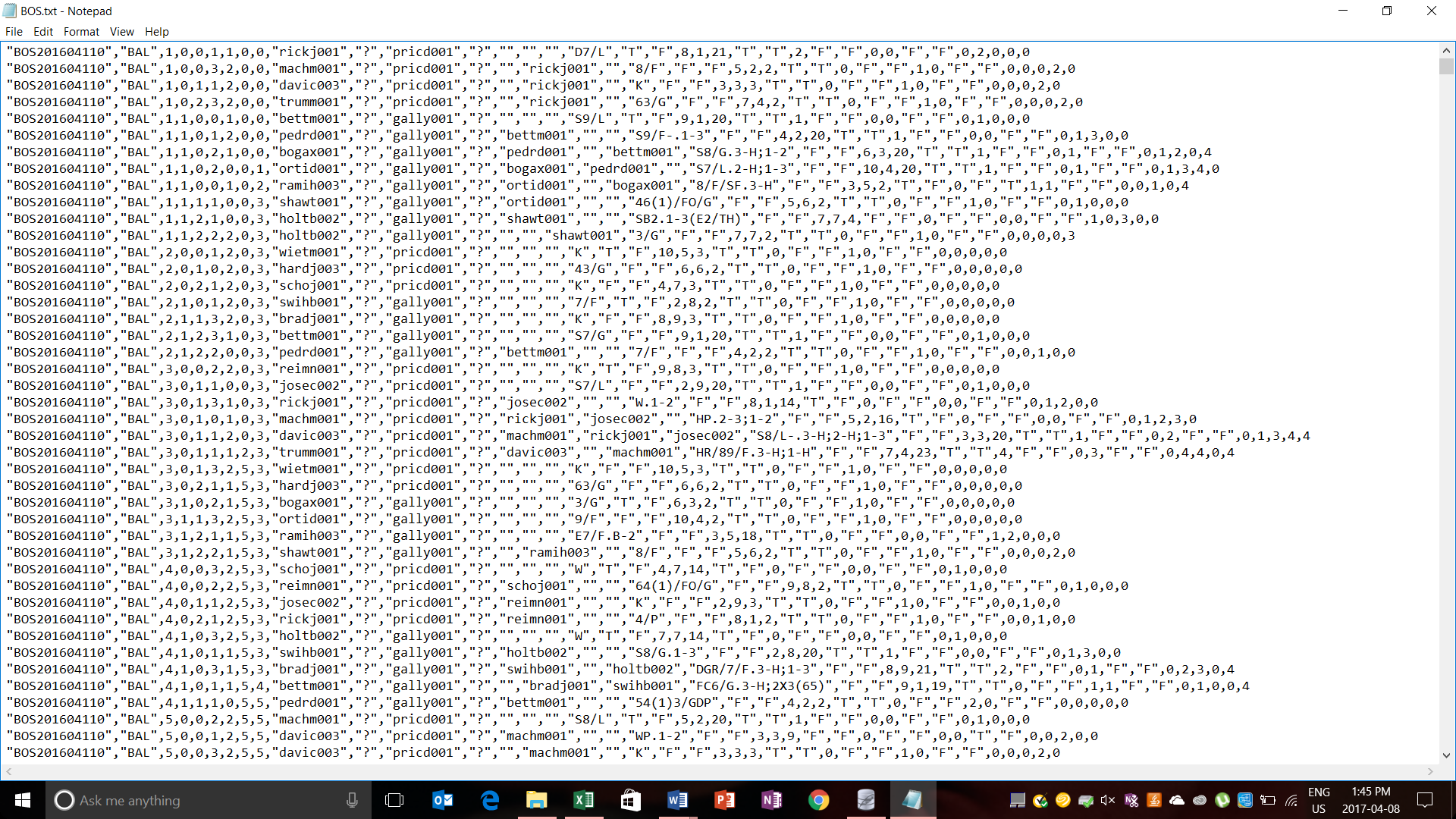


Figure 1: Raw data downloaded from retrosheet.org

Second, by utilizing a custom-made program within python, we could parse the data and insert the newly formatted data into an Excel sheet. The data was then imported into SQL allowing us to aggregate the data for every game by utilizing queries:

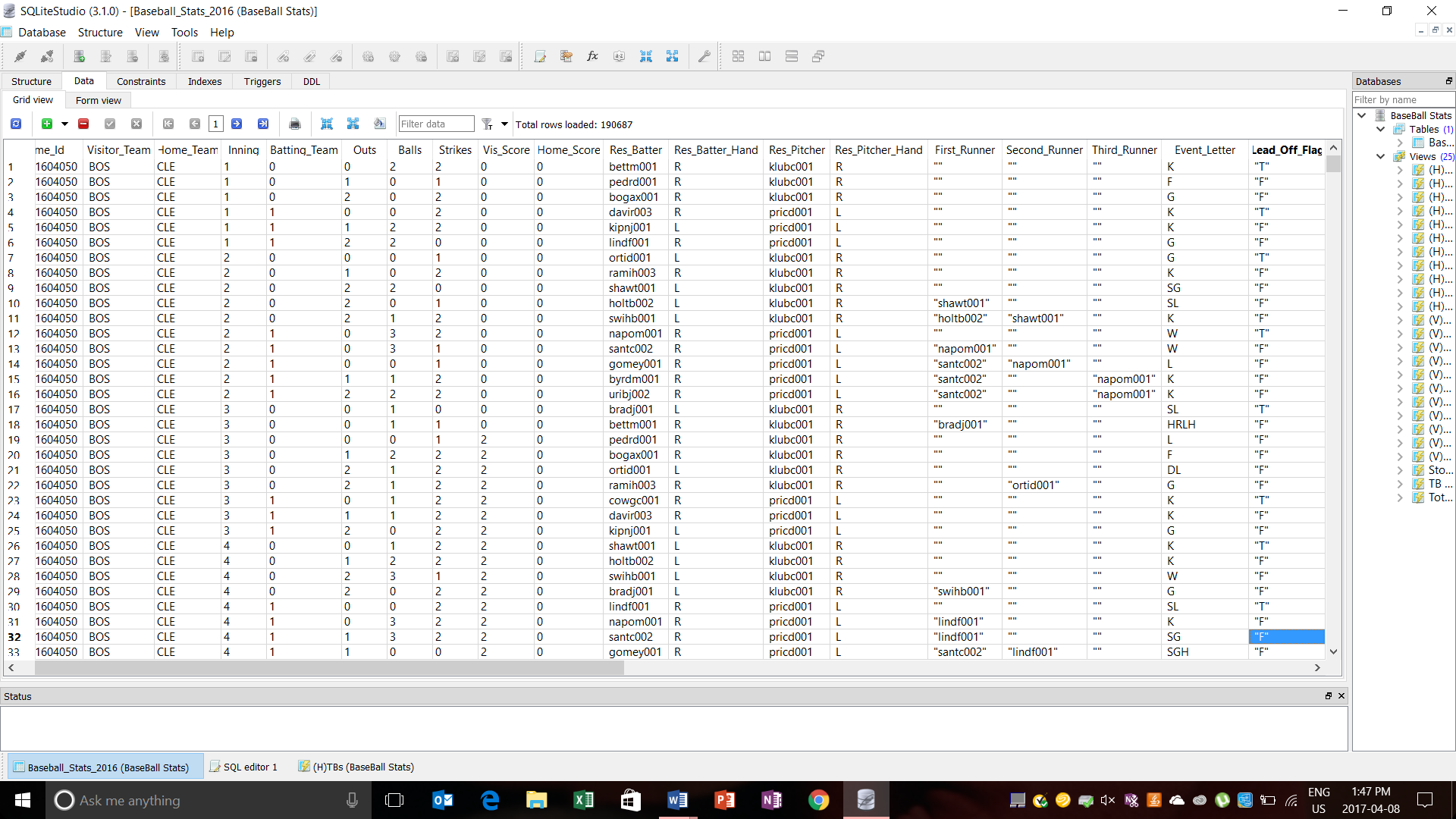


Figure 2: Parsed data ready to be aggregated in SQL

Finally, the data was exported from SQL into an Excel sheet to be analysed. The following image illustrates the transformation of raw data into a usable Excel file containing 2430 data entries, which represents all the games played in the 2016 season:

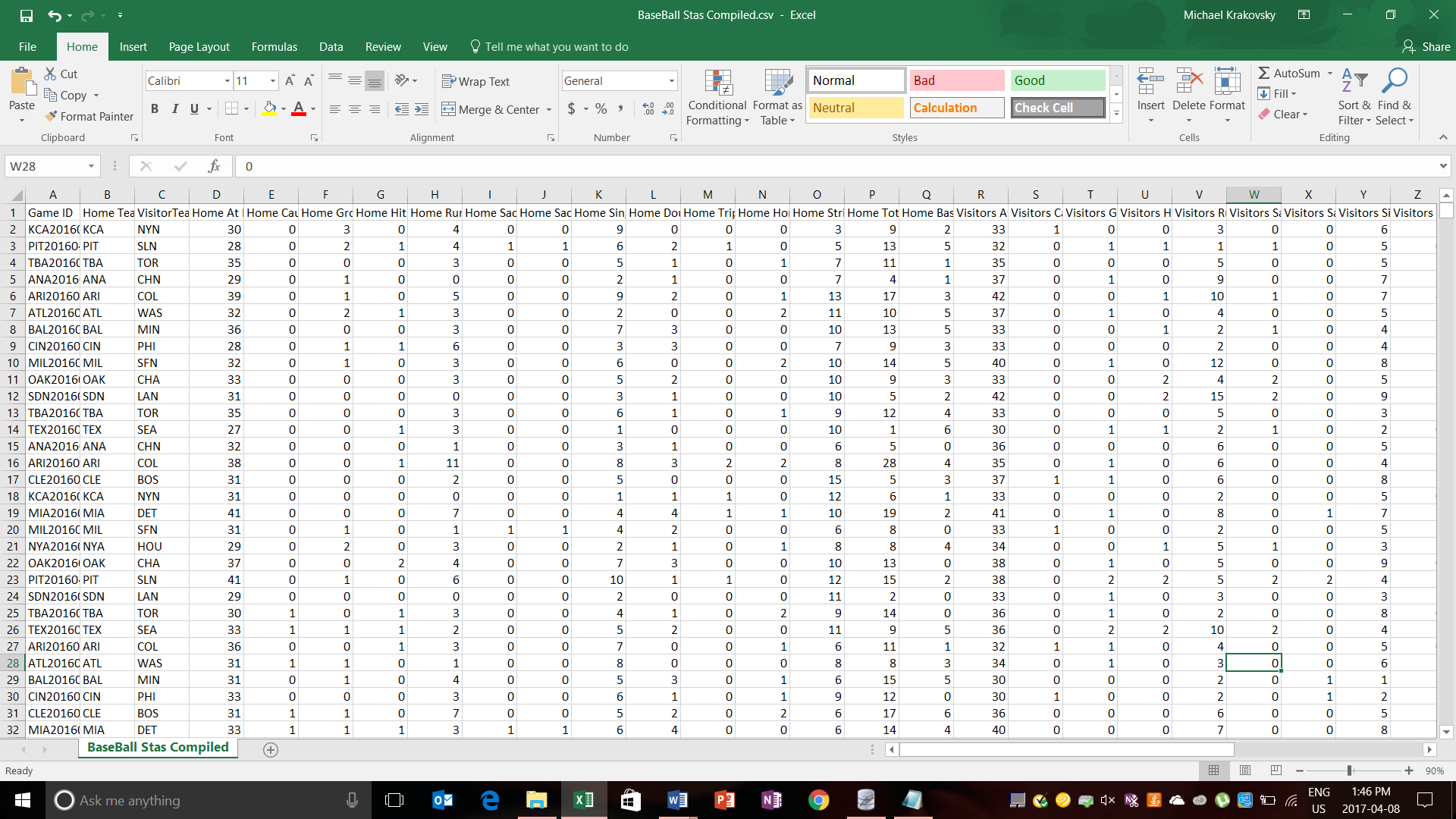


Figure 3: Final Excel sheet used in Gretl, R and Stata

Formatting the data into an Excel file allowed the data to be imported into three main software programs where we performed our statistical analysis. These programs were Gretl, Stata, and R. In the interest of time, we will only display the results from one of these programs.

1. Predicting a Winning Team

The first model we decided to create is intended to find the determinants of a winning team. Therefore, we modelled whether the home team won or lost against the determinants mentioned above. Since runs scored against are just as important as runs scored, we included both the home team’s performance and the visitors team’s performance within our model. In addition, we included a holdout period of 400 games to test the validity of our model.

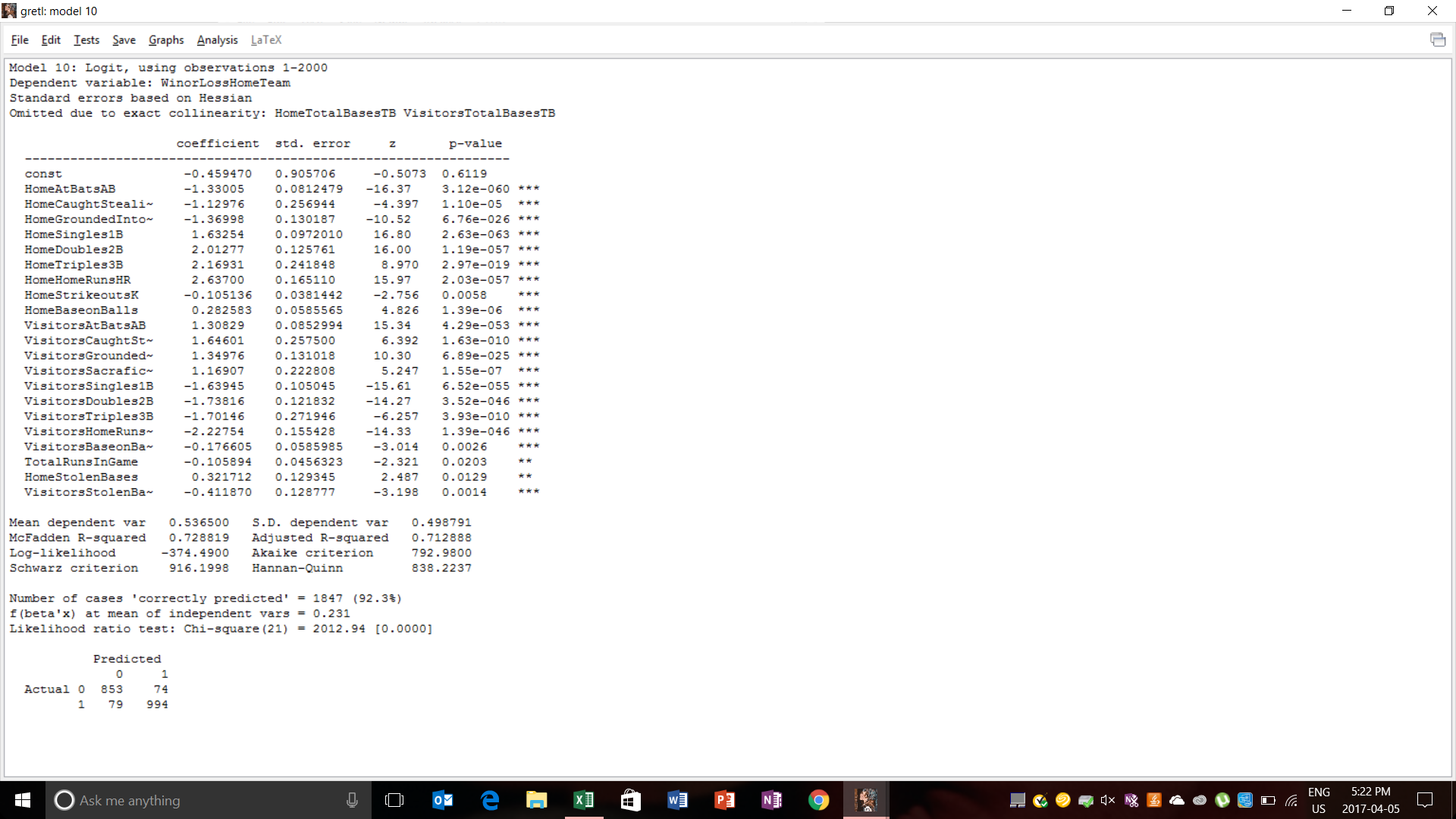


Figure 4: Baseline model to predict the winning team

The model was created under the criteria to maximize the log-likelihood. Both the Adjusted R-squared and McFadden R-Squared were above 70% thus indicating a decently strong model. By utilizing the model created above, we attempted to predict the final 400 hundred games of the year based on the subsequent indicators.

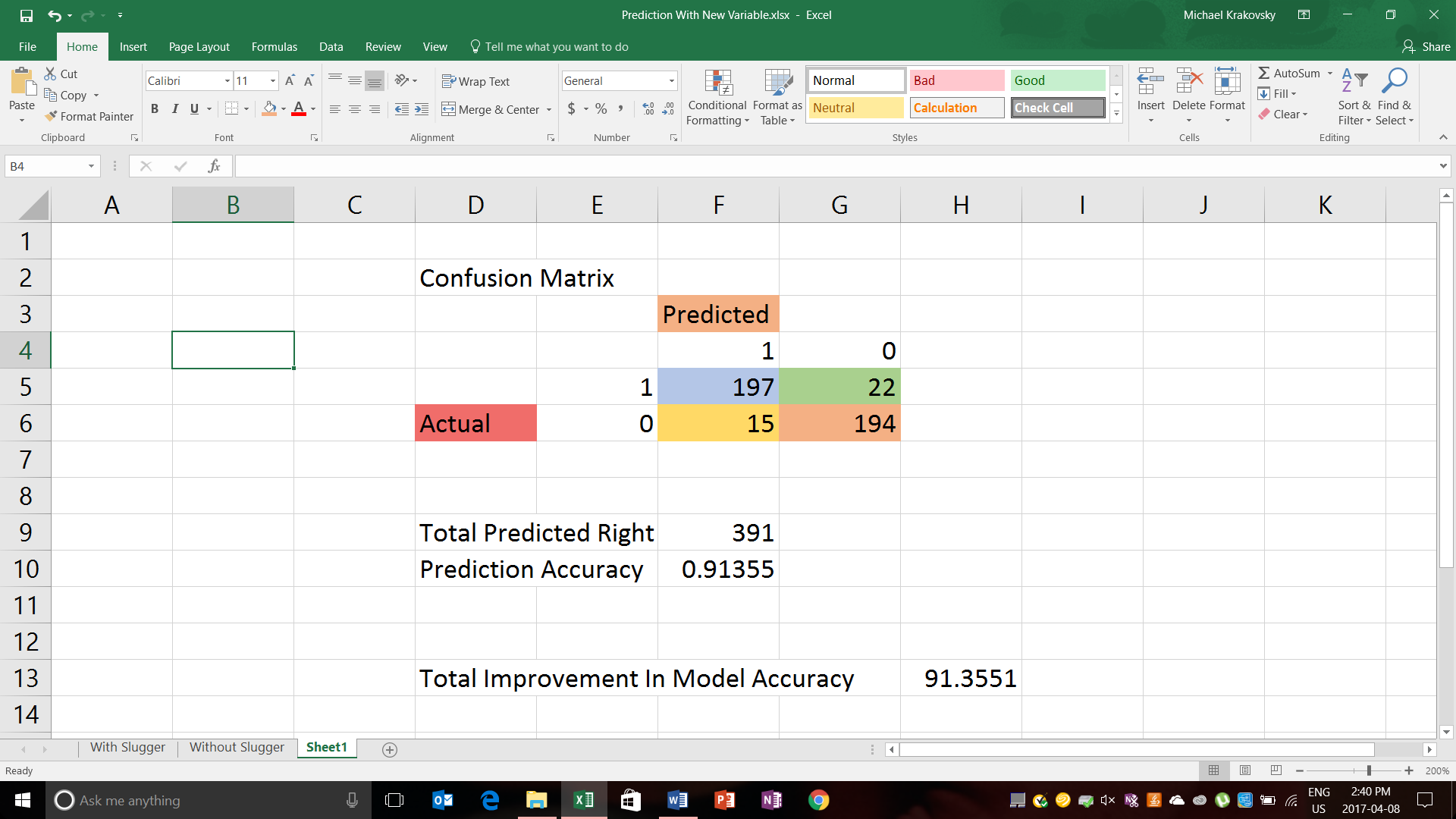


Figure 5: Results from the holdout period

The numbers that are highlighted blue and orange indicate the games that were correctly predicted. The number that is highlighted yellow indicates the number of false positives in the game. Therefore, the home team was predicted to win, however, the actual result was a loss. The number highlighted in green indicates the number of false negatives. A false negative in our model indicates that the home team was predicted to lose although the game ended with a home team win. The total accuracy of the model was 91.35% indicating a sufficient model. Next, we will attempt to improve the model even if by a couple percentage points.

1. Impact of Payroll and Fan Attendance

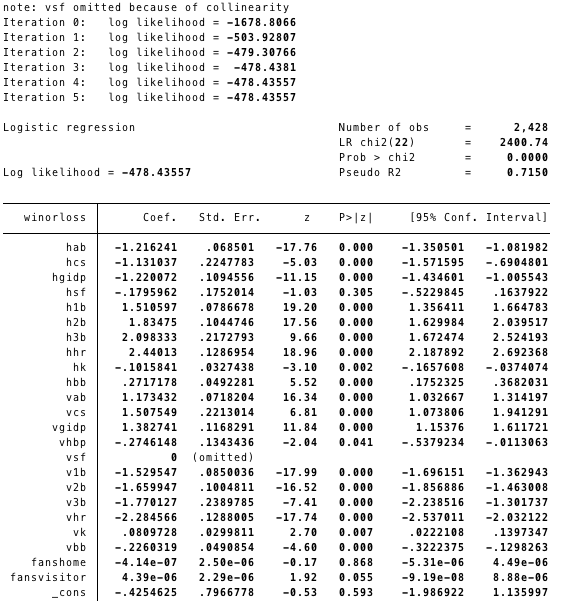
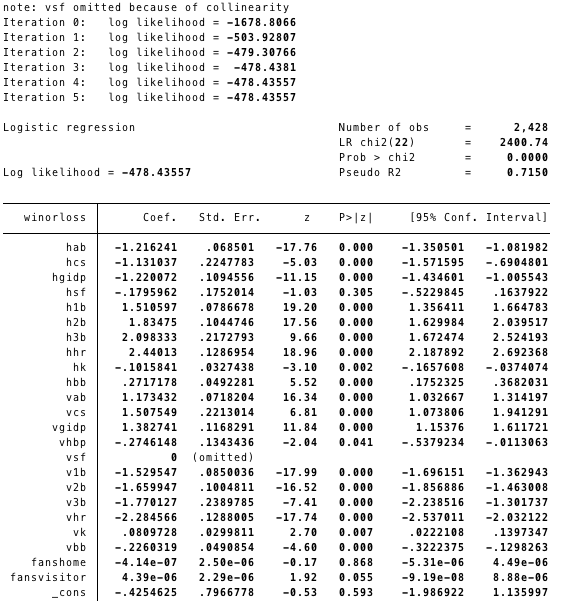
To study the influence of the teams’ payroll and attendance on the result of match we ran a logit regression on these factors along with the factors we defined in the first simple model. The results we got are as follows:

Figure 6: New model with only attendance data



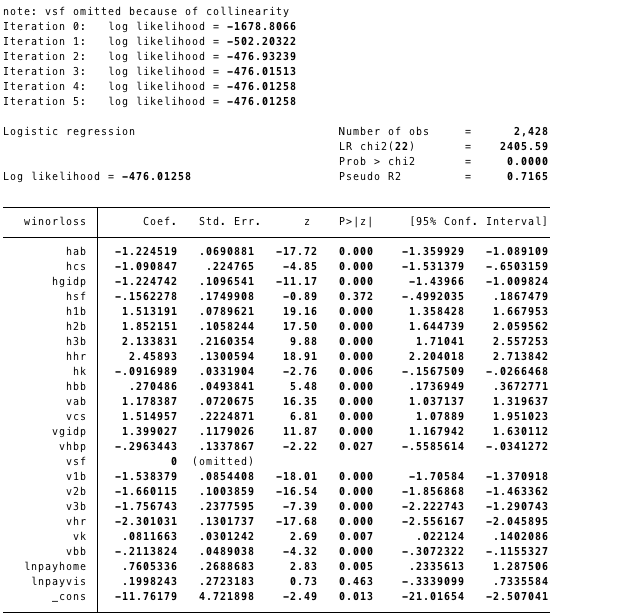
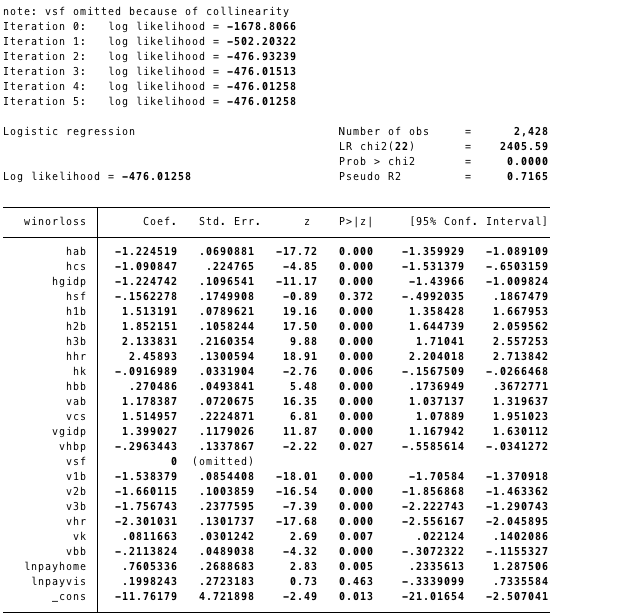
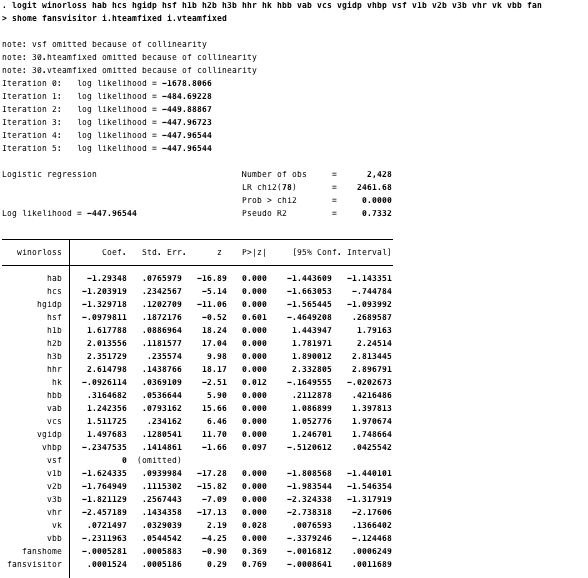
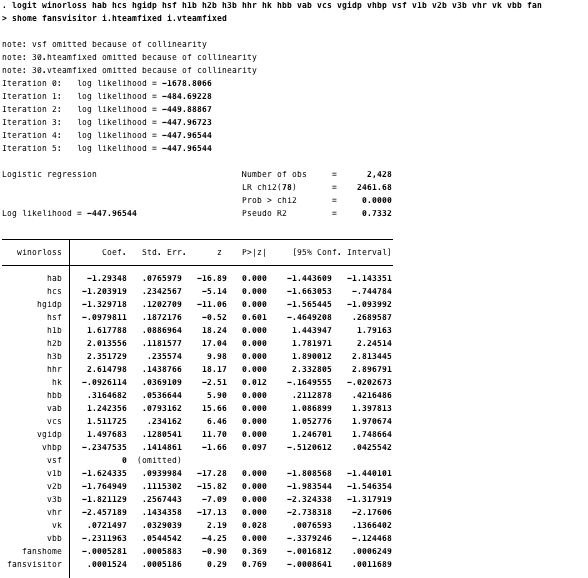


Figure 7: New model including payroll



We can see from the results that payroll of the home team and fan attendance of the visitor’s team has a significant influence on the result. However, we realized that payroll and number of fans may not have a direct influence on winning or losing. Instead, a stronger team will certainly receive a higher payroll, attract more fans, and have a higher chance of winning the game. Thus, the influencing factor will be the team itself rather than the payroll and attendance. To examine this assumption, we assigned one binary variable to each team and regressed on the existing variables together with the binary variable for home team and visitor team. We numbered all 30 teams and treating the numbers as factor variables. This will exclude the effect of the team itself and isolate the influence of payroll and attendance. The new regression results are as follows:

Figure 8: Model that isolates the influence of attendance

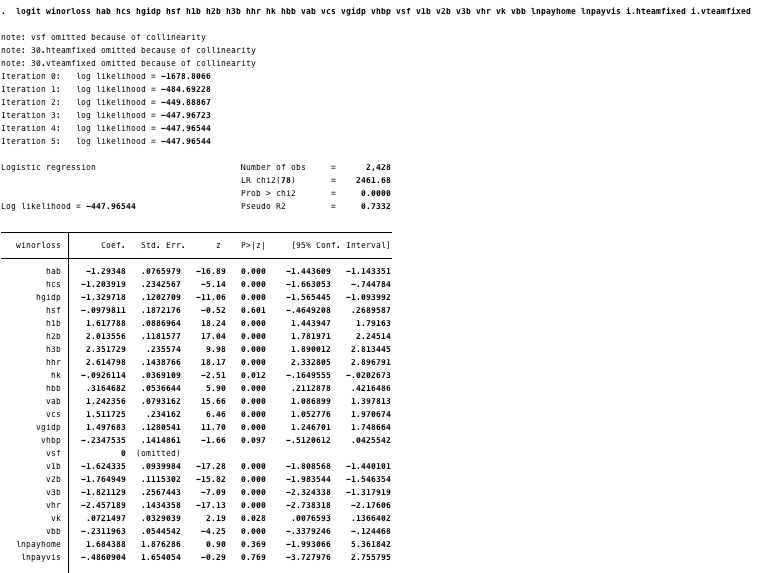
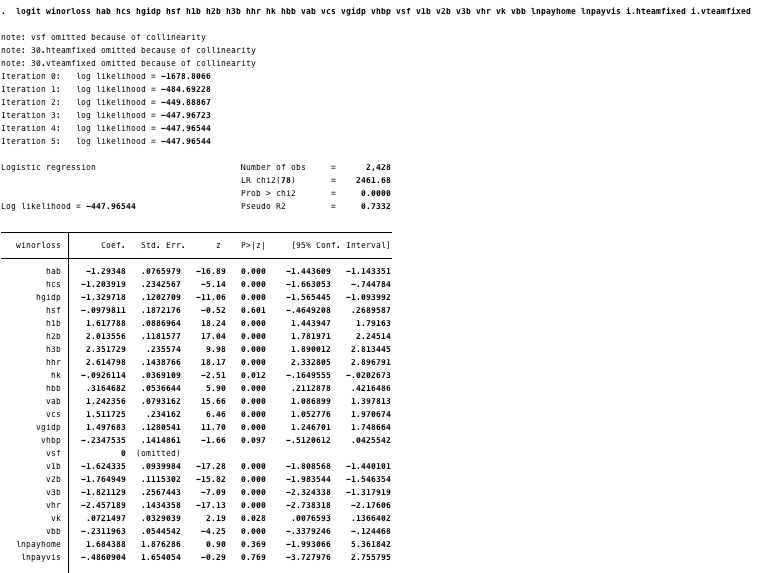


Figure 9: Model that isolates the influence of payroll



After creating the new models, we can see that none of the new indicators are different from 0. Therefore, there is no need to test the validity of the model since it is the exact same model from the previous section.

1. Impact of a ‘Slugging’ Strategy

The final section of the report will attempt to model the impact of a specific strategy on the outcome of a game. Within baseball, it is possible to classify offensives into two types. First, there are offensives that are focused on getting extra base hits (i.e 2Bs, 3Bs, and HRs). These teams are typically slower when running on the base paths and strike out more because of their tendency to hit for power. Second, there are offensives that are focused on getting on-base through singles and walks. Once on the base paths, these teams rely on speed and sacrifice hitting to eventually move the runner home. We will call the teams oriented around power ‘Slugging Offensives’ while teams oriented around threading together singles ‘Small-Ball Offensives.’

7.1 Classifying the Teams

The method of classifying a baseball team strategy begins with calculating their “Hit Ratios.” Hit Ratios is a new formula we created specifically designed to classify teams based on the type of hits they produce:

The following image displays the Hit Ratio of every team in the 2016 season ordered from largest to smallest. The number of hits and total bases were calculated by aggregating the teams data within SQL. Teams with higher hit ratios means the composition of their hits are mainly extra base hits while lower hit ratios indicate teams who hit more singles. We also decided to include both home and away games in the calculation to display the true nature of the team’s offence. We included every game within the calculation because teams may get more extra base hits in opponents ball parks if their home part has unfavourable hitting conditions.

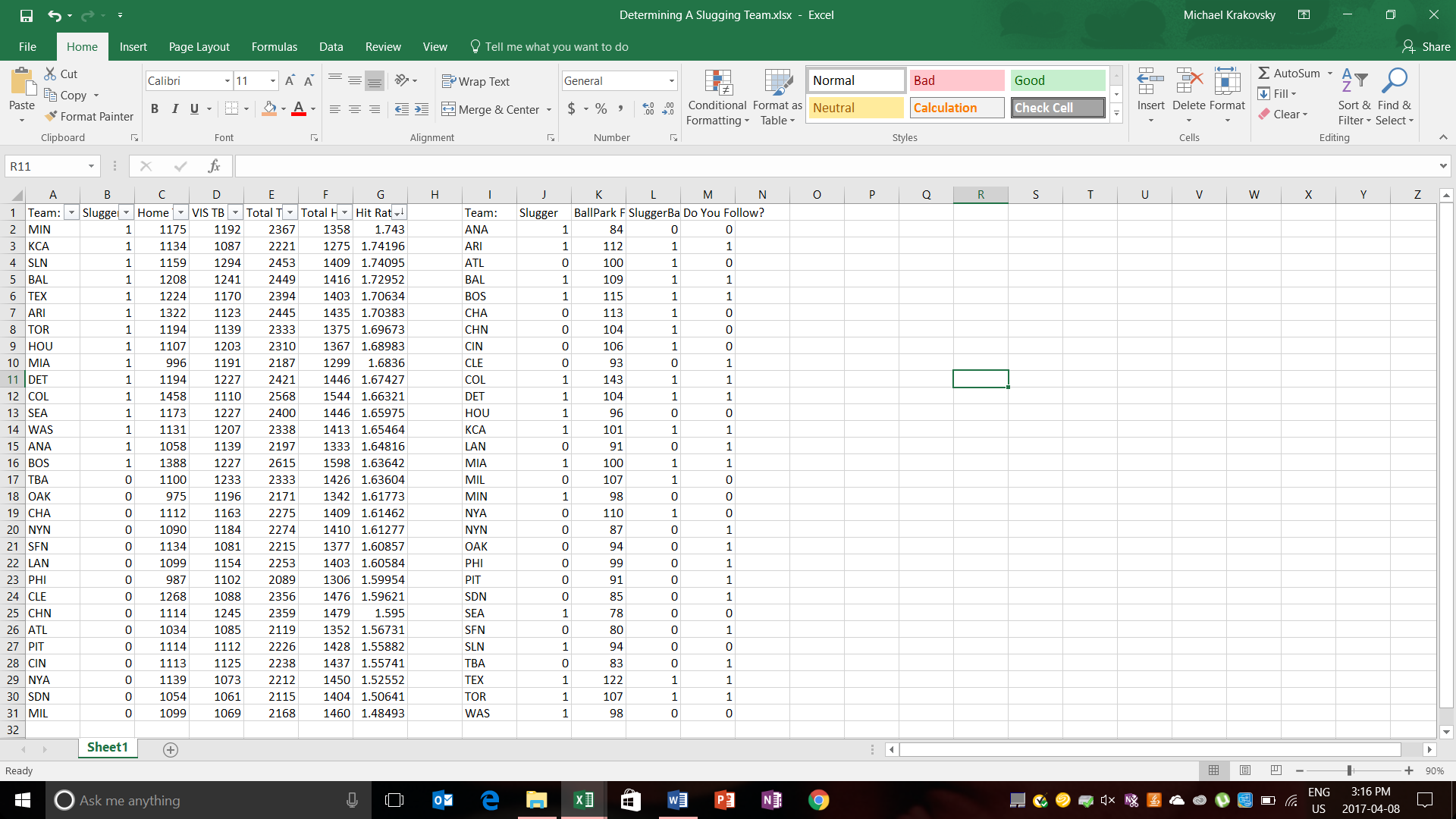


Figure 10: Calculating Hit Ratios

After calculating the hit ratio of every team, we assigned binary variables to classify the strategy the team utilizes. A ‘1’ indicates the presence of ‘Slugging Team’ while a ‘0’ indicates the presence of ‘Small-ball Team.’ We assumed that the 15 teams with the highest hit ratios use the power approach while the 15 teams with the lowest hit ratios use the small ball approach. After classifying the teams, we added the variable into our model; however, the variable’s coefficient was insignificant from 0 at the 5% confidence interval. Therefore, using a slugging strategy alone provided no advantage when it comes to winning games.

* 1. Strategy Dependent on Ballparks

Baseball is extremely unique in the sense that every ballpark provides a different experience because the dimensions of every ballpark varies. For example, Rogers Centre is a smaller field compared to AT&T Park in San Francisco. Therefore, to measure how favourable the ballpark is towards hitters we use a known statistic called a Ballpark Factor:

A ballpark with a higher factor indicates favourable conditions towards hitting while lower factors indicate unfavourable conditions to hitters. Like the previous section, we rated the Ballparks with binary variables. A ‘1’ indicates a field favourable to sluggers while ‘0’ indicates the opposite. Finally, we determined whether a team was utilizing the proper strategy within their ballpark. If the team had slugging characteristics and was in a hitter friendly ballpark, then they were assigned a ‘1’ indicating the use of a proper strategy. If the team had small ball tendencies but was in unfavourable conditions for hitting, they are utilizing the proper strategy thus receiving a ‘1.’ Teams that did not have a style of play relating to their ballpark where given a ‘0’ indicating poor strategy.

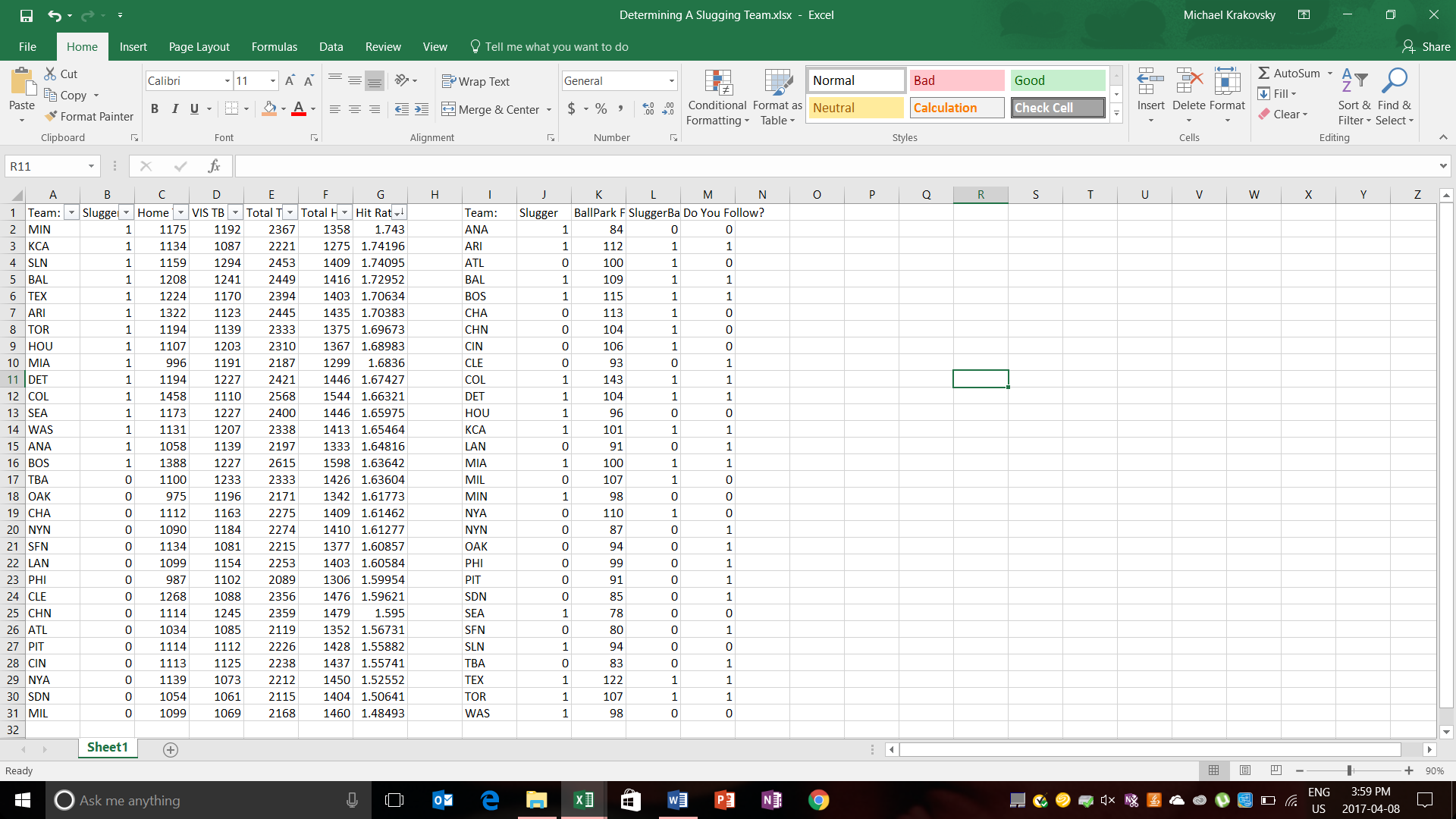


Figure 11: Deciding which teams are choosing the right strategy

Their might be speculation about whether Ballparks Factors are correlated to Hit Ratios since higher factors should lead to more home runs. However, no relationship exists between the two variables; therefore, we can proceed with modelling the strategies.

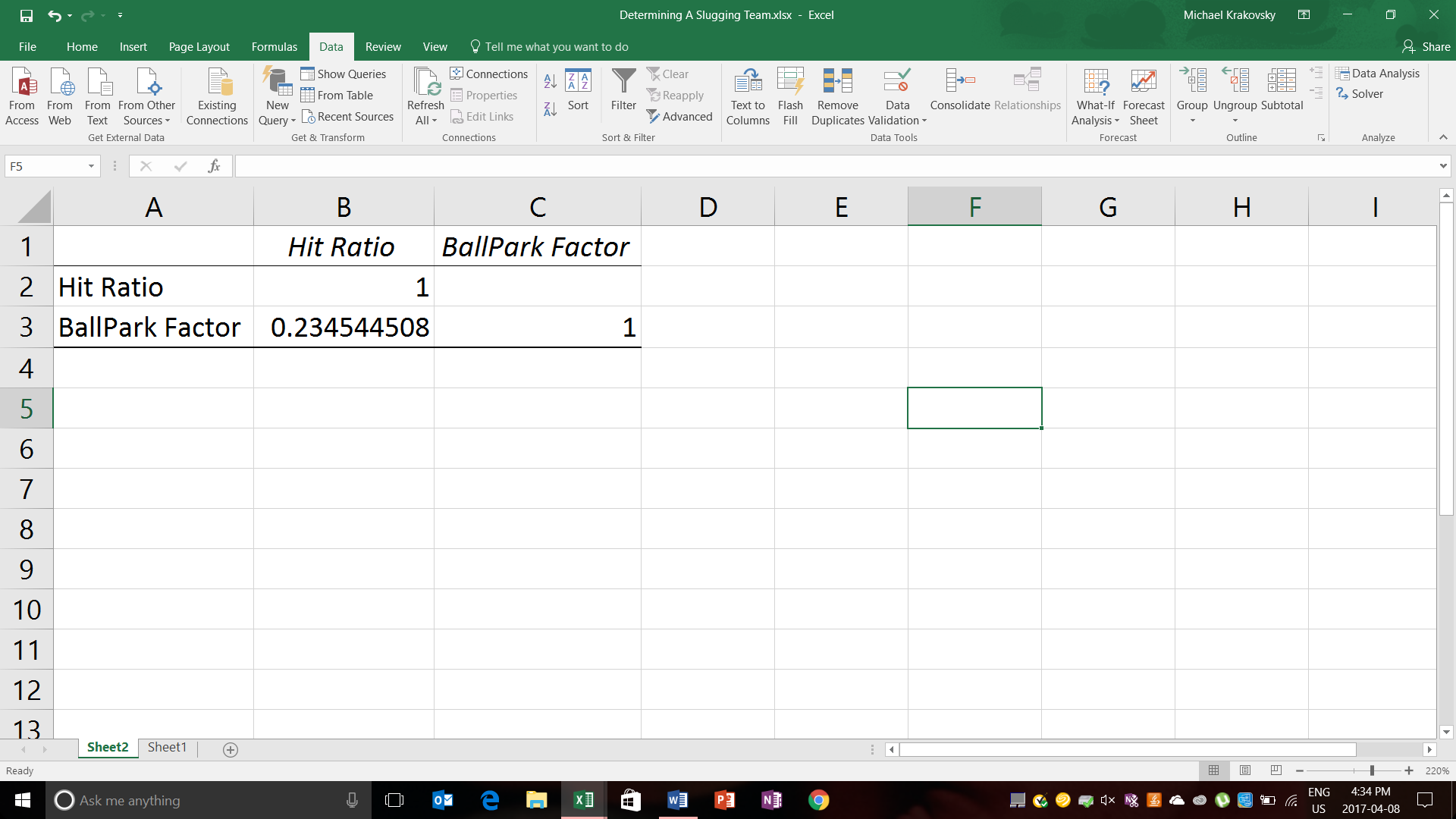


Figure 12: Correlation Matrix with Hit Ratio and Ballpark Factor

* 1. The Final Model

After incorporating the previously discussed variable into our model, we see that it is significant from 0 at a 5% level of significance. Therefore, choosing the right strategy based on your ballpark will influence your ability to win games. Like the baseline model, we attempted to predict the final 400 games of the season by using a holdout period.

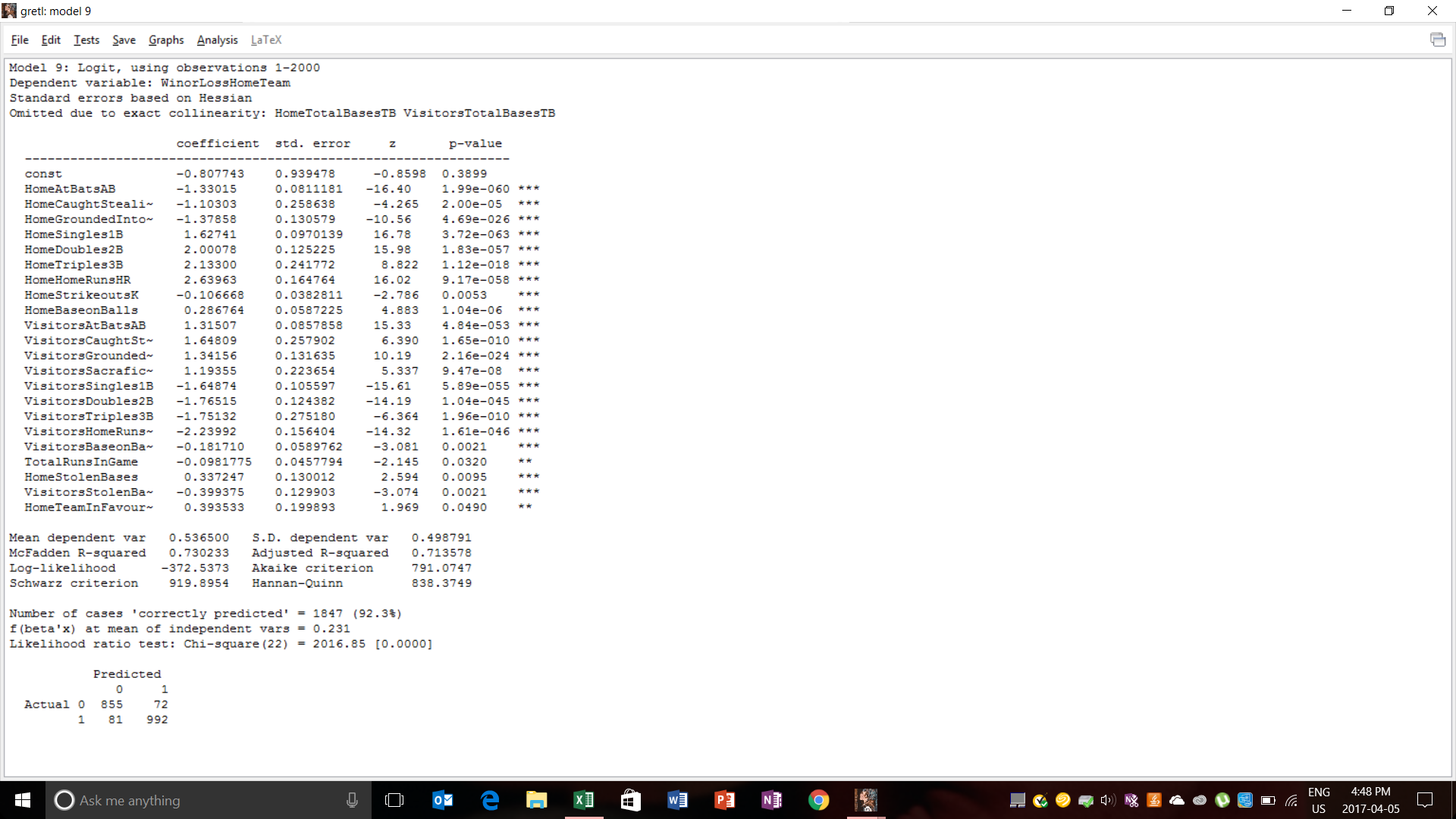


Figure 13: Model proving the importance of strategy

As you can see, the McFadden R-Squared and the Adjusted R-Squared increased compared to the previous model. The following image is the confusing matrix pertaining to the predictions of the final 400 games of the baseball season.

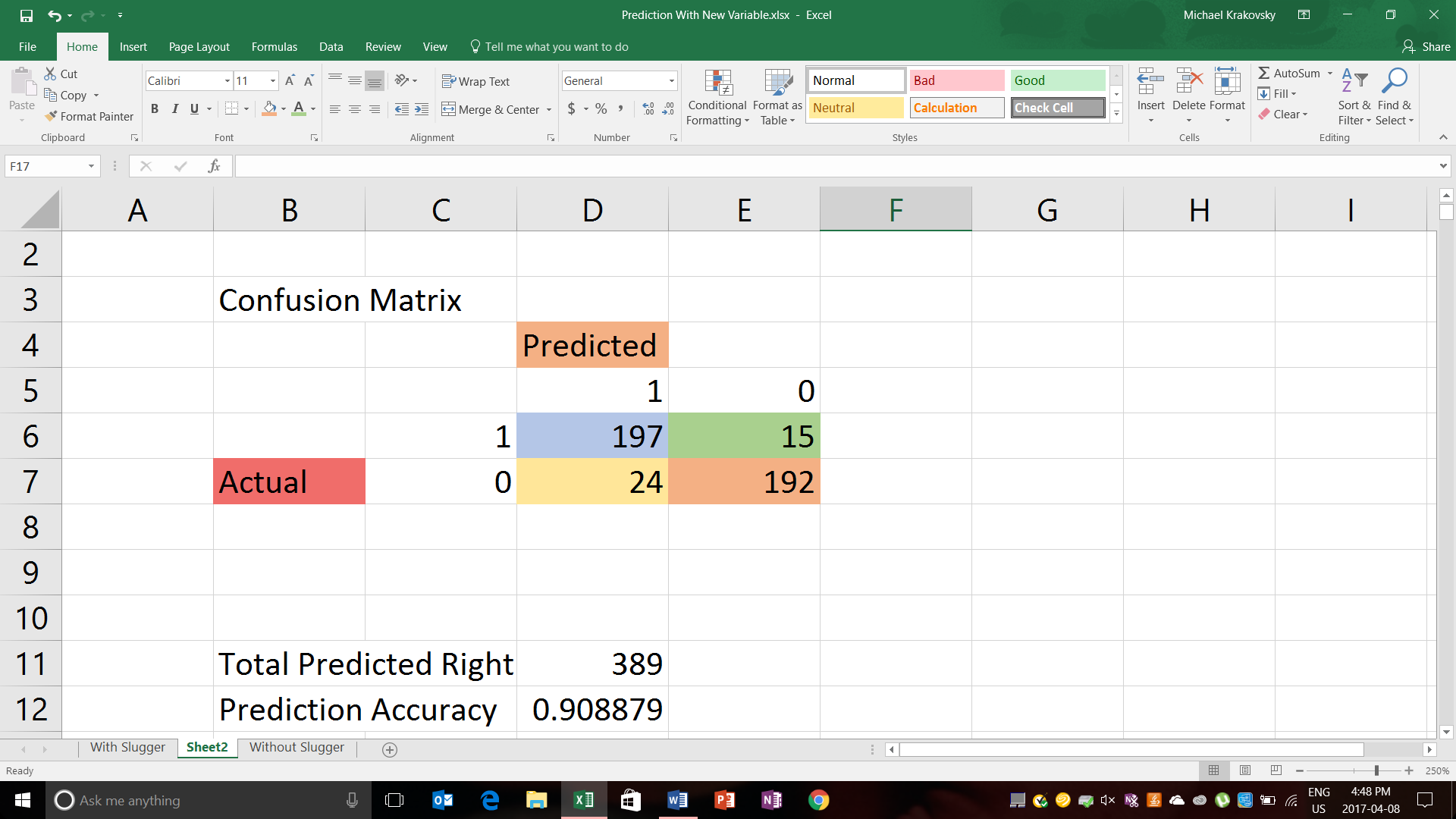


Figure 14: Confusion matrix of new model

Unfortunately, the new model predicted two less games correctly; nonetheless, the newer model explains more variance within the model. In addition, the newer model performs better regarding RMSE. Therefore, we can conclude that the newer model is slightly more reliable.

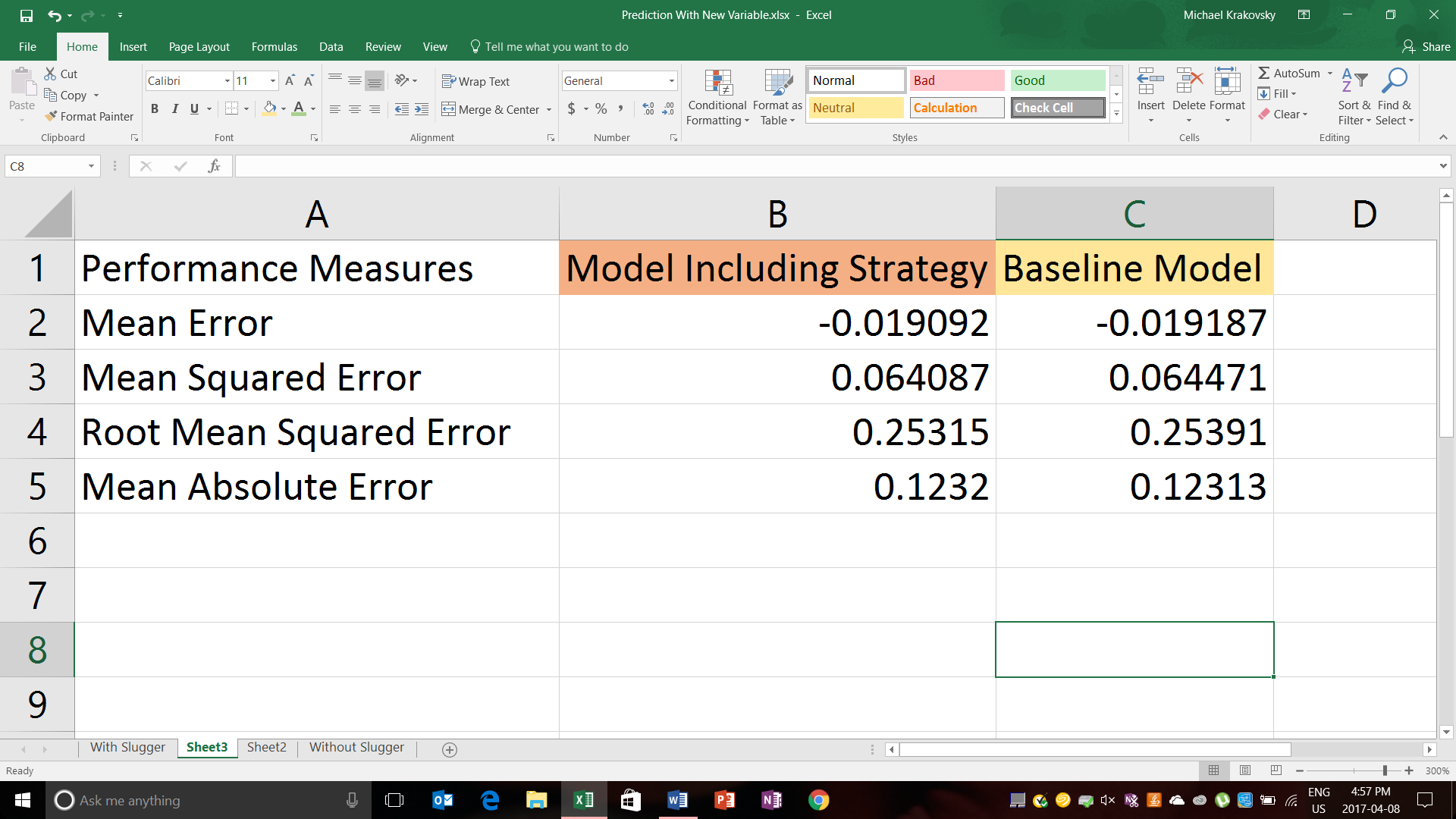


Figure 15: Comparison between models

1. Utilizing the Information

The information collected above can be used to bet against odd makers; however, much more work needs to be done. The purpose of the report is to explain both obvious and hidden indicators to winning a game. Finding the indicators is only half the work, the other half of the work is to predict how the teams will perform in the next game. After being heavily involved in the sport for the past 5 years, we intuitively noticed that the game of baseball is extremely cyclical in nature. Since there are 162 games every season, teams are bound to go on multiple hot streaks and cold streaks. If you can create a model that accurately predicts the variables indicated in the baseline model while also identifying teams that are streaking, an accurate model can be created to predict the winning team. With the complete information, you can identify inaccurate odds within the marketplace thus betting against these odds. Otherwise, we will act as consultants to management stressing the importance of tailoring your team towards the dimensions of your ballpark.

1. Conclusion

Thank you for taking the time to read our report on baseball analytics. Baseball teams should build their teams according to the dimensions of their ballparks and should place higher value on players who match the strategy dictated by the ballpark. The is a beautiful game so we hope you will take the time to watch a couple games in the 2017 season. May the Toronto Blue Jays be successful this year!

1. References

Glossary. (2016, January). Retrieved April 06, 2017, from http://m.mlb.com/glossary

Retrosheet Event Files. (n.d.). Retrieved April 09, 2017, from http://www.retrosheet.org/game.htm

1. Definitions were rephrased from (Glossary, 2016) in the references section. [↑](#footnote-ref-1)