**Final Analysis: Predicting MLB Home Runs**

Kyle Hirsch and Ian Suleski

University at Albany

CINF 624: Predictive Modeling

Professor Cusano

December 6, 2021

**Abstract**

The goal of this project was to create and test a predictive model that would be able to accurately predict home run output for a given MLB player in a future season. The model was created by and tested with two different sets of player data, both of which were identical in structure but included different players and their respective variable statistics. Both sets of data were tested against various validation methods to ensure accuracy and reliability, including tests against multicollinearity, partial least squares, autocorrelation, heteroscedasticity, and more. Each of these validation methods produced nearly identical results for both sets of data. Although the model may certainly be improved upon in future iterations, the current model has been successfully proven to be accurate and reliable in predicting future home run outputs for an MLB player, given that player has participated in each of the previous four seasons with at least 200 plate appearances in each season.

**Final Analysis: Predicting MLB Home Runs**

Baseball has always been an analytically driven sport with public statistics for virtually every action a player can take, and players being judged mainly on averages and offensive output. With a total of 162 games per regular season for each team, player consistency is key, and player value is determined mainly on consistency. Consistency at high levels in certain statistics hold more weight, such as statistics that measure offensive output. Players with higher measures of offensive output generate more runs for their team, which means more wins. As a tool for identifying player value, this study will attempt to construct a model that is able to predict a players’ home run output for an upcoming season. Home runs have been a staple of baseball as the most emphatic moment in the sport going back to the times of Babe Ruth. Home runs are defined in baseball as when a batter hits a fair ball and is able to score without being put out. In almost every occurrence, a batter hits a home run by hitting the ball over the outfield fence, awarding them all four bases and the score (“Home Run,” 2021). Home runs are now more present in baseball today than ever before. In 2019, 6,776 home runs were hit across all teams which broke the record for the number of home runs in an MLB season, and the previous record was only set in 2017 (“Major League Baseball finishes,” 2019). Home run totals do not paint a complete picture of offensive output for a player, but they are certainly a major factor. They are a very exciting part of the sport for spectators, can often win games in clutch moments, and also strongly correlate with total offensive output which is why they are the main focus for this study.

The study *Physicochemical Investigation into Major League Baseballs in the Era of Unprecedented Rise in Home Runs* conducted by Beals et al. presents a deep chemical and physical analysis of the baseball itself to understand why home runs have been more frequent. The motive behind this study is speculation that the baseballs have been “juiced” to make the ball travel farther in the air. In support of the speculation, CT scans were used for physical examinations of baseballs that concluded there was a severe decrease (56.7%) in the density of the baseball core between the 2014 and 2017 seasons (Beals et al., 2019). This notable difference in the ball density is in line with the concept of the balls being “juiced”, however the reason why is uncertain. Nevertheless, Beals et al. (2019) points out the effects of this with the number of home runs increasing a staggering 46.5% between the 2014 and 2017 season. This study does not claim the change in ball density as the sole reason for the surge of home runs, but it establishes the significance they have in today’s game.

Analyzing player performance data in the MLB has never been as advanced as it is today. Beneventano et al. explain the origins of this in their article *Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics*. Sabermetrics is a type of analysis in baseball created by Bill James in the 1970s that uses a detailed variety of player performance data to make informed decisions about particular players (Beneventano et al., 2012). The performance data represents a quantitative measurement to gather information on a player rather than qualitative methods like judging character or style of play which are subjective. Beneventano et al. made a stepwise regression model using 12 different statistics (including number of home runs) to find the best predictor of total runs scored by a team. This model had an R-squared value of 0.95 meaning 95% of the variation in the number of runs for a team was explained by the variation in the independent variables (Beneventano et al., 2012). The model shows an example of how performance statistics can be used to predict offensive output which will be important to consider in making a model to predict the number of home runs.

In an article titled *Baseball and Machine Learning: A Data Science Approach to 2021 Hitting Projections*, John Pette shows how he used statistical learning to predict certain baseball hitting statistics for the 2021 MLB season. Pette acquired his data from FanGraphs, which is also the website to be used in this study to acquire data. Pette’s model predicted ten different hitting statistics for the 2021 season, including runs, home runs, and plate appearances. Pette had a similar approach regarding data congregation, where he used two sequential seasons of data to predict data in the following season. Using multiple and sequential seasons of data helps establish a trend showing recent player performance and has a large impact on the final prediction results. Through Pette’s predictive model, he was able to accurately predict hitting statistics for each player as well as discover which predictive variables had a higher level of influence on the prediction (Pette, 2021).

The task of the model to be created in this project will be to predict home runs per plate appearance in an upcoming season for any MLB player given their offensive stats from the 3 sequential seasons immediately preceding. It is essential to predict home runs as a percentage of total plate appearances because total plate appearances can vary drastically from season to season. Plate appearances can be affected by player injury, substitutions, or outside forces such as COVID-19 in the 2020 season. Predicting a percentage will allow the model to highlight player consistency and have the ability to be applied to a wider range of data. The completed model could potentially be used by professional organizations to identify players with high offensive value.

**Methods**

**Data Overview and Collection**

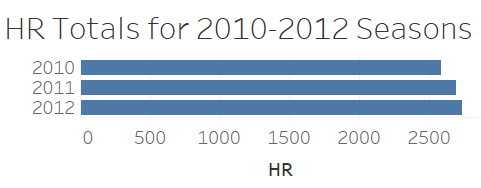
The data in this study covers MLB player hitting statistics. More specifically, the statistics included in this study are home runs (HR), runs batted in (RBI), walk percentage (BB%), strike-out percentage (K%), batting average (AVG), on-base percentage (OBP), slugging percentage (SLG), and plate appearances (PA). Home runs (HR) are the dependent variable this study is trying to predict. The subsequent variables are independent variables that may have some influence on HR. Runs batted in (RBI) are defined as when a player hits the ball or gets on base and allows a base runner(s) to score. Walk percentage (BB%) is the percentage of at bats that a player will receive a walk. Walks are awarded to a player when the pitcher throws 4 balls or the opposing team intentionally walks the batter. Similarly the next variable, strike-out percentage (K%) is the percentage of at bats that a player will strike when a pitcher throws three strikes against a batter. The K% does not include when a player gets out by a flyball or a ground ball. The batting average (AVG) is one of the more important statistics in baseball when evaluating a player’s offensive performance. The AVG is determined by the number of hits a player gets per at-bat. For example, if a player gets 2 hits out of 4 at bats in a game their AVG is 0.500. The AVG is then calculated from every game to determine a player’s season production (MLB, 2021).

Next, on-base percentage (OBP) is the number of times the player gets on base via a hit, walk, or hit by pitch per PA. The OBP is similar to the AVG except it also includes walks which is a better representation of how often a batter gets on base to potentially score. Slugging percentage (SLG) is the next variable in this study that represents the number of bases a player records per at-bat. This means that SLG accounts for extra base hits like doubles, triples, and home runs where the batter advances multiple bases. The calculation of the AVG uses the same amount of hits as SLG, but SLG also measures the output of those hits in terms of bases recorded and the AVG views all hits the same. Lastly, PA is every time a batter gets a turn at the plate (MLB, 2021). PA shows the number of opportunities at the plate a player has had which gives them more chances to hit home runs. These specific statistics were determined based on which variables are expected to have an impact on total home runs hit per season and should prove to show positive or negative correlations.

Since baseball has been established as a major sport in the late 1800s, over one hundred official seasons have been played, which means over one hundred seasons worth of data have been collected (“Major League Baseball Schedule,” 2021). For the purpose of simplifying the data set, this study will only focus on a four season interval: 2010-2013. This interval should provide sufficient data to work with and justify the findings. This study aims to use four seasons worth of hitting data from a specific player (2010-2012) to predict that player’s total home runs in the future season.

In order to avoid blatant outliers and inaccurate data, players included in this study must have had at least 200 PA in each season they participated in. 200 PA is an estimated value that requires about 65 games (out of 162) in a season for a player to complete, assuming they get three PAs per game, which is reasonable for a main-lineup starting player. This allows the model enough data to work with while avoiding anomalies and outliers. Also, each player must have participated in each of the four seasons during the four season interval. This allows our model to work with complete data and all players who don’t meet this criteria have been removed.

**Figure 1**



*Note*. Figure 1 shows the HR totals of all the players used in this study for the 2010-2012 seasons. It can be seen that the HRs increase each season starting in 2010 with 2,587, then 2011 with 2,697, and lastly 2012 with 2,739 HRs. This is a representation of all the HRs by MLB players that are going to be used in this study.

# **Data Generation Process**

## The data used in this study represents offensive production by players in the MLB between the 2010 and 2013 seasons. Players produced the data via official PAs, in which they recorded the offensive hitting outcomes/variables included in this study. All hitting statistics were recorded and collected by the MLB. This data has been structured and presented on FanGraphs, a website that specializes in providing baseball data and statistics, which is where the data for this study has been acquired (FanGraphs, 2021).

**Figure 2**

## 

*Note.* Figure 2 illustrates the AVG of players by how many PAs they had in the 2010 season. This is a helpful visualization for the reasons that this study requires a minimum of 200 PAs in a season. The scatter plot on the left represents the AVG of players who had less than 200 PAs. The lower the amount of PAs is shown to increase the distribution of the AVG between the players. Meaning that with the smaller sample size of PAs, there is so much variation in batting AVG that it is not accurate to assess a player’s performance for a whole season with. For example, both the highest (0.364) and lowest (0.036) AVGs are from players below 200 PAs. The scatter plot to the right represents the AVG of players who had more than 200 PAs which shows less outliers and a better distribution to try and make accurate predictions with.

# **Biases**

An initial bias in the data of this study was from MLB players being included who had very little plate appearances for the season. These players' stats could be disproportionate to what they would have produced over more PAs and affect the accuracy of the model. To solve this, this study includes players who have had a minimum of 200 PAs in a given season to control for the outliers with very few PAs. A normal MLB season includes 162 games, which means that 200 PAs equates to about 1.2 PAs per game not including postseason. 1.2 PAs per game is a very reasonable requirement for a large number of main-lineup players.

Another potential bias is that the data only includes players in the MLB rather than players from different leagues across the world. This is because the MLB represents the highest level of baseball that every professional player strives to make it to. Therefore, this model may not be accurate when applied outside of the MLB, such as in the Minor Leagues or Nippon Professional Baseball (NPB) in Japan. Leagues other than the MLB may include different factors that could influence results and some variables may hold a different weight. For example in the NPB, the strike zone is smaller allowing less space to throw a pitch as a strike and they also use a smaller ball than the MLB (“Baseball in Japan,” 2021). Furthermore, NPB games can end in a tie which is not allowed in the MLB. Games in the MLB that are tied will go to extra innings where the batters get extra PAs to try and score more runs.

Figure 3 shows a comparison between the MLB and the Japan Central League in Nippon Professional Baseball. The individual charts are portraying the percentage of total HRs by each PA bracket. For example, in the MLB in 2010, roughly 20% of total HRs were hit by players who had within 630-660 PAs (see Appendix A). Using three seasons for each league allows us to see trends throughout seasons and better distinguish the MLB from the Japan Central League. The MLB data shows that higher PA brackets consistently hold a majority of HRs hit and there are clear upwards trends. Although the Japan Central League data also shows that higher PA brackets hold the majority of HRs hit, the charts show weaker trends with less consistent data spikes. This comparison shows that MLB data is more predictable with more clear trends. At the very least, both leagues present differing trends, which means that a model fit to MLB data may not translate well to Japan Central League data.

Another initial bias of this study involves outlier seasons of player performance. For example, a player may perform poorly in a given season which could predict poor results for the subsequent season. In a model that only uses statistics from one season to predict another, this may not be sufficient enough data to accurately predict their upcoming season, especially if this poorly performed season is a major outlier compared to previous seasons. In order to combat this bias, this study will utilize statistics from four consecutive seasons to predict the outcome of a subsequent season. This strategy helps find a four-season average for the player’s statistics which should allow higher accuracy for the model by placing less weight on outlier seasons. Using a longer season stretch also allows the model to identify trends for each player, which should have a large impact on prediction as well.

Figure 4 shows an example of a player, Adam Dunn, who had an outlier of HR production in the 2011 season (11 HRs) relative to the 2010 season (38 HRs) before and 2012 season (41 HRs) after (see Appendix A). In this case, the model will use the average years to make a prediction about the fourth year so the outlier season does not have as major of an impact on the final HR prediction.

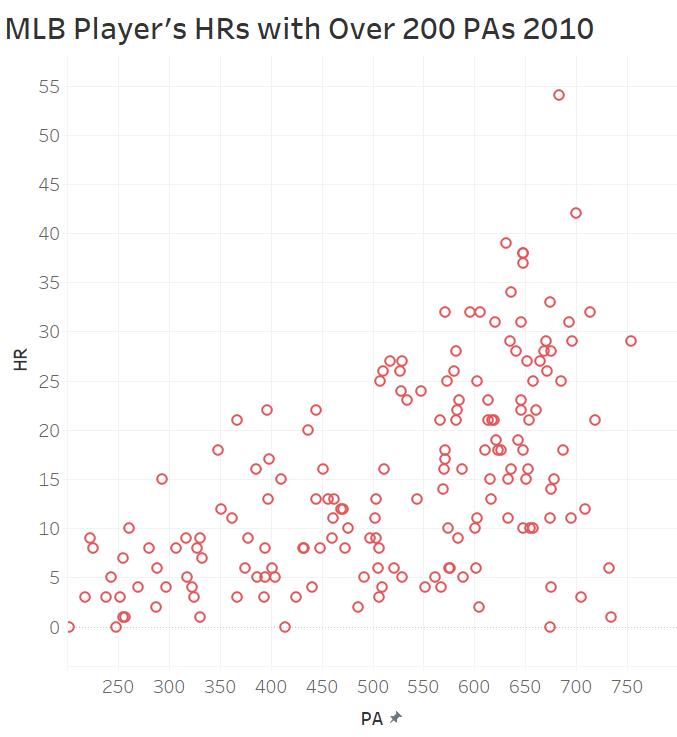
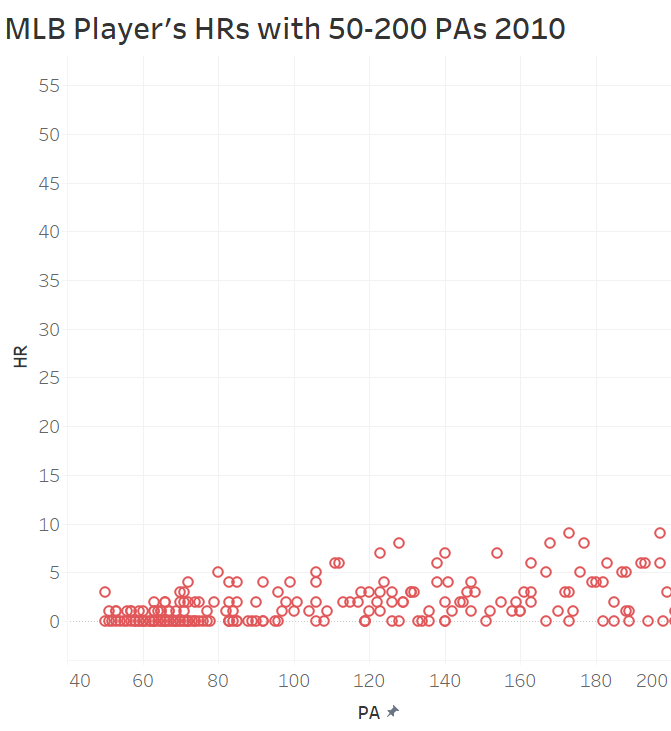
**Exploring the Data**

Through the process of descriptive analysis regarding the presented model idea, many important questions and concerns were answered. The questions that were explored in the descriptive analysis are as follows: Is four consecutive seasons the most efficient amount of seasons in order to create an accurate model? Would more or less consecutive seasons be more efficient? Is it even necessary to use more than one season? Is two hundred minimum PAs the most efficient number? Will more or less plate appearances prove to be more efficient or more accurate? Are there any prevalent correlations, positive or negative, between certain hitting variables and home run totals? Which variables have no correlations and will not help with prediction?

Figure 5 helps illustrate the importance of using multiple seasons worth of data to construct the model. These line graphs use Robinson Cano as an example, as he qualifies for the dataset to be used in the study through multiple consecutive seasons (see Appendix A). Clearly, using only one season worth of data does not tell a story or show any trends. Predicting data would be virtually impossible for any player if only one season is used as a predictor. Using four seasons of data is more ideal as it starts to show a trend of how Robinson Cano performed in multiple seasons. An average can be derived from these values and a prediction can be attempted. Four seasons of data was determined to be the sweet spot, as requiring more consecutive seasons of data for each player cuts down the data set exponentially and gives the model less data to work with while not giving much more information.

Figure 2 shown previously helps illustrate why a minimum of 200 PAs is required for players to qualify for the data set. Players with less PAs produce outlier data more frequently that can potentially throw off the accuracy of the created model. Requiring 200 PAs allows the model to work with players who have established consistency and will produce more reliable statistics. The exact number of 200 PAs is a ballpark number, as outlier data seems to fall off slightly earlier than 200 PAs but this number gives the model enough data to work with.

**Figure 6**



*Note*. Figure 6 clearly shows that, naturally, more PAs generally lead to more HRs.

The general conclusion on variable correlation was that most variables have no obvious correlations to HR count, other than PAs. This realization led to the decision to have the model predict HR percentage based on PAs, rather than just a total HR count prediction which was the previous model plan. PA count can vary wildly between seasons which can lead to inaccurate predictions. Predicting HR percentage of PAs will allow model application to more scenarios.

**Initial Model Build**

The data sets used in the initial build include the training set and the test set. The training set and test set both consisted of offensive player data for all MLB players who had a minimum of 200 PAs in each of the four MLB seasons between 2010 and 2013. In total, this includes 173 records, with about 80% of the records being used in the training set and 20% of the records being used in the test set. The variables that were initially used include HR, RBI, BB, K, AVG, OBP, and SLG. To reduce the variance in statistics caused by players who had more PAs than others, each variable used in the model is calculated as a percentage of that players’ PAs. Furthermore, instead of using 28 different variables for each player (all 7 variables mentioned previously for each of the four years in the time period), it seemed more practical to use the variable averages across the four year period. Using the averages for each variable takes away the “trend” aspect of the predictive model, where it was previously assumed that the model would be able to utilize previous trend information to predict the future HR value. This may cause the accuracy of the model to take a dip, but there are possible solutions should this be the case.

**Figure 7**

*R Code for Model Creation*

****





*Note*. Figure 7 displays the code used for model creation. Figure 8 displays the R packages used in this study (see Appendix B).

First, the CSV file of the data is imported. Then the training set is defined, using the first 138 rows (80% of total rows) and all 7 variables. The test set is also defined, using rows 139 to 173 (20% of total rows). Then, a multivariate linear regression model is created using HRs as the response variable, all other variables as the predictors, and the training set as the data. The multivariate linear regression is the preferred machine learning algorithm to use for this study because it produces the needed quantitative output of the number of HR hits in a season. Opposed to a classification or logistic regression algorithm which outputs categorical qualitative data. The multivariate linear model equation shown in Figure 7 was reproduced.

HR = B1 + B2RBI + B3BB + B4K + B5AVG + B6OBP + B7 SLG

In the model equation, B1 represents the predicted y-intercept of the regression line. The subsequent B2-7 in front of each independent variable represents the predicted slopes. These slopes signify that when an increase of the independent variable occurs the dependent variable (HR) will increase/decrease depending if the slope is up or down. Now that the necessary algorithm is selected and the HR model has been built, steps for evaluating and validating the performance of the model can be taken.

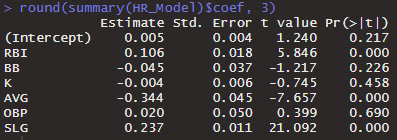
**Validating the Initial Model**

This section will evaluate and validate the HR Model created for this study based on various assumption checks and tests to judge its accuracy and appropriateness for use. There are assumptions for linear regressions which need to be tested for and corrected as necessary. First, the overall summary of the HR Model will be assessed with significance in the independent variables (P-value), the R-Squared value, and the residual or error terms of the model. Second, outliers in the data will be accounted for as they can have an influence on the model. Third, the assumption that there is no multicollinearity or influence amongst the independent variables will be measured and corrected as needed. Within this step the partial least squares method will also be used to determine the best number of variables to use. The fourth step is to check the assumption that heteroscedasticity (high variation) does not exist in the residual values of the model. Fifth, the assumption that autocorrelation does not exist in the model’s data will be tested. If these assumptions are not met then the quality of the predicted values can be affected. Lastly, the correlation accuracy of the predicted and actual values are measured to find the accuracy percentage for the HR Model.

Figure 9 is the summary of the original HR Model regression (see Appendix B). The summary results show a wealth of information regarding how the variation of the dependent variable can be explained by the variation of the independent variables. This is shown by the Adjusted R-squared value in the summary which reads 0.967 for the HR model. In this case, it means 96.7% of the variation in HRs is explained by the independent or control variables. This is a very high R-squared value meaning the independent variables and the variation in HRs are closely aligned.

**Figure 10**

*Coefficients Table*

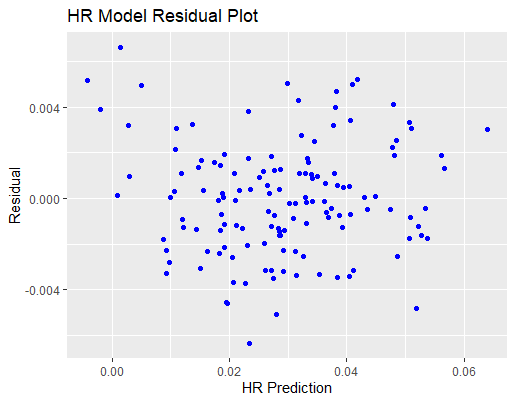


HR = 0.005 + 0.106 RBI + -0.04 5BB + -0.004 K + -0.344 AVG + 0.020 OBP + 0.237 SLG

*Note*. Figure 10 displays the Coefficients table from the summary in Figure 8 which has the statistics broken down for each of the independent variables. Also included is a recreated HR Model equation using the intercept and estimated slopes for each variable.

As noted before, the slopes signify that when an increase of the independent variable occurs HR will increase/decrease. For example, AVG has a negative slope so an increase in the AVG is estimated to have lower HR numbers. It is important to analyze the Sig. or P-Values. The P-Value means if you reject that the true slope of the variable is zero you will be wrong X% of the time. This is why interpreting a variable’s slope with a high P-Value is trivial due to the great probability that the true slope is zero. In other terms, the great probability that the independent variable didn’t affect the dependent variable at all. The HR Model summary shows three out of six of the independent variables proved to be statistically significant and interpretable at less than .05 level of significance. These variables include RBI, AVG, and SLG while BB, K, and OBP proved to have little significance and should not be interpreted.

**Figure 11**



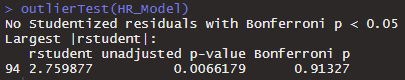
*Note.* Figure 11 shows the residual plot of the HR Model that has the residuals on the y-axis and the HR predictions represent the fitted values from the model on the x-axis.

The residuals or error terms are the difference between the predicted HR values from the model and the actual HR values in the dataset. The greater the residual means the predicted value is farther away from the actual. The data points in Figure 11 do not show much of a trend other than a large group of predictions with very low residual values near 0.000. Furthermore, there are more predictions with a negative residual than a positive one meaning the HR Model generally predicted less HRs than the actuals. The points that are more spread out in the top and bottom of Figure 11 could be from potential outliers influencing the residuals.

It is important to check for outliers in the regression because an outlier could have high leverage and influence on the regression coefficients. On the other hand, it is not correct to drop any point that exhibits these qualities as more insight could be gathered from interpreting why that specific point is an outlier.

**Figure 12**

*Outlier Test*



*Note.* Figure 12 is an outlier test to determine if there is any unwanted influence in the model.

The output of the outlier test signifies that there are no Studentized residuals with Bonferroni p < 0.05. The Bonferroni test determines if the largest Studentized residual is significantly different from the other data points (Wehde, 2020). In this case, there are no outliers that are worth looking into because there are no residuals that are significant. However, the largest Studentized residual is still reported at the bottom of Figure 12 which represents the most extreme outlier at row 94 who was Jonathan Herrera. Jonathan’s HR numbers are relatively lower than the other players, but this does not mean that it is an outlier as proven in Figure 12.

**Multicollinearity Test**

Multicollinearity refers to an intercorrelation between the independent variables in a study. This could cause a problem in regression models because the independent variables may affect each other rather than the dependent variable and by doing so possibly increasing the variance of the coefficient estimates. This would make interpreting the results of a regression with uncorrected multicollinearity inaccurate.

**Figure 13**

*Condition Index*

****

*Note.* Figure 13 displays the results of the Condition Index run on the HR Model to check for multicollinearity.

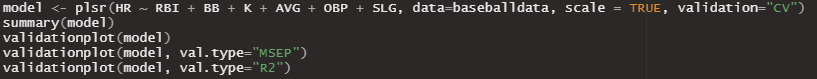
The Condition Index is a way to show the degree that multicollinearity exists in the regression model (Kim, 2019). This provides an efficient overall test not individualized to specific variables to first identify if this regression has interdependence or not. The important number here is 253.17 far to the right in Figure 13 which is a representation of the multicollinearity that exists in the whole regression. For reference, a 10-30 Condition Index says there is moderate to strong multicollinearity and 30+ is severe (Kim, 2019). Being that 253.17 is many times beyond what is considered severe it can be determined that the HR Model regression has a very severe problem with independent variables influencing each other. Further testing will have to be done to isolate the contributing variables.

**Partial Least Squares (PLS)**

Using the method of partial least squares allows the model to reveal the most efficient amount of variables that are needed to make an accurate prediction.

**Figure 14**

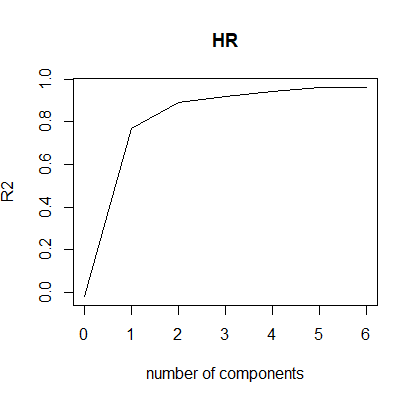
*Partial Least Squares Data Import*



*Note.* Figure 14 shows how the model is imported into the partial least squares method with model outputs that illustrate the accuracy of model predictions corresponding to the number of variables included.

**Figure 15**

*R Squared per Number of Components for Predicting Home Runs*



*Note.* Figure 15 shows how the number of components relate to the R Squared value. It is clear that the more variables or components included, the higher the R Squared is. But it is also clear that there are diminishing returns after 2 components.

The PLS method also allows the calculation of the RMSE value for each component shown in Figure 16, which reveals the accuracy of the model for each number of components included, shown by the ncomp value (see Appendix C). Clearly, as the number of components increases, the model becomes more accurate. It may seem beneficial to just include all variables in the model, but it was discovered through other methods (namely the Correlation Matrix) that some variables may be more harmful than beneficial. Due to this reason and the fact that there are already diminishing returns after 2 variables, it was determined that the sweet spot for the amount of components that should be included in the model is 4. While the PLS method reveals the most efficient number of components, it doesn’t describe which components to use. To answer this question, a Correlation Matrix was constructed.

**Figure 17**

*Correlation Matrix*



*Note*. From the results of the Condition Index, it is necessary to run a Correlation Matrix shown in Figure 17.

The output of the Correlation Matrix gives a value between zero and one for every independent variable compared with each other. The closer the number is to zero the less multicollinearity is present and vice versa. For the purposes of this study, the ranges of the Correlation Matrix coefficients are defined as 0.0 - 0.29, 0.3 - 0.59, and 0.6 - 1. The lowest range is when there is low multicollinearity, the middle range can be described as moderate multicollinearity where there is a potential impact and the upper bound is where multicollinearity is severe showing that the variables have a high correlation and can affect the results. It is to be noted that these ranges can be negative, and the same information from them regarding multicollinearity can be determined.

In assessing Figure 17, it is obvious that OBP and SLG both have high degrees of multicollinearity with the other independent variables. SLG is severely collinear with both OBP and RBI while OBP is severely collinear with BB and AVG. This intuitively makes sense as players with a higher SLG will get more extra base hits to have higher OBP and RBI. Then players with a high BB and AVG will generally get on base more to have a higher OBP. The high levels of multicollinearity exhibited in Figure 17 call for correction in the model to reduce to negative impacts as much as possible.

**Correction for Multicollinearity**

After consideration and many trials, the correction for severe multicollinearity in this model was made by removing OBP and K. This was determined by considering the Correlation Matrix results and going back to the coefficients in Figure 10 which showed OBP and K had insignificant P-values making their estimates not interpretable. The summary statistics for the new model (HR Model2) without OBP and K are included in Figure 18 and the model equation was updated.

HR = 0.004 + 0.106 RBI + -0.0325 BB + -0.319 AVG + 0.235 SLG

The summary shown in Figure 18 shows a similar Adjusted R-squared value to the original HR Model at 96.7% (see Appendix C). Unlike before, all the independent variables have significant P-values (< 0.05) meaning their results are worth interpreting in the new model. Further impact of removing OBP and K in regard to multicollinearity can be seen by running a Condition Index on HR Model 2.

**Figure 19**

*Condition Index without K and OBP*



*Note.* The Condition Index in Figure 19 was calculated with K and OBP removed as can be seen from the formula. The final number 59.72 far to the right is still over the 30+ threshold that signifies multicollinearity in a regression. However, the first HR Model was reported to have a Condition Index of 253.17 which shows the effectiveness of the correction to reduce the number by almost 200.

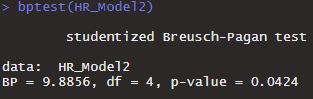
Even though HR Model2 is shown to have multicollinearity in Figure 19, the drastic reduction from the first model justifies removing OBP and K as the best remedy for this issue.

**Heteroscedasticity**

Heteroscedasticity arises when there is an unequal spread in the residuals of the predicted values. This will potentially make the prediction of the linear model’s true slope inaccurate and not worth interpreting. The Breusch-Pagan test was conducted to check for this.

**Figure 20**

*Breusch-Pagan Test*

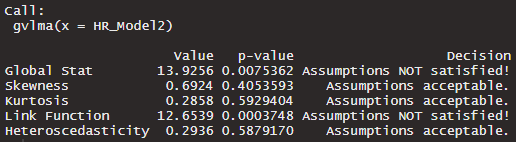


*Note*. Figure 20 is the Breusch-Pagan test where a p-value that is greater than 0.05 means that heteroscedasticity exists from the changing variance in the residuals, thus rejecting the null hypothesis that the variance of the residuals is unchanging (Zach, 2020).

Using the Breusch-Pagan test, the P-value reported for the new HR Model is 0.042 which is a positive sign. To look into this further, the Global Validation of Linear Models Assumptions (gvlma) package in Rstudio checks for Heteroscedasticity among 4 other assumptions to see if they are satisfied by the model.

**Figure 21**

*Global Validation of Linear Model Assumptions*



*Note*. This figure displays a gvlma output to check for Heteroscedasticity in the model.

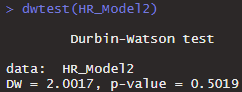
The gvlma output in Figure 21 confirms that heteroscedasticity in the HR Model2 is at an acceptable level for use in this study and no further action is necessary to correct it. The other assumptions in the gvlma output each represent a different test for the model. Skewness and Kurtosis are also acceptable in the model. The Skewness accounts for the distribution of the data and Kurtosis entails high and shallow peaks that can be an influence in the data which can be caused by outliers. The Global Stat and Link Function assumptions are not satisfied. The Global Stat is used when there are many covariates to measure how associated they are with the dependent variable and the Link Function determines if the underlying data is categorical or continuous (Goeman et al., 2021). Therefore, measures will have to be taken if the Global Stat and Link Function are to be accepted in the future.

**Autocorrelation**

Autocorrelation is when the error term of an observation relates to that of another observation. This study will use the Durbin-Watson test to check for this assumption. This value can range from 0-4, and the closer it is to 2, the lower the chance that autocorrelation is affecting the data.

**Figure 22**

*Durbin-Watson Test*



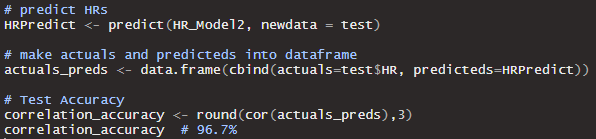
*Note.* This figure displays a Durbin-Watson test to check for autocorrelation in the model.

The results of the Durbin-Watson test in Figure 22 show a nearly perfect number at 2.00 meaning autocorrelation does not exist in this data and there is no need to correct for it. The accepted range for the Durbin-Watson number is between 1.5 and 2.5 (Glen, 2021). The data used for this model is the average of multiple seasons, and this makes it not conducive to autocorrelation. Time-series data generally has issues of positive autocorrelation because the variables over the years relate to each other. Overall, the Durbin-Watson test clears that this study is not being impacted by autocorrelation.

**Accuracy Test**

**Figure 23**

*Model Accuracy Test*

****

*Note.* Figure 23 presents the code to measure the accuracy between the predicted values and the actual values.

Shown in Figure 23, the HR predictions were first created using the HR Model2 and the test data which has been purposely separated from the HR Model from the beginning to ensure the predictions were being made using unseen data. Next, a data frame was built to combine the actuals and the predicted values to then be able to assess the correlation accuracy. The bottom of Figure 23 shows where the correlation accuracy was calculated using the actuals\_preds data frame built in the prior step. The accuracy result of 96.7% means the variation in the predicted values are extremely related in that they increase and decrease concurrently (Prabhakaran, 2021). This end result speaks to the close relationships present between HR and the independent variables of this study that allow the precision of the predicted values to be so accurate.

**Subsequent Model Build and Validation**

The subsequent model built for this study will use four different years of MLB player hitting statistics including different players to compare the results to the initial build and determine if the initial findings can be represented in a broader timeframe. The initial model was created using MLB player statistics from the 2010-2013 seasons, so the subsequent model is built using the 2014-2017 seasons. The variables included in the subsequent model are HR, RBI, BB, AVG, and SLG which mirrors the build of the initial model to compare the two timeframes. Note that K and OBP were excluded from the subsequent model to mitigate expected multicollinearity found in the initial model. The dataset used for the 2014-2017 seasons was arranged from FanGraphs with the same 200 PA minimum requirement to filter out players with little opportunity (FanGraphs, 2021). Only players who had recorded data for all four seasons were included and then the average of their seasons was calculated to produce one value. This analysis aims to analyze if the findings of the initial model also hold true in the subsequent model built with different seasons.

A multivariate linear regression model is also the chosen algorithm for the subsequent model because the study still requires a numerical output of the dependent variable HR which makes changing the algorithm to a classification problem unnecessary. Furthermore, using a regression will allow for better comparison with the initial model. The model includes HR as the dependent variable and RBI, BB, AVG, and SLG as the independent variables. First, the new data was imported in Figure 24 and then training and testing sets were defined (see Appendix D). The split between the training and testing set is the same as before with 80% of records in the training set and 20% of records in the testing set. However, the new dataset contains more players at 183 records so the sizes of the training and testing sets were adjusted proportionally. Then the subsequent model is created in Figure 24 as the Final HR Model using the new training data and variables from the 2014-2017 seasons (see Appendix D).

After the Final HR Model is made, the summary is run in Figure 25 to show the R-squared values, significance, and estimated slopes of the independent variables (see Appendix D). The R-squared value is 0.964 meaning 96.4% of the variation in HRs is explained by the variation in the independent variables of the subsequent model. All the independent variables also have statistical significance (*P* < .05) meaning they have influence on HRs and their estimated slopes are worth interpreting.

HR = 0.003 + 0.094 RBI + -0.025 BB + -0.334 AVG + 0.249 SLG

In the Final HR Model equation, the slopes for RBI and SLG are positive meaning when they increase, HRs also increase. Alternatively, the slopes for BB and AVG are negative so they make HRs decrease. These results of the R-squared, significance, and slopes are almost identical to the results found in the initial model build. Therefore, the summary here starts to meet the goal of this subsequent build in that it determined much of the initial findings are also consistent in another time frame of different data and players. Nevertheless, it is still necessary to further validate the Final HR Model to ensure the preliminary summary results are accurate.

The residual plot of the Final HR Model in Figure 26 was created to show the residuals and the HR predictions to see if there is anything to be seen graphically from the error terms (see Appendix D) As mentioned previously, the greater the residual means the predicted value is farther away from the actual. Opposed to the initial model, the subsequent model residuals show more of a spread throughout the graph which means the error terms are less accurate than before. This could potentially have an impact on Heteroscedasticity. However, there are still many points near or around 0 so the slight difference in variance between the initial and subsequent model should not heavily influence the model. The Final HR Model Residual plot still shows some potential outliers that are far away from the other predictions. In testing further it can be determined that outliers are in fact not influencing the Final HR Model with the Bonferroni test in Figure 27 confirming there are no Studentized Residuals with a *P* < .05 (see Appendix D). Outliers were not expected as the Final HR Model has already shown similar results to the initial model. Now that the summary of the model has been analyzed, it is necessary to test for the assumptions of multicollinearity, heteroscedasticity, and autocorrelation which can be influencing the results of the Final HR Model.

To reiterate, multicollinearity refers to an intercorrelation between the independent variables in a dataset that causes them to influence each other rather than the dependent variable. The Final HR Model is shown to have multicollinearity present through the condition index of the whole model in Figure 28 (see Appendix E). The condition index of this model has almost the same number (59.5) as the initial model. This begins to confirm the idea that the relationships between the variables found in the initial model of 2010-2013 data also hold true in the subsequent model of 2014-2017 data. These relationships between the independent variables are better illustrated in the correlation matrix of the Final HR Model in Figure 29 (see Appendix E). The correlation matrix does not appear to show the severe levels of multicollinearity found in the previous models, but this does not include OBP and K like the correlation matrix before. SLG and RBI are the two variables with the highest correlation at 0.83 which is in the severe category (no other variables are in that range). Multicollinearity remains present in the Final HR Model, but the approach of dropping OBP and K from the initial build proves to still be effective in limiting the effects of intercorrelation as much as possible.

Heteroscedasticity is an issue that was potentially shown in the Final HR Model Residual plot in Figure 26 with the error terms appearing to be unevenly spread out (see Appendix E). This can cause the overall predictions of the Final HR Model to be inaccurate from the inconsistent error terms. The Breusch-Pagan test was run to highlight potential heteroscedasticity present in the model. The test result on the Final HR Model in Figure 30 has a P-value of .57 which signifies that heteroscedasticity may be present in the model and influencing the predictions (see Appendix E). This is concerning because the desired result of the Breusch-Pagan test is a P-value less than .05 and the initial model did not have the same results. Testing the Final HR Model further for heteroscedasticity will be needed using the gvlma function.

Even though the results of the Breusch-Pagan test showed signs of heteroscedasticity at first, the output of the gvlma function on the Final HR Model in Figure 31 clears any issues as the heteroscedasticity assumption is acceptable (see Appendix E). Furthermore, the other assumptions in the gvlma output are all accepted with the exception of the Link function that is not satisfied. Most importantly, the Global Stat is accepted in this model which was not the case in the initial model. The Global Stat measures the relationship between the covariates and dependent variables in the model (Goeman et al., 2021). This is the first major difference between the initial and subsequent model that shows the new data may be better suited to predict HR numbers.

The final assumption of autocorrelation is not expected in the Final HR Model as the observations between players are unlikely to relate to each other in a cross-sectional dataset. This is confirmed by the result of 2 from the Dubrin-Watson test in Figure 32 which shows that essentially no autocorrelation exists in the data (see Appendix E). The same result was also observed in the initial model and this goes back to using the average of the seasons rather than using the data from each season separately. Each of the validation checks for the Final HR Model set up the final accuracy test to determine the end results from using the 2014-2017 seasons.

The accuracy test for the Final HR Model in Figure 33 is coded the same as the accuracy test for the initial model with the predictions first being made using the testing data and then a data frame is created to find the correlation accuracy between the predicted and actual HR values (see Appendix E). The accuracy result of 95.3% means the variation in the predicted values are extremely related in that they increase and decrease concurrently as found in the initial model (Prabhakaran, 2021). The close relationships between the variables in the Final HR Model result in the high accuracy percentage. The similarities between the two models of this study show that there were no relative differences in the data between the 2010-2013 and 2014-2017 seasons.

**Results and Discussion**

This project involved creating and testing a predictive model that would be able to accurately predict home run output for a given MLB player. Before model creation, certain restrictions and limitations were initially applied to the data set to ensure the model would work as intended. Multiple variables were chosen as predictors and it was deemed that a multivariate linear regression model would be the best option for model creation.

An initial run through of model validation revealed that heavy collinearity existed within the data set used for model creation. As an attempt to mitigate collinearity, a couple of harmful variables (K and OBP) were removed from the data set. Despite clear improvements, validation methods still suggested that a near unacceptable level of collinearity existed. The improvements were deemed acceptable enough, and an accuracy test for both data sets still shows that the model is near perfect prediction accuracy (~95%).

The proposed model has proven to be reliable, but there are certainly limitations. The data sets used in model creation and validation only consisted of four year time spans (2010-2013 and 2014-2017). This may seem problematic if the model were applied to longer or shorter time spans, however, the model uses variable averages across the used time span. Therefore, the model should retain its accuracy and reliability when used for longer time spans, but it may lose reliability when used for shorter ones. Longer time span averages will help filter out outlier seasons, while shorter ones will not. Future research may test this theory.

It is possible that other variables that were not considered are more ideal as predictors in this model. There are many offensive statistics in baseball, most of which are certainly collinear with one another as proven by this project. Perhaps variables such as age or physical attributes may aid in home run prediction, which is something future research may also test.

In conclusion, the model proposed in this project has been proven by multiple sources of validation to be accurate and reliable. The model was tested against two different sets of data consisting of the same number of years in a different time frame. Both sets of data were tested against various validation methods, including tests against multicollinearity, partial least squares, autocorrelation, and heteroscedasticity. Each of these validation methods produced nearly identical results for both sets of data. The model has been proven to be accurate and reliable in predicting future home run outputs for an MLB player, given that player has participated in each of the previous four seasons with at least 200 plate appearances in each season.

**References**

Baseball in Japan. (2021, October 22). Wikipedia. Retrieved October 22, 2021, from <https://en.wikipedia.org/wiki/Baseball_in_Japan>.

Baseball Reference. (n.d.). *2012 Japan Central League Batting Leaders*. Retrieved October 20, 2021, from <https://www.baseball-reference.com/register/leader.cgi?type=bat&id=d45883da>.

Beals, N., Zhang, L., Law, M., Hwang, D., Acharya, J., & Basu, S. (2019). Physicochemical investigation into Major League baseballs in the era of unprecedented rise in home runs. *ACS Omega*, 4(23), 20109–20117. <https://doi.org/10.1021/acsomega.9b00405>

Beneventano, P., Berger, P. D., &amp; Weinberg, B. D. (2012). Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. *International Journal of Business, Humanities and Technology*, 2(4), 67–75. Retrieved September 12, 2021, from <http://www.ijbhtnet.com/journals/Vol_2_No_4_June_2012/7.pdf>

FanGraphs. (2021, September 9). *Major League leaderboards " 2010 batters " Dashboard: FanGraphs baseball*. Retrieved September 10, 2021, from <https://www.fangraphs.com/leaders.aspx?pos=all&amp;stats=bat&amp;lg=all&amp;qual=y&amp;type=8&amp;season=2010&amp;month=0&amp;season1=2010&amp;ind=0&amp;team=0&amp;rost=0&amp;age=0&amp;filter=&amp;players=0&amp;startdate=&amp;enddate=>.

Glen, S. (2021, August 15). *Durbin Watson Test & Test Statistic*. Statistics How To. Retrieved November 7, 2021, from <https://www.statisticshowto.com/durbin-watson-test-coefficient/>.

Goeman, J., Oosting, J., Finos, L., Solari, A., & Edelmann, D. (2021, October 26). *The Global Test and the globaltest R package*. Bioconductor. Retrieved November 5, 2021, from <https://www.bioconductor.org/packages/release/bioc/vignettes/globaltest/inst/doc/GlobalTest.pdf>.

Home run (HR). (2021). MLB.com. Retrieved September 10, 2021, from <https://www.mlb.com/glossary/standard-stats/home-run>.

Kim, J. (2019, July 15). *Multicollinearity and misleading statistical results*. Korean journal of anesthesiology. Retrieved November 3, 2021, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/>.

Major League Baseball Schedule. (2021, November 17). Wikipedia. Retrieved December 7, 2021, from <https://en.wikipedia.org/wiki/Major_League_Baseball_schedule>.

Major League Baseball finishes with 6,776 homers, 11% above previous record. (2019, September 30). USA Today. Retrieved September 10, 2021, from <https://www.usatoday.com/story/sports/mlb/2019/09/29/mlb-home-run-record/3816155002/>.

Pette, J. (2021, May 17). *Baseball and Machine Learning: A data science approach to 2021 hitting projections*. Medium. Retrieved September 10, 2021, from <https://towardsdatascience.com/baseball-and-machine-learning-a-data-science-approach-to-2021-hitting-projections-4d6eeed01ede>.

Prabhakaran, S. (2021, June 5). *Linear regression - a complete introduction in R with examples. Machine Learning Plus*. Retrieved November 5, 2021, from <https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/>.

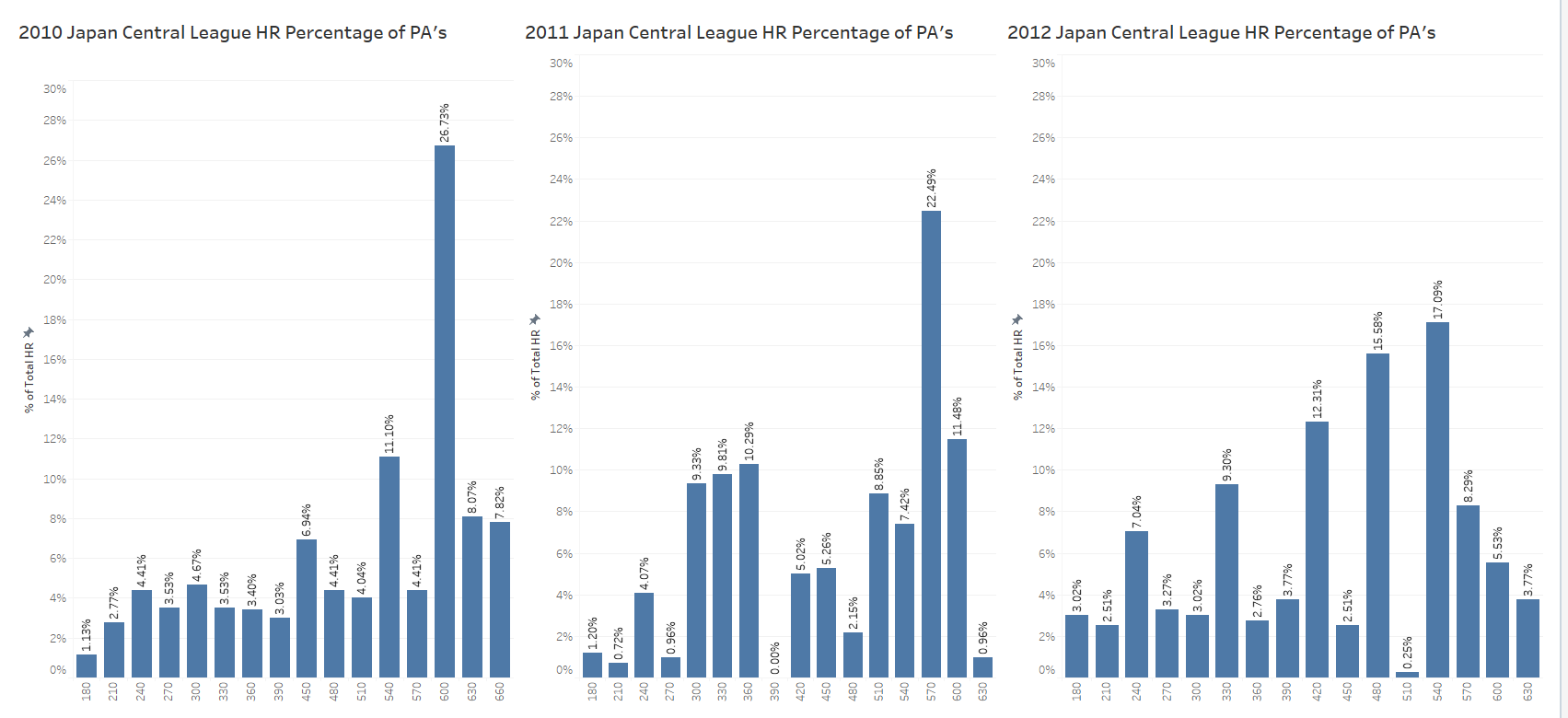
Wehde, W., Jenkins-Smith, H., Ripberger, J., Copeland, G., Nowlin, M., Hughes, T., Fister, A., & Davis, J. (2020). *Quantitative research methods for political science, public policy and public administration for undergraduates: 1st edition with applications in R. Bookdown*. Retrieved November 4, 2021, from <https://bookdown.org/wwwehde/qrm_textbook_updates/ols-assumptions-and-simple-regression-diagnostics.html>.

Zach. (2020, December 31). *The Breusch-Pagan test: Definition & Example*. Statology. Retrieved November 3, 2021, from <https://www.statology.org/breusch-pagan-test/>.

**Appendix A**

Data Biases and Exploring the Data

*HR Percentage in the MLB vs the Japan Central League*



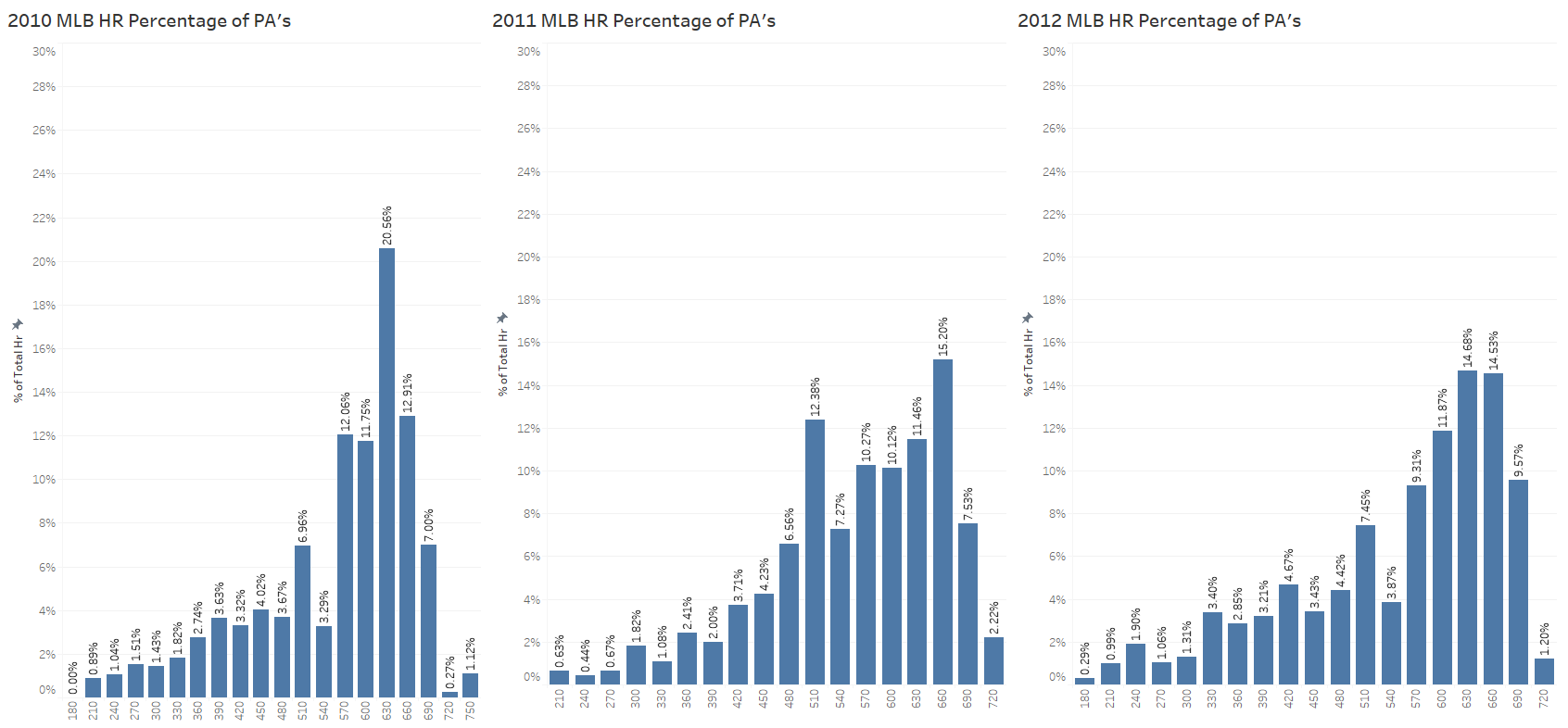


Figure 3. These bar charts show how the HR percentage of PA differs between the MLB and Japan Central League.

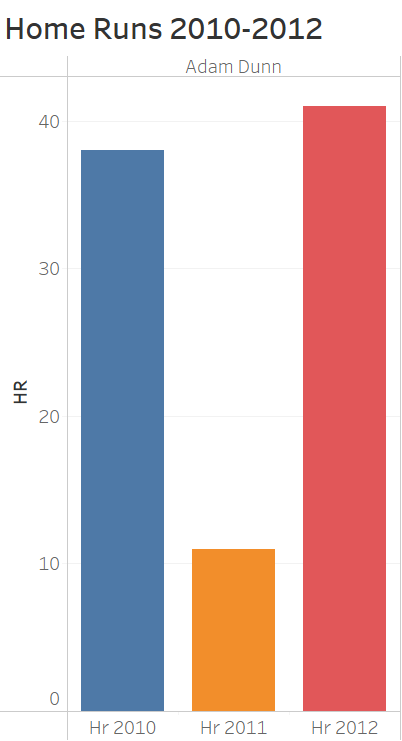
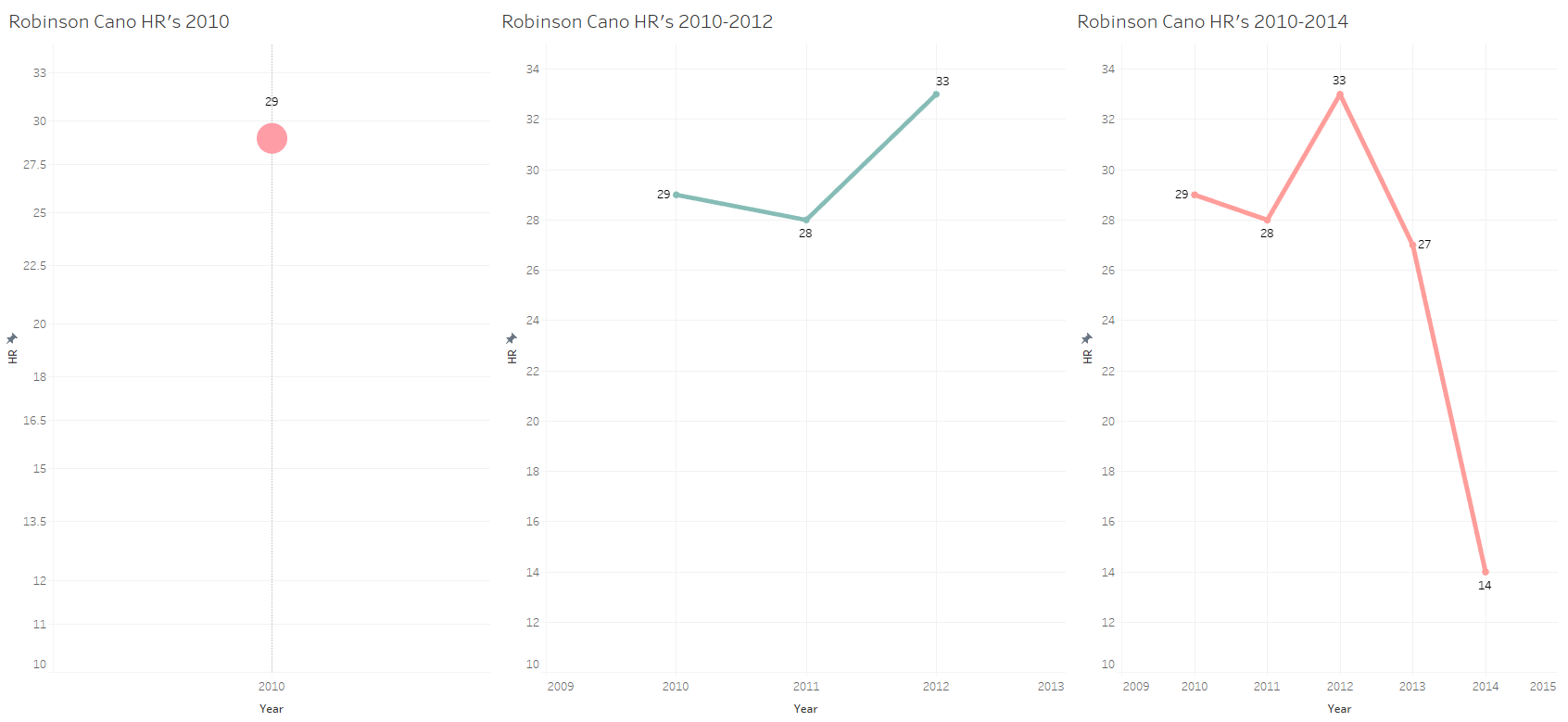


Figure 4. This bar chart illustrates how HR values for Adam Dunn varied between each season from 2010-2012.

*HR Count with One, Three, and Five Seasons for Robinson Cano*

Figure 5. These line graphs show Robinson Cano’s HRs using one season, three seasons, and five seasons

**Appendix B**

Initial Model Build

*Libraries loaded into Rstudio*

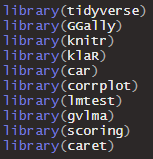
****

Figure 8. This figure displays the libraries that were used in R to code the model.

*Summary of the HR Model*

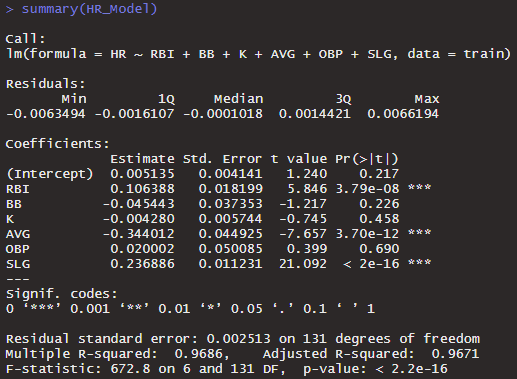
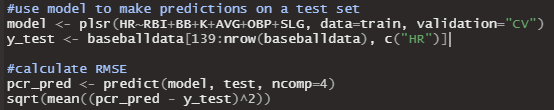


Figure 9. This figure is the summary of the original HR Model regression run in this study.

**Appendix C**

Initial Model Validation

*Root Mean Square Error*



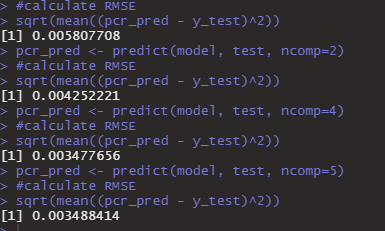


Figure 16. This figure displays the R code used to create a model and test set, and calculate an RMSE value for each number of components, represented by ncomp. The outputs are shown in the bottom figure, showing that a higher ncomp value leads to a lower RSME value. This means that including more components in the prediction process leads to higher model accuracy.

*Summary of the HR Model2*

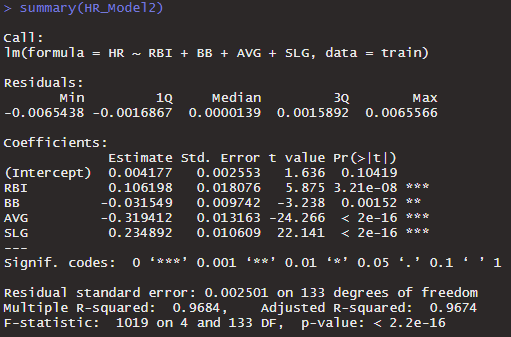


Figure 18. This figure shows an updated summary of the model after removing the variables OBP and K.

**Appendix D**

Subsequent Model Build and Summary

*R Code for Creation of the Subsequent Model*

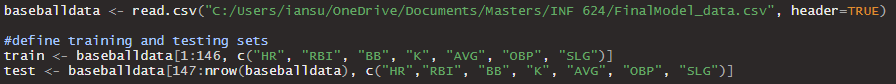




Figure 24. This figure displays the code used for the subsequent model creation.

*Summary of the Final HR Model*

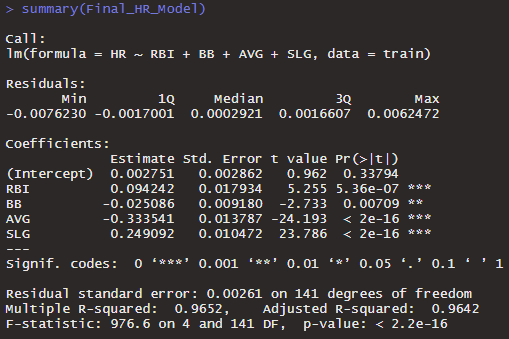


Figure 25. This figure displays the summary of the subsequent model (Final HR Model).

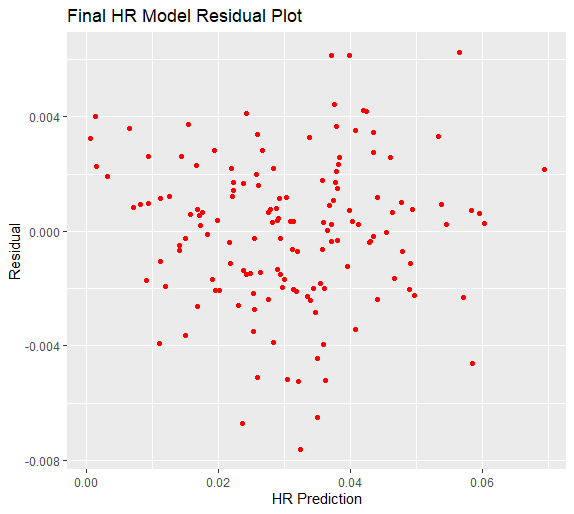


Figure 26. This figure shows the residual plot of the Final HR Model that has the residuals on the y-axis and the HR predictions on the x-axis.

*Outlier Test of the Subsequent Model*

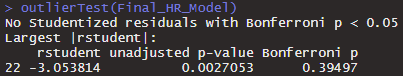


Figure 27. This figure is an outlier test to find any unwanted influence in the model.

**Appendix E**

*Condition Index of the Subsequent Model*



Figure 28. This figure displays the Condition Index of the subsequent model (Final HR Model).

*Correlation Matrix of the Subsequent Model*

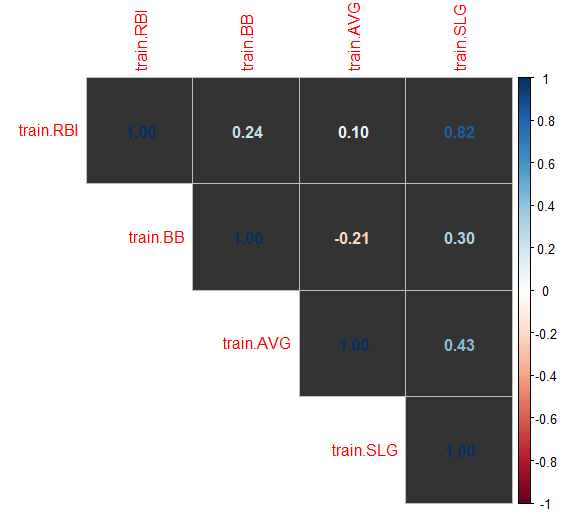


Figure 29. This figure represents the correlation matrix of the independent variables in the subsequent model (Final HR Model).

*Breusch-Pagan Test of the Subsequent Model*

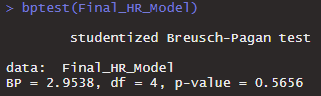


Figure 30. This figure displays the Breusch-Pagan test of the subsequent model (Final HR Model).

*Global Validation of Linear Model Assumptions for the Subsequent Model*

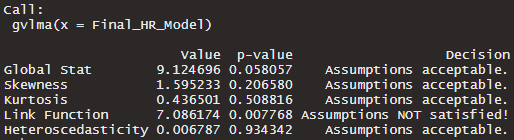


Figure 31. This figure is the output of the gvlma function testing various linear model assumptions including Global Stat, Skewness, Kurtosis, Link Function, and Heteroscedasticity.

*Durbin-Watson Test of the Subsequent Model*

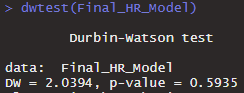


Figure 32. This figure displays the Durbin-Watson test checking for autocorrelation in the subsequent model (Final HR Model).

*Prediction Accuracy of the Subsequent Model*

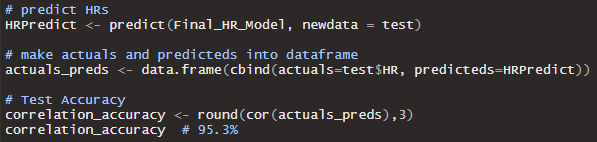


Figure 33. This figure presents the code to measure the accuracy between the predicted values and the actual values for the subsequent model, along with the accuracy output. (Final HR Model).

**Allocation of Work**

Below are the team members for this project and the role each one played in the assignment.

It is to be noted that a plan for all sections and tasks were discussed amongst the members

before completion.

**Kyle Hirsch** - Initial Project Plan: Completed the task/problem statement, 1 literature review, data source and data description, evaluation, descriptive analysis, and analysis.

DGP/Bias Check: Completed data set overview, data generation process, and 1 bias.

Descriptive Analysis: Data organization, dataset description, 2 biases, exploring the data, data quality section, 2 questions, conclusion, and 3 figures.

Initial Model Build: Data cleaning and work in Rstudio, Introduction, Data Overview and Collection, 2 biases, Exploring the data, Modeling the data, Validating the Model (Partial Least Squares), Conclusion, Appendix.

Initial Model Build Presentation: Objectives, Model Introduction, Data Overview, Collection, and Generating Process, Descriptive Analysis (4 slides), Partial Least Squares, Conclusion.

Final Analysis Paper: Introduction, Data Overview and Collection, 2 biases, Exploring the data, Modeling the data, Validating the Model (Partial Least Squares), Abstract, Results and Discussion, Appendix, References.

**Ian Suleski** - Initial Project Plan Completed the Introduction, 2 literature reviews, Goals, Bias Check, Initial Model Build, and the Conclusion.

DGP/Bias Check: Completed the introduction and 2 biases.

Descriptive Analysis:data organization, introduction, dataset description, DGP, 1 bias, exploratory visual analysis, 1 question, and 12 figures.

Initial Model Build: Data cleaning and work in Rstudio, Data Overview and Collection, DGP, 1 bias, Validating the Model (Summary statistics, Multicollinearity, Heteroscedasticity, Autocorrelation, Accuracy Test), Appendix.

Initial Model Build Presentation: Modeling the Data, Predictive Model Algorithm, Validating the Model, Model Summary, Residual Plot, Outliers, Multicollinearity Test, Correlation Matrix, Multicollinearity Correction, Heteroscedasticity Test, Auto Correlation Test, Model Accuracy, References.

Final Analysis Paper: Data cleaning and work in Rstudio, Data Overview and Collection, 1 bias, Validating the Model (Summary statistics, Multicollinearity, Heteroscedasticity, Autocorrelation, Accuracy Test), Subsequent Model Build and Validation.