STATS 202A Final Report

Nicklaus Kim

Fall 2021

1 Introduction

Statistics have been a huge part of baseball for a long time. They have been a crucial factor to understanding the game arguably since 1845, when the first box score appeared in the *New York Morning News* with batters' columns including only runs and outs [1]. In modern times, baseball stats have evolved quite a bit past that archaic beginning, and they have a very wide spectrum of use cases, from simple counting for designating league leaders to advanced "sabermetrics" and predictive modeling in Major League Baseball (MLB) teams' analytics departments. Our area of emphasis in this paper is home runs, one of the most relevant statistics today and a popular barometer for the skill level and popularity of big name players.

We shall look at ways of modeling a player's total number of home runs hit during a regular season using his other batting statistics as explanatory variables; we explore not only the trends and relationships visible on the surface but also look for an underlying framework to model the causal link between home runs and other stats. Our analysis in this project lies somewhere in the middle part of that aforementioned spectrum of baseball statistics analysis; one simple practical application of the results we aim to reach is to the world of fantasy baseball leagues (or more cynically, to sports betting and gambling). Our methods will include some exploratory data analysis, linear regression models, kernel density estimation, and kernel regression models. We present all plots and figures related to these methods for additional commentary in the Appendix. In these endeavors we hope to find surprising and useful results which would prove to be illuminating to some degree to the baseball mind, however seasoned.

2 Obtaining and Preprocessing the Data

Data for our analysis was collected from Baseball Savant [2, 3], an MLB-owned official hub for player and team data, including Statcast[1] data. The data we

[1] Statcast is a high-speed, high-accuracy automated tool developed to analyze player movements and athletic abilities in MLB. By combining camera and radar data, dozens of physical metrics relating to every aspect of the game (pitching, hitting, base running, and fielding) can be obtained.

chose to search and download was for the 2021 MLB regular season; in addition, the data search was constrained to only include qualified hitters; that is, hitters who had a statistically significant number of plate appearances throughout the season (roughly speaking, 3.1 PA per game).

The data contained a plethora of player batting statistics, such as home runs, hits, batting average, slugging percentage, and many more. In total, 132 players were included in the data set.

3 Exploratory Data Analysis

When considering the question of predicting home run (HR) output from other variables, there are a couple of intuitive yet naive hypotheses one may have. One idea is the use of a common batting measure, such as batting average or more simply, the number of hits. More generally, we could even consider the use of the number of singles, doubles, triples, i.e., non-HR hits. After all, intuition may suggest to some that these could work since many top players have the best ability to hit the ball overall; they often have the highest HRs, BA, hits, etc.

We can check this guess by looking at correlation plots; we plot home runs versus batting average and home runs versus total hits [Figures 1-2]. Unfortunately, we see that there is very little to no linear relationship between either of these two variables and a player's home run total. This lack of relationship can be easily verified by attempting to use linear regression to model it. We demonstrate an example of this with a linear model using similar "box score" variables: the number of singles, and doubles, and triples as the features. The end result is quite poor predictive performance as expected; the model has a very low R^2 score of 0.2356 and is not appropriate to use here.

The possibilities are plentiful with how many baseball metrics exist nowadays; we list a few of them experimented with (with no real success): batting average (BA), slugging percentage (SLG), runs batted in (RBI), strikeouts, walks, etc. The end result and takeaway from models using any combination of these aforementioned, more traditional box score statistics is that they are simply not well-suited for predicting home runs, since baseball just has extremely different, sometimes juxtaposing types of hitters. There exist "power hitters" and "contact hitters," for instance. It makes sense, in the end, that knowing how often a batter gets hits or RBIs should not have a bearing on his home run proclivity; we would still have no idea what *type* of hitter he is. So, we must look deeper, below the surface to discover more about a batter's type of hitting and in turn, his efficacy at knocking the balls into the outfield stands.

4 Methods

4.1 Linear Regression

A new approach would be to consider the data supplied via the previously mentioned Statcast. These data, derived from camera observations, include raw physical measurements of batting events, and we will discover the use of these newer, more advanced statistics in this section.

Statcast captures a mind-boggling amount of raw data; on average, an estimated *7 terabytes* of data is collected in a single MLB game. The possibilities for potential feature variables for our home run model are almost endless. After doing some pre-screening, we settle on a couple that we are interested in: exit velocity (EV) and launch angle (LA). These are also some of the most common and ubiquitous measurements discussed when evaluating a hitter today. Exit velocity is simply how fast, in miles per hour, a ball was hit by a batter; launch angle is how high, in degrees, a ball was hit by a batter. It should be mentioned that for these two measurements, the averages are calculated only from swings resulting directly in hits.

We can see from correlation plots (Figures 3-4) that these two new variables would appear to have moderately strong correlation with home runs, so we turn again to a linear regression model. In implementing linear regression, we first try a model using EV and LA as our two predictors, to serve as a baseline model of sorts. (Interestingly, EV and LA are practically uncorrelated (Figure 5) and so we are safe to proceed without the possibility of multicollinearity issues). From this model, we get an R^2 score of 0.6398, which is already a massive improvement on the previous model which used the more traditional variables, batting average and total hits. The residual plots (Figures 6-7) reflect that the usual regression assumptions are being upheld, so our model is valid. This model actually exhibits some predictive ability, and confirms our suspicion that there exists a solid linear relationship between EV/LA and home runs.

We can take things a step further and produce even more robust linear models by introducing other predictors. After all, more sophisticated models often perform better, especially when the existing predictors are all deemed significant (which we have in this case). Again, we can do a lot of pre-screening and experimentation with using different combinations of predictors (since there are so many to choose from). But, we eventually settle upon one which is still quite simple yet comes with an interesting point.

We first introduce one last Statcast measurement, barrels. The underlying mechanism for classifying a batted ball as a barrel is somewhat convoluted and therefore will not be discussed in large detail here but essentially, a barrel boils down EV and LA into a single, more effective metric; a batted ball counts as a barrel if its EV and LA qualify above a certain threshold. With that being said, our final model that we present here is home_runs barrels + doubles. The predictors chosen here are both deemed statistically dependent from the model's summary, and they are quite interesting for a couple reasons. We see that the number of doubles, a more traditional, “counting” statistic, does become a significant predictor when coupled with our less traditional Statcast variable, barrels. What is even more interesting is that we can report from repeated trials that trying the same model with other hit types — singles and triples — does not lead to a similar result. This is likely due to the fact that doubles are perhaps the most indicative of a “solid piece of hitting,” thus players with more affinity for hitting doubles could have greater home run success. By contrast, singles typically come from shorter hit balls (or even grounded/mishit balls); meanwhile, obtaining triples is heavily reliant on the player's speed when running the bases[1].

Our model is again somewhat basic but does a good job of modeling the linear nature of the relationship; the R^2 here is 0.7452. Adding more predictors could of course increase the model's score even more, but we present an overall parsimonious model in this paper so as to not over-complicate the underlying domain-backed analysis. We can also go further by noticing that while our causal relationship does seem roughly linear in nature, we can still turn to building kernel regression models to attempt to capture any nonlinear trends in the data.

4.2 Kernel Density Estimation

As we have seen, our new predictor variables do a significantly better job at capturing the variation in our outcome variable, home runs. We now turn to alternate methods of analyzing these relationships, specifically in an effort to catch any nonlinear trends. We will consider each of the Statcast features previously discussed — exit velocity, launch angle, and barrels — and model each's influence on home run output.

First, we look at non-parametrically estimating the distribution of each of our new variables, particularly the output, home runs; we achieve this using kernel density estimation. For greater computing efficiency, this method is programmed in C, then this C code is called in R. We are chiefly interested in visualizing each of the variable distributions, so we create some quick kernel density estimation plots (Figures 8-11). From these, we observe a few clear trends. First, it is clear that while EV and LA share roughly the same distribution (approximately normal, it seems), barrels exhibit a wildly different pattern. This KDE plot shows a heavy right skew, meaning that most players do not produce many barrels at all, compared to the elite few who may tally up quite a few more. These indications will be interesting to explore further in the next section, when we consider kernel regression.

4.3 Kernel Regression

We now run kernel regression models using these three new statistics as the independent variables and home runs as the dependent variable (Appendix). Again, we accomplish this using C to reduce the computational cost. Included in the kernel regression plots accompanying the models (Figures 12-14) are the 95% confidence interval bands, so we can also look at a more general approximation of the predicted densities.

We find that in each of the cases, the number of home runs steadily increases as the value of the feature variable increases. Intuitively, this makes sense since batters who can hit the ball harder and higher should be expected to produce

[1] Also, we note that the number of triples in MLB is significantly lower than for the other hit types, and so there is not enough variation here to be a very reliable predictor

more home runs; after all, a certain distance and height is required to hit the ball that far. We now have a good general idea of how home run output may change based on a player's EV or LA, as well as on his barrel rate, which is just the collective measure of EV and LA. This gives us some great takeaways on the reliance of home runs on such Statcast measures, especially as opposed to depending on the other traditional, box score metrics.

5 Conclusion

As we have seen, the matter of predicting home run totals is one that is extremely intricate and complex, with so many possible influences existing. However, we have shown through a handful of statistical models that good progress can still be achieved using some key intuitive ideas from baseball. First of all, we have demonstrated that many or most of the lauded traditional metrics, such as batting average, slugging percentage, or even total hits do not have much or any real bearing on home runs, at least when considered by themselves. Overall, we have seen that raw physical measurements like exit velocity and launch angle are much more indicative of home run output than other traditional statistics like batting average, hits, etc. Of course, the overall conclusion that the quality and force of contact can predict HRs makes sense intuitively; the harder and higher a player can launch the ball on average, the more likely he is to churn out more home runs. The *surprising* part is that we are effectively ignoring any swings that did not directly result in hits. This means we are not considering most of the swings of the bat that occur in the game; for example, a player could *whiff* all the time at the plate but it does not make a difference to the data we collected or to the models we built. We are not considering some incredibly valuable aspects of evaluating a player's quality and skill, like his discipline or ability to make contact in the first place

There exist many possible avenues for exploring this topic in higher detail; of course, time and space is limited so we have only presented a few pivotal paths for model building that can then in turn be further improved and explored. Here are just a few of the ways in which one may alter our approach to this problem and possibly find more refined models:

- Find other similar Statcast data and add more features for a more sophisticated model. The plentiful nature of this data, as mentioned before, means there are too many possibilities to name, but a few of the more likely ideas include sweet spots, expected BA, fly ball rate, etc.
- Split EV/LA data by type of play so that we can isolate balls hit into the air — after all, who cares about how fast balls are crushed into the ground, or perfectly angled balls that have no force behind them

Lastly, of course, there is an entire other half to the story: pitching. We have basically ignored the various influences that the opposing pitching may have on overall home run output. For instance, the distributions of the types of pitches that different players receive throughout the season differ depending on the opposition's evaluation of said player's hitting tendencies. We have essentially assumed in all of our cases here that any pitching fluctuations may cancel out/become insignificant when taken over all of the games against every team throughout the entire season. So the home run outcomes could for all we know change dramatically with the influence of the other half of the battle; after all, just as there would be no home runs without the batters hitting them, they also would not exist without the pitchers giving them the balls to hit.

6 Appendix

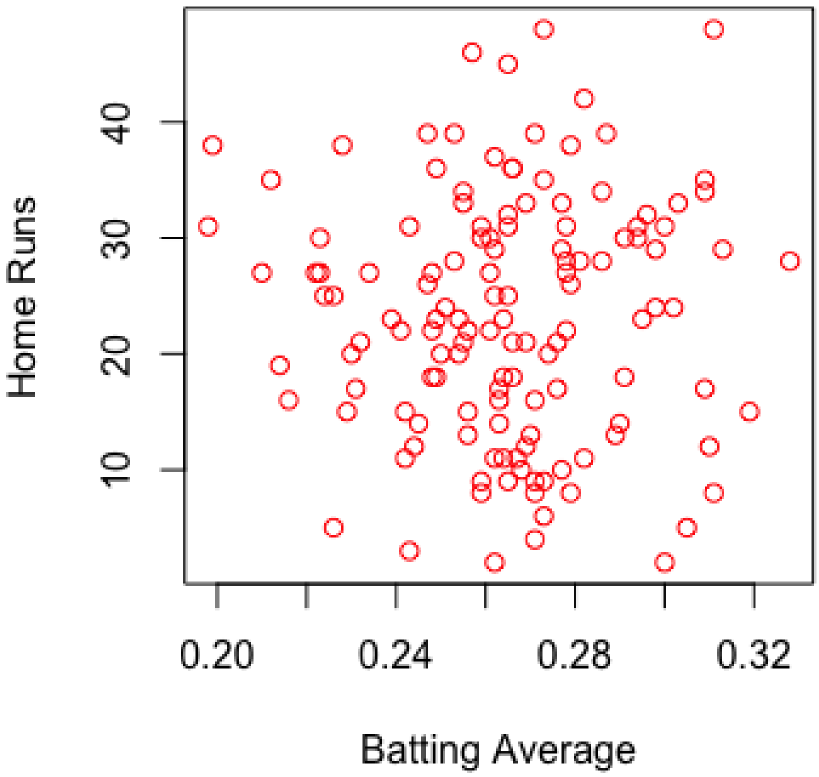

Figure 1: Scatter plot

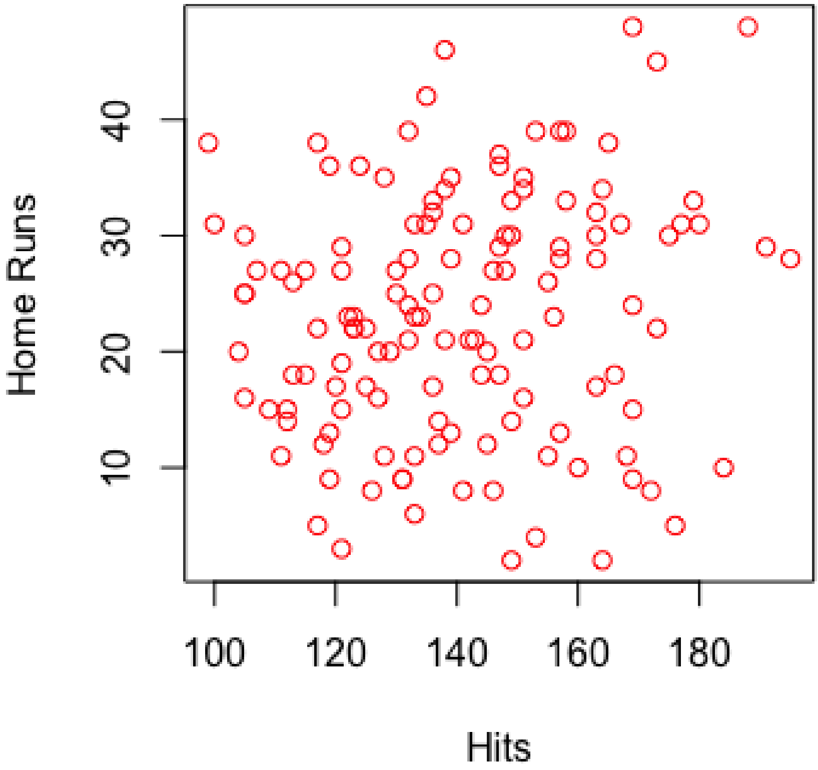

Figure 2: Scatter plot

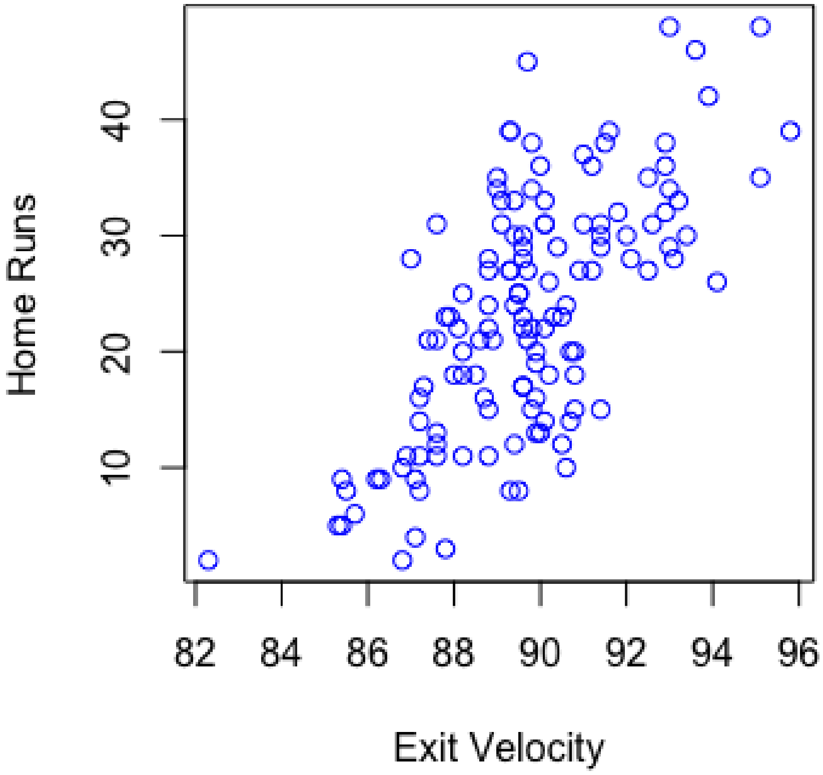

Figure 3: Scatter plot

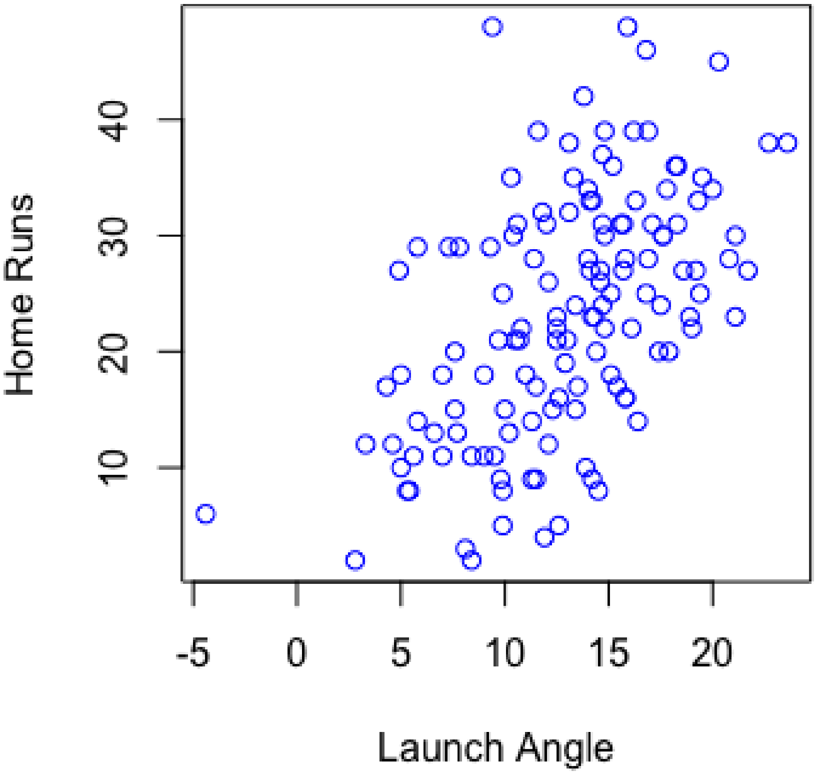

Figure 4: Scatter plot

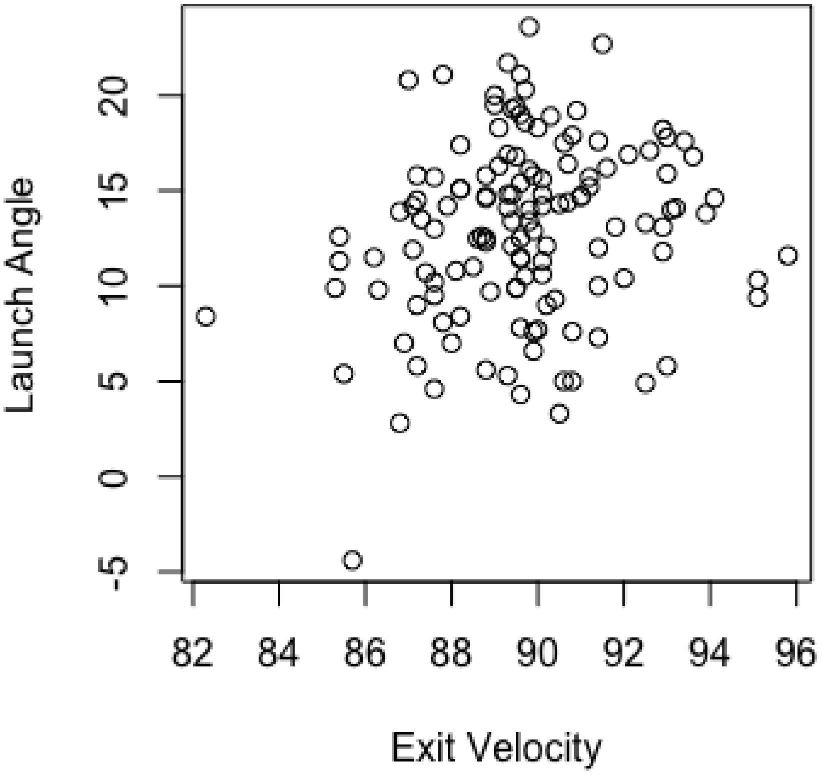

Figure 5: Scatter plot

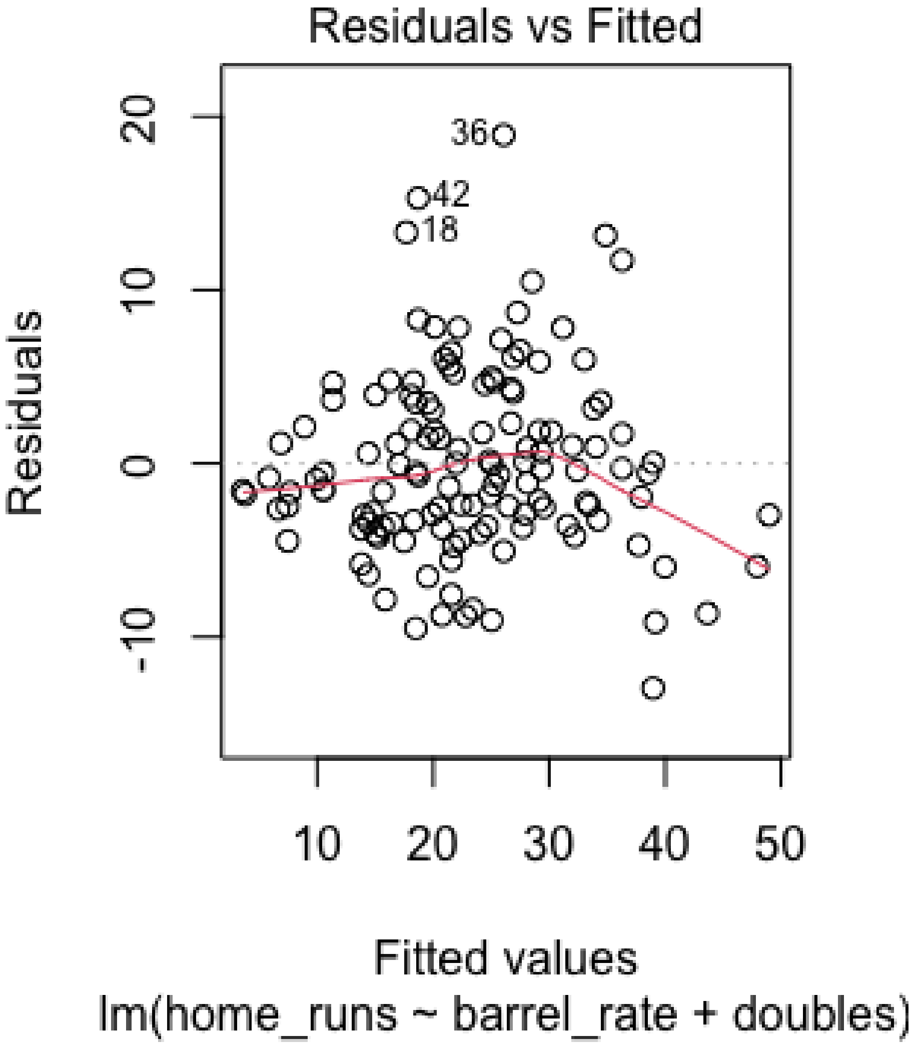

Figure 6: Residual plot

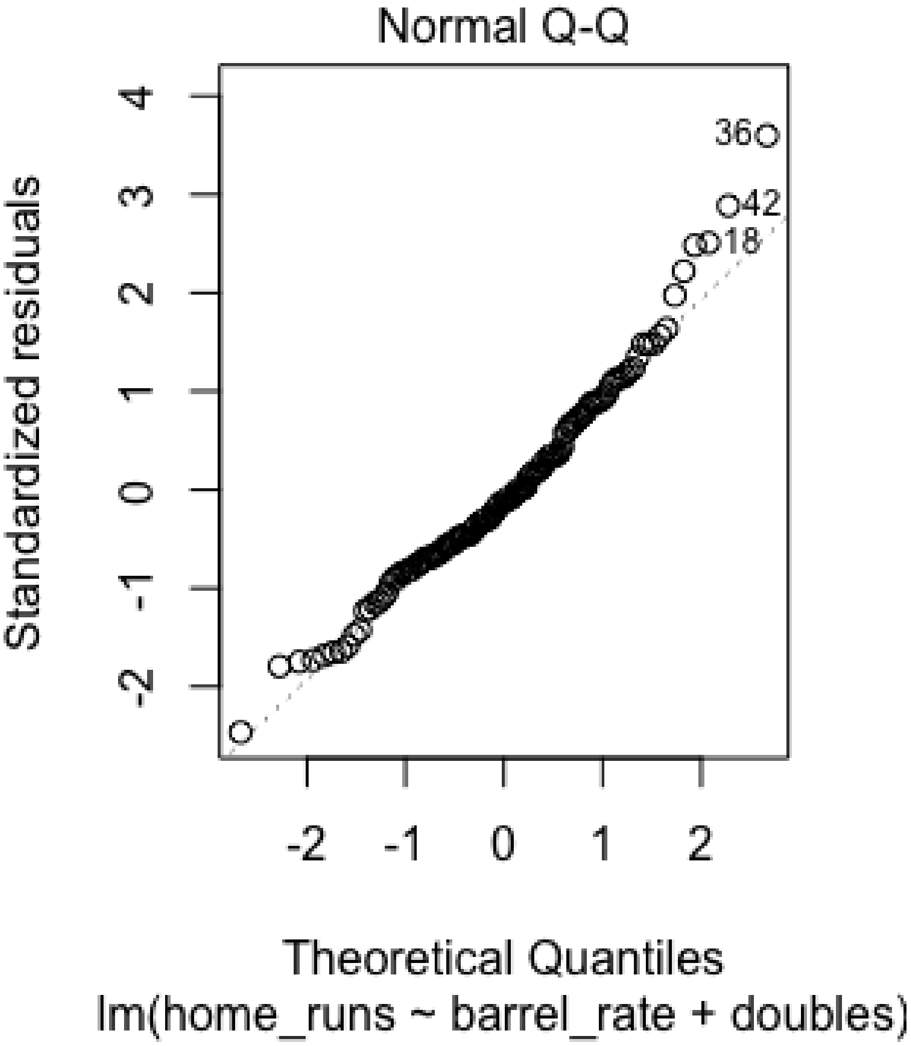

Figure 7: Residual plot

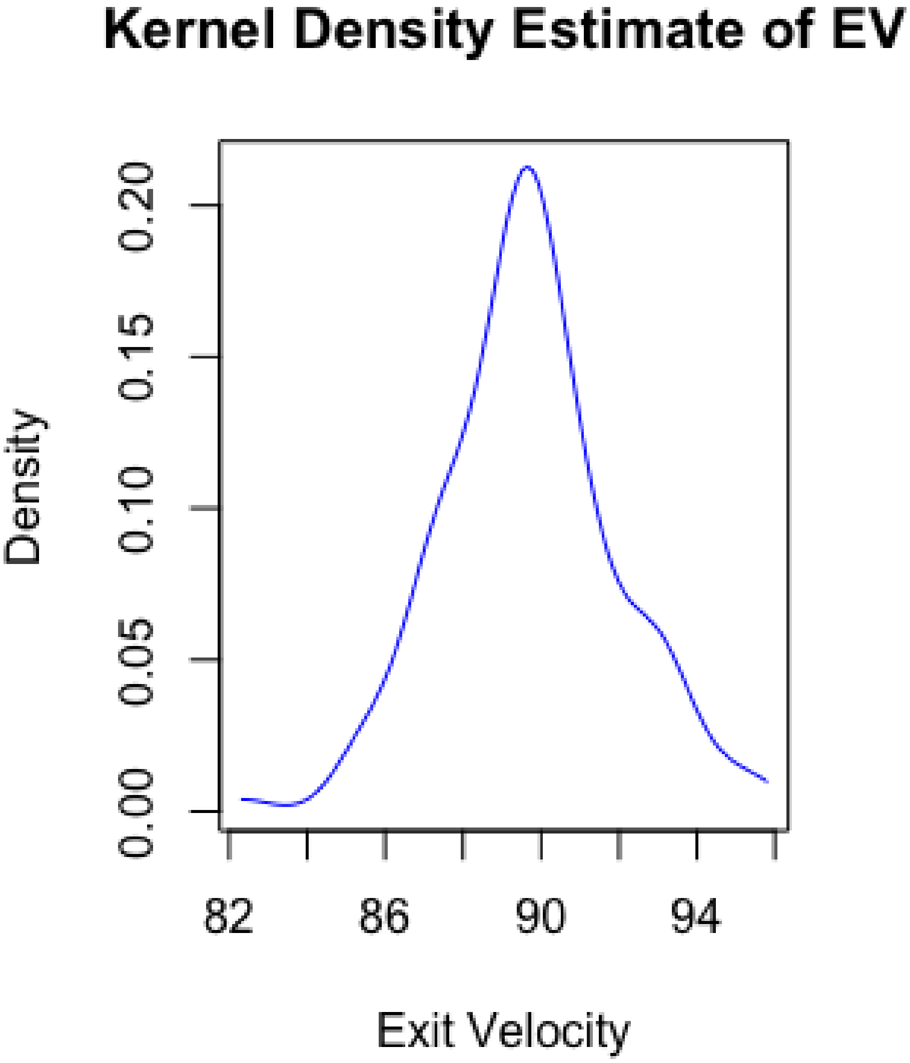

Figure 8: Kernel density plot

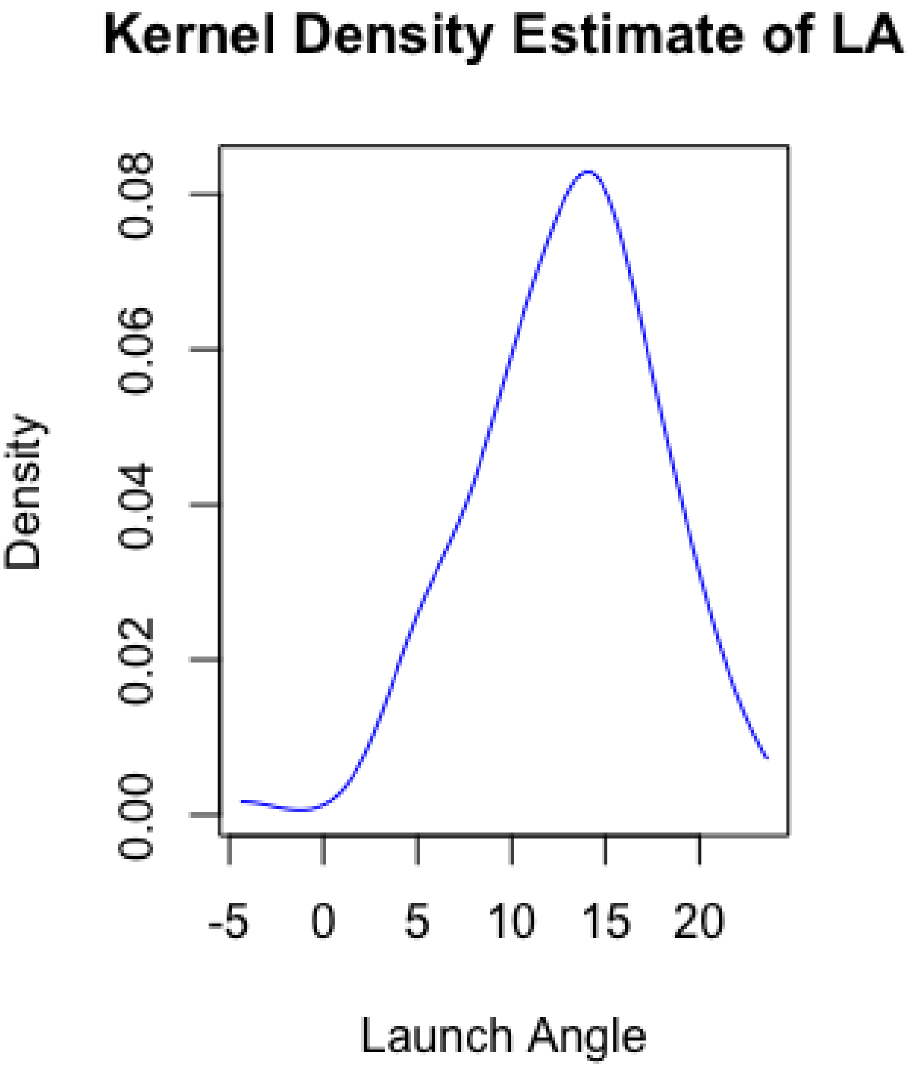

Figure 9: Kernel density plot

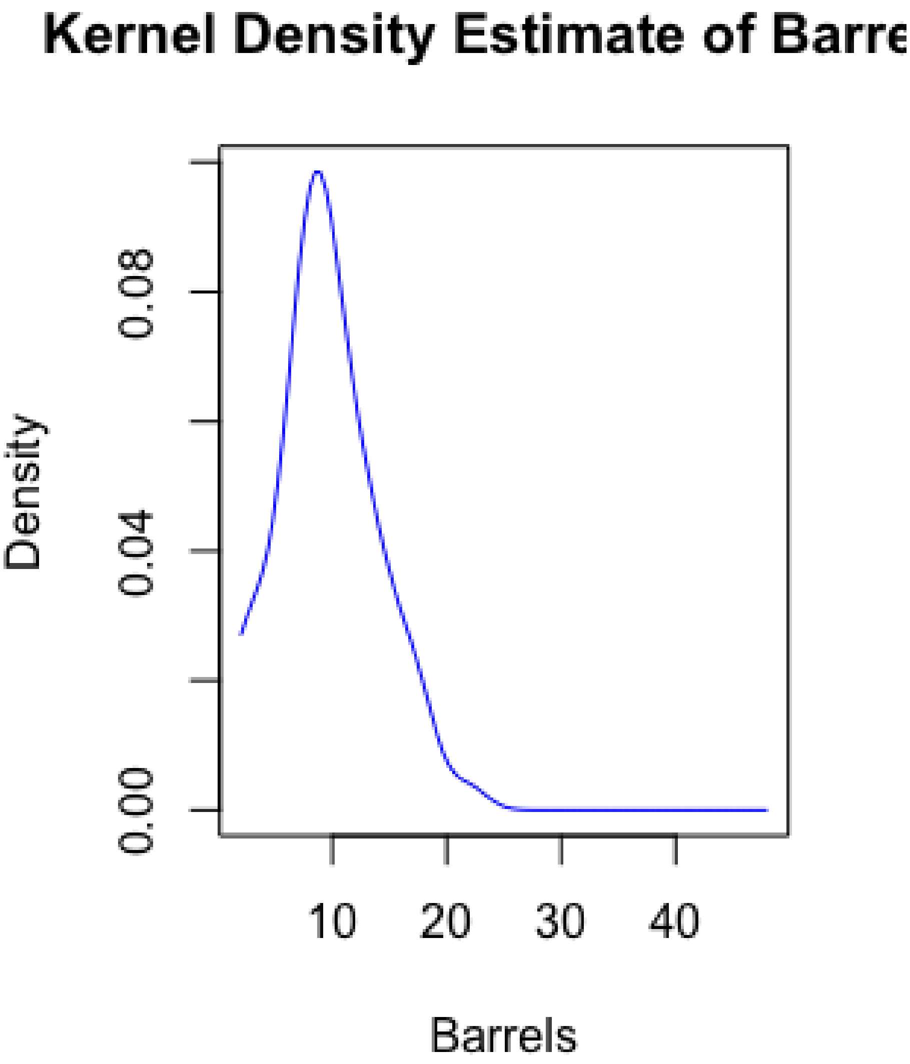

Figure 10: Kernel density plot

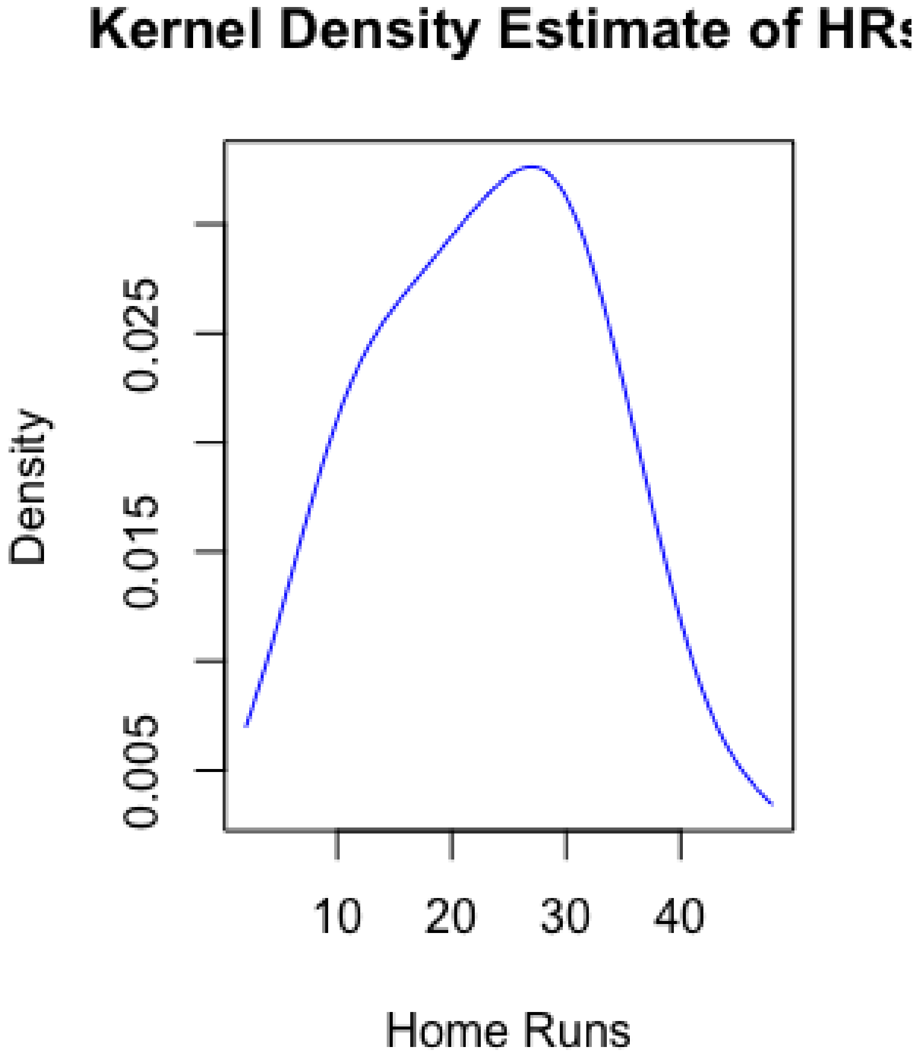

Figure 11: Kernel density plot

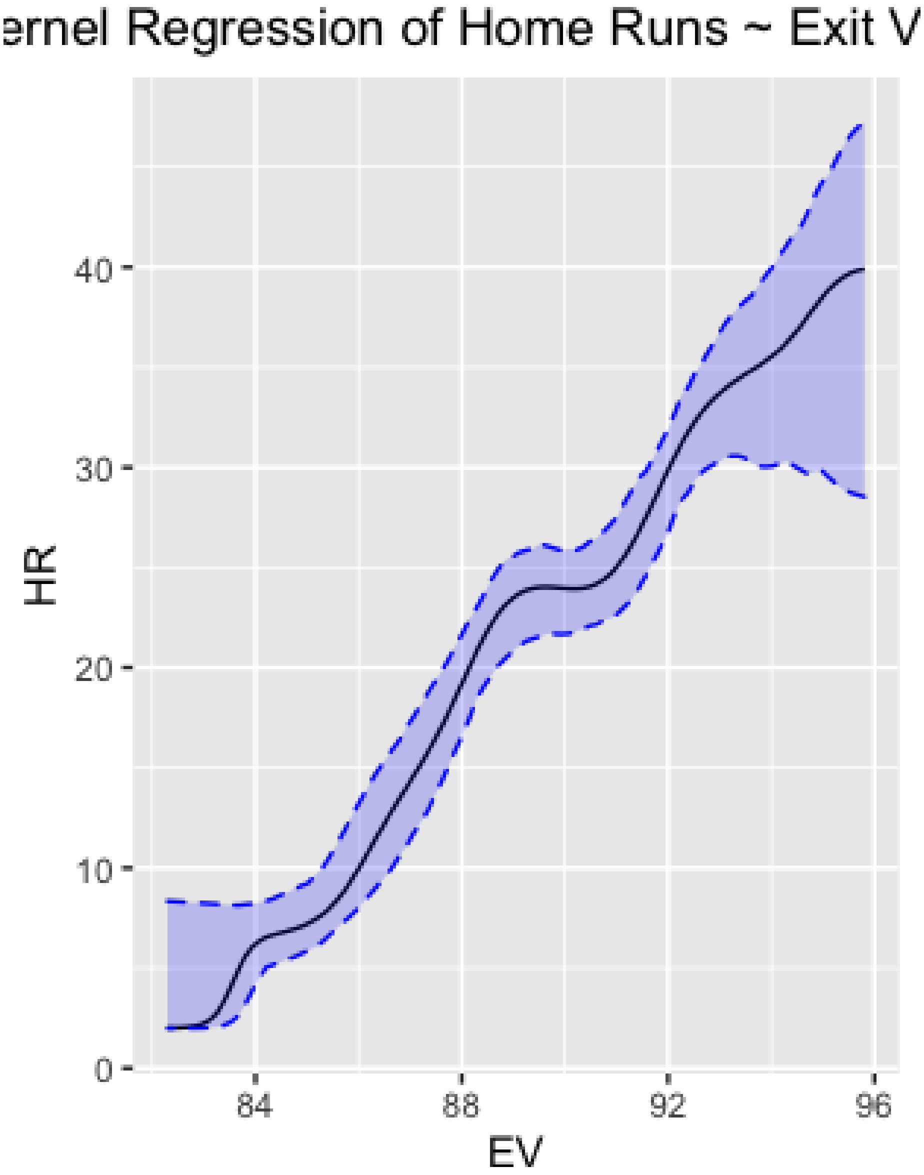

Figure 12: Kernel regression plot

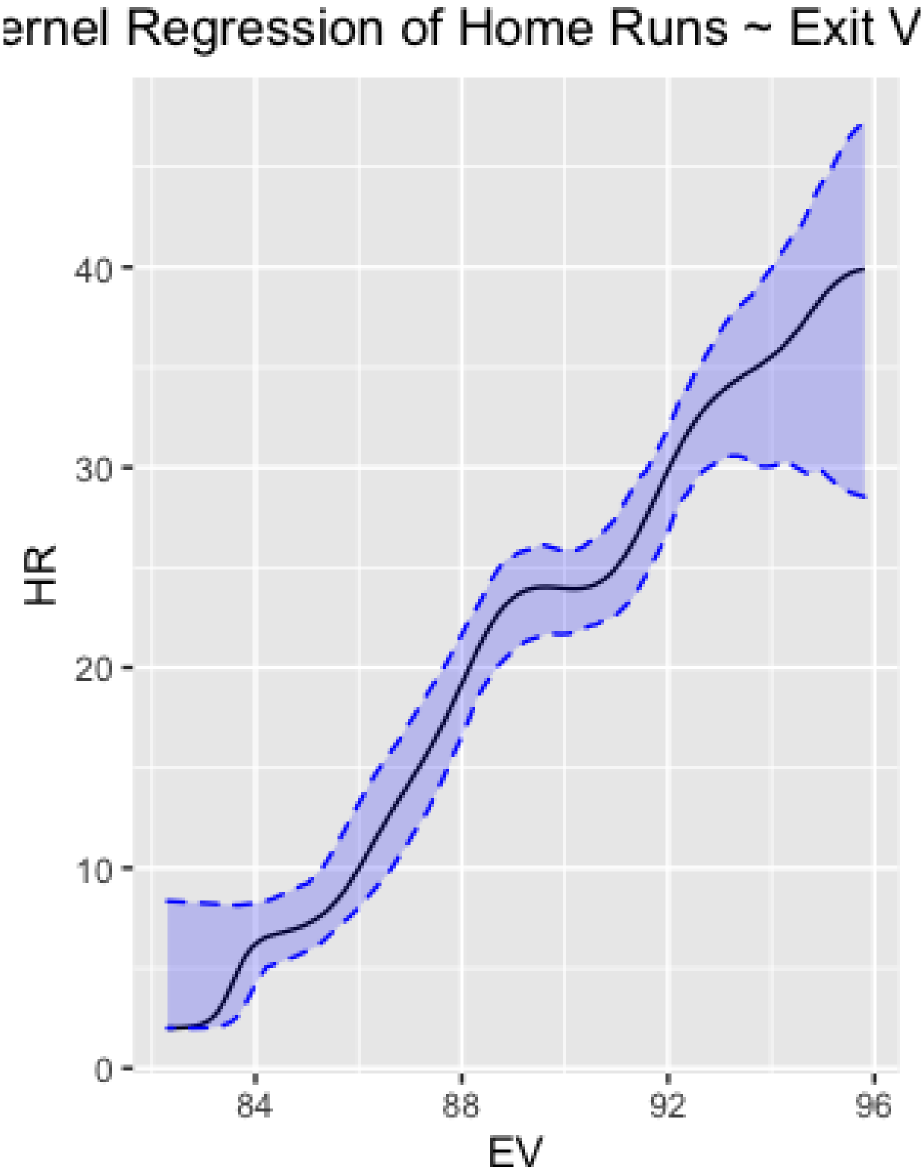

Figure 13: Kernel regression plot

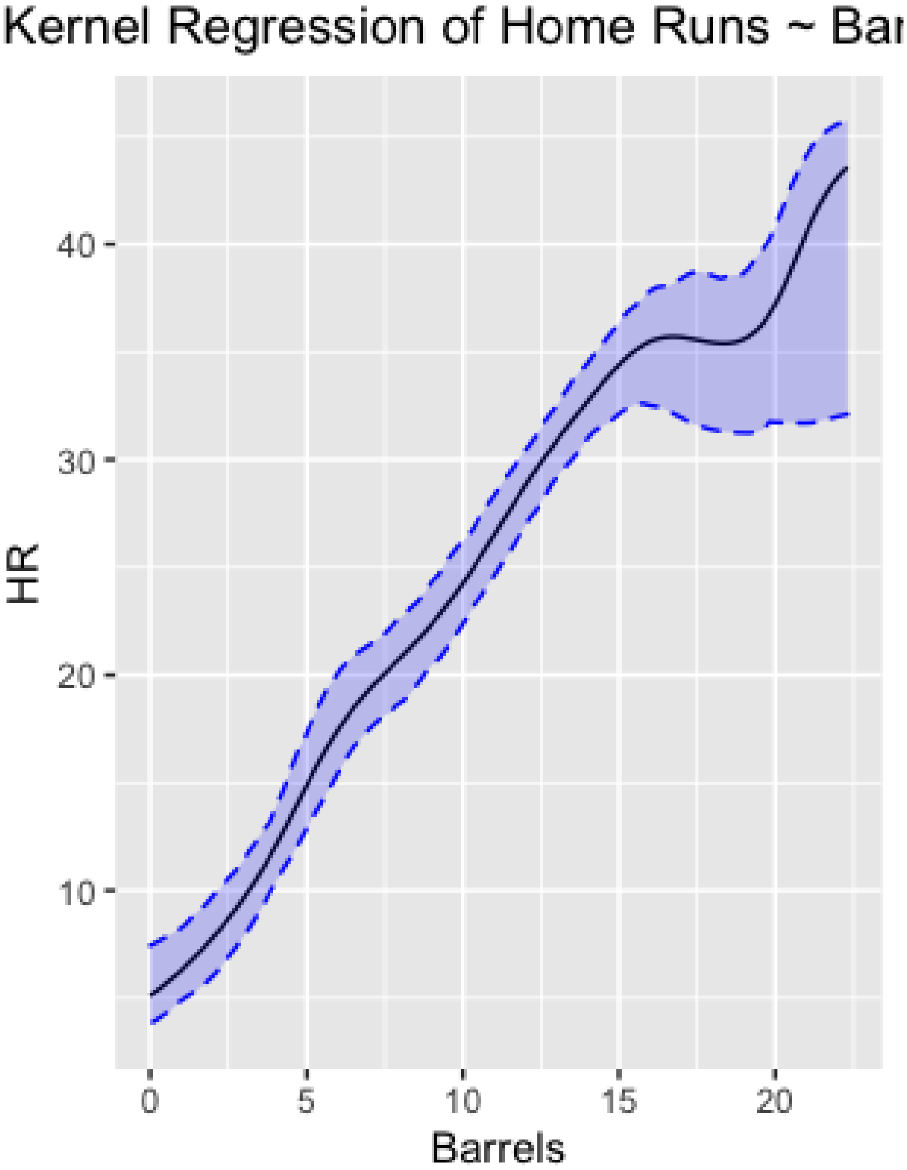

Figure 14: Kernel regression plot

7 References

1. https://www.espn.com/mlb/columns/story?columnist=schwarz_alan&id=1835745

2. https://www.baseballsavant.mlb.com

3. https://www.mlb.com/glossary/statcast

4. https://www.washingtonpost.com/graphics/sports/mlb-launch-angles-story/

5. https://blogs.fangraphs.com/expected-home-run-rate-2020-edition/