An Examination of Home Runs

Jake Coleman

1 Executive Summary

In this analysis I aim to investigate the causes of home runs, both at the pitch and pitcher level. A logistic regression that models the probability of hitting a home run suggests that the largest effects are those from the stadium and from inside location of the pitch. To model home runs allowed at the pitcher level, I used a hierarchical Bayesian overdispersed Poisson regression. While model diagnostics demonstrated decent fit, the PITCHf/x covariates did not appear to have strong effects on the mean home run rate. While there are other possible PITCHf/x variables to construct and test, this analysis suggests that home runs are best predicted at the pitch level, and most likely affected by factors specific to each pitch. Though results from this paper are largely intuitive, this analysis provides a statistically principled result that quantifies effects in a rigorous fashion.

2 Introduction

This analysis investigates the question, "Why do pitchers give up home runs?" Proneness to home runs, often measured by home run to fly ball ratio, is frequently used in the evaluation of a pitcher. However, acknowledging that pitcher is plagued by home runs is different from understanding why he allows them. Intuition gives some suggestions - command, movement of pitches, ball park effect - but the magnitude of these effects has not been public studied in a statistically rigorous fashion. Section 3 of this analysis will be conducted at the pitch level, investigating which pitch attributes are correlated with the probability of that pitch being hit for a home run. In Section 4 I model home runs allowed at the pitcher level, fitting a Bayesian Poisson regression model that allows for overdispersion of the data.

3 Home Runs on Pitches - Logistic Regression

Here I look to model the probability of allowing a home run, using logistic regression. See the Appendix (Section 6.2) for a brief review of logistic regression - where it comes from and the meaning of the coefficients. The motivation for this section is something simple and explainable with a principled approach.

3.1 Data

The data was taken from pitches during the 2013-2015 seasons, here subset to pitches in which a fly ball was hit in order to model home run to fly ball percentage. Observations with pitch type "Unknown" had no PITCHf/x information, and thus were discarded - an imputation scheme could have been implemented, but this affected less than 0.2% of all data points. Covariates considered were pitch type (as measured by PITCHf/x), stadium, whether the pitcher and batter had the same hand, whether there was a man on base or not, release velocity, movement in the x plane, movement in the z plane, total movement, whether the pitch was in the top of the zone, and whether it was inside to the batter. Total movement (change in Euclidian distance in the xz plane) was considered rather than individual x- and z-movement due to high correlation between total and individual movement and low correlation between the individual movements.

I am most interested in the variables controlled by the pitcher, so that, all else constant, I might estimate the effect of those variables on the probability of a home run.

There are some expected figures here - Miami and San Francisco suppress the home run, while Colorado and Toronto are among the league leaders in stadiums allowing home runs. It is worth noting that in 2015 the Mets changed the dimensions of their ballpark, and while this certainly resulted in more home runs being hit at Citi field (130, 37th percentile in 2014 compared to 177, 67th percentile in 2015), the three-year count is right in the middle (53rd percentile).

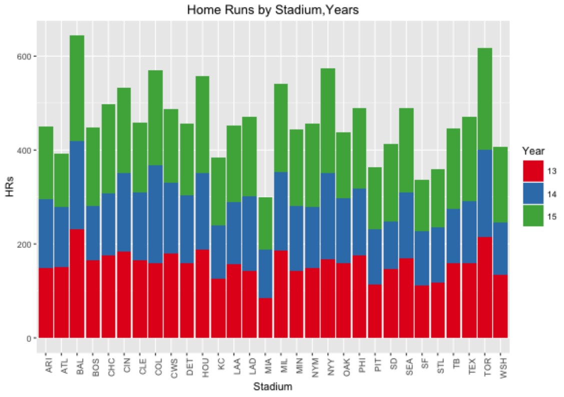

Figure 1: Home run counts by stadium

Pitch type is also another variable to consider - Figure 2 depicts all the home runs allowed in 2015 to both right-handed and left-handed batters.

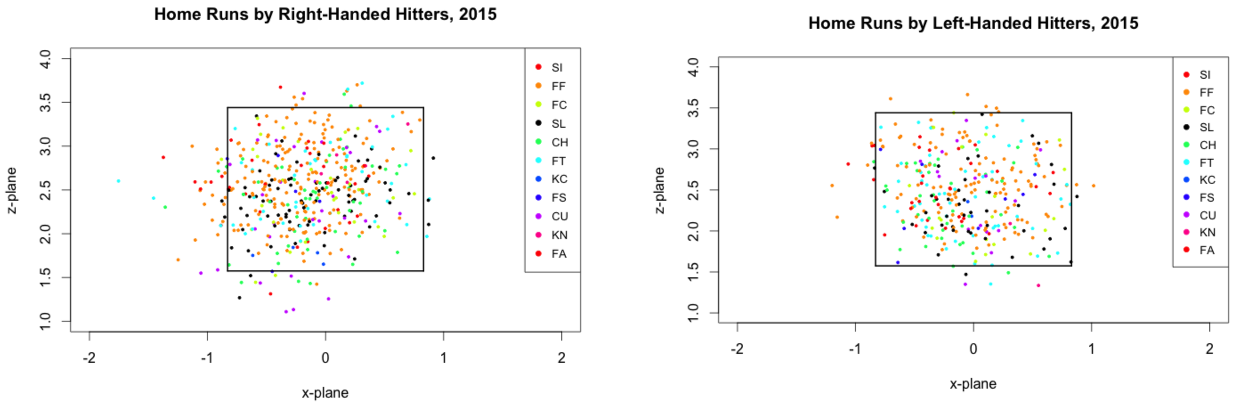

Figure 2: Home Runs in 2015, visualized. The average strike zone is depicted by the black lines, and the home runs are shown from the catcher's perspective.

This figure conveys both location and pitch-type of home runs - one can immediately see that right-handed hitters seem to hit most home runs on the inside part of the plate, while that effect is less pronounced for left-handed hitters. Four-seam fastballs ("FF") appear to dominate the home runs, but it is also probably the case that hitters see more four-seam fastballs than other pitches. Change-ups ("CH") also appear to have a high prevalence, especially for off-speed pitches. While this visualization subset to one of the three years considered, it represents a sizable sample of the total pitches without swamping out the pitch-type information.

3.2 Model Fit

After fitting a logistic regression model with pitch type, stadium, binary of same hand, binary of man on base, binary of top of the strike zone, binary of inside to the batter, and total movement, the coefficients and significance levels were obtained. The table can be found in the Appendix in Section 6.1, Table 1. Model validation is important but somewhat tricky with logistic regression - as a sense check I plotted a histogram of predicted probabilities in Figure 3.

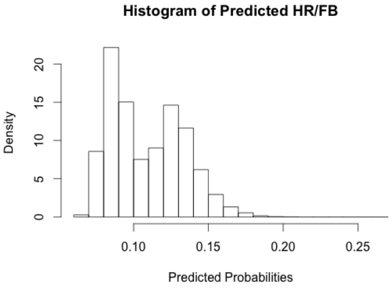

Figure 3

As one can see, these probabilities are around 0.05-0.15 - this is right around the home run to fly ball percentage that one would expect (recall that I subset the data to fly balls, and are trying to estimate the probability of that fly ball going for a home run). A goodness-of-fit test also revealed no evidence against lack of fit - thus, with sensible predicted values I proceed with the analysis of the coefficients.

For pitch types, the only significant coefficients were those for the curveball and knuckle ball - if a fly ball was hit on these pitches, the log odds ratio compared to a changeup dropped by 0.2 and 0.4, respectively (though the standard error is large for the latter). There were quite a few stadiums which significantly affected a fly ball's probability of leaving the park - as expected, Miami and San Francisco (among others) led to a decrease in the log odds ratio, while familiar hitter-friendly parks led to an increase in log odds ratio (Baltimore, Cincinnati, New York, Colorado to name a few). Somewhat surprisingly, Los Angeles and Seattle were both modeled as more hitter-friendly than Arizona - coefficients for both stadiums were positive, and Seattle's is relatively large (0.3). Among those not significant that were notable were Boston, Oakland, and San Diego.

The inclusion of stadiums and pitch types helps account for these variables and condition on them. It is clear that having a man on base (proxy for pitching from the stretch) does not significantly impact the probability of a home run, but all others do. The signs of the coefficients are what one expects for the most part - same hand, velocity, and total movement are negative, and inside is positive. It is interesting that pitches in the top of zone lead to a lower log odds ratio, but the magnitude is among the smallest for significant variables (0.07). The magnitude of coefficient for velocity is also quite small (0.03). The log odds for total movement is -0.15, meaning that holding all other variables constant, a unit increase in movement leads to a multiplication of the odds of a home run by $e^{-0.15} = 0.86$. However, throwing inside increases the log odds ratio by 0.491 (odds ratio increased by a multiplicative factor of 1.63), the largest magnitude of significant coefficients.

The model tells us that the largest effect on the probability of hitting a home run are throwing inside, and possibly the stadium, but it also helps quantify the magnitude of these effects while conditioning on the other effects. While this is nothing groundbreaking, it provides insight on home run tendencies, backed by rigorous statistical methods.

4 A Pitcher's Propensity to Home Runs - Poisson Regression

Poisson regression is a useful tool for modeling the effects of covariates on count data. In this section I will attempt to estimate the effect of various coefficients on the average number of home runs allowed by a pitcher.

4.1 Model Setup

In the last section, I used a frequentist logistic regression to model the probability of a home run - here I will explore a Bayesian approach using Poisson regression. This analysis will not be a comparison of Bayesian versus frequentist methods, but rather an illustration of how a Bayesian model might apply to this data. For a brief overview on Bayesian data analysis, see Appendix Section 6.3

For Poisson-distributed data, the variance is equal to the mean; when the variance is higher than the mean (as is often the case with real data) we say that the data is *overdispersed*. In order to allow for the possibility of overdispersion, I will add addition parameters η_i and λ_i for each pitcher. The full (hierarchical) model is

$$\begin{aligned} Y_i &\sim Poisson(\mu_i * \eta_i) \\ log(\mu_i) &= log(p_i) + \mathbf{X_i}\beta \\ \eta_i &\sim Gamma(\lambda_i, \lambda_i) \\ \lambda_i &\sim Gamma(2, 2) \\ \beta_j &\sim Normal(0, 100) \end{aligned}$$

This induces a marginal distribution that has a mean of μ_i and a variance of $\mu_i(1 + \frac{\mu_i}{\lambda_i})$. I am still most interested in the posterior β values, but the introduction of the η and λ parameters gives the model added flexibility.

4.2 Data and Model Validation

For simplicity, I have used pitchers in 2015 with at least 100 IP - this (somewhat arbitrary) cutoff subsets to starting pitchers, and removes some of the low-inning variance. I also used starters with at least 100 pitches of each of four-seam fastballs, curveballs, and changeups (a crude classification) . The mean home run for this data was around 18 and the variance around 34 - this suggests overdispersion is possible. The velocity, usage percentage, absolute x-movement, and absolute z-movement were calculated for each pitch type and added as covariates to the model. The goal is to see which pitch type and PITCHf/x interactions significantly affect the average home run rate.

To obtain samples from the posterior distribution, I use the program JAGS, which is accessible through R. To validate the model, I use posterior draws to recreate the data - the model is believable if the generated data is indistinguishable from the true data. Figure 4 shows the true distribution of home runs next to two recreated datasets.

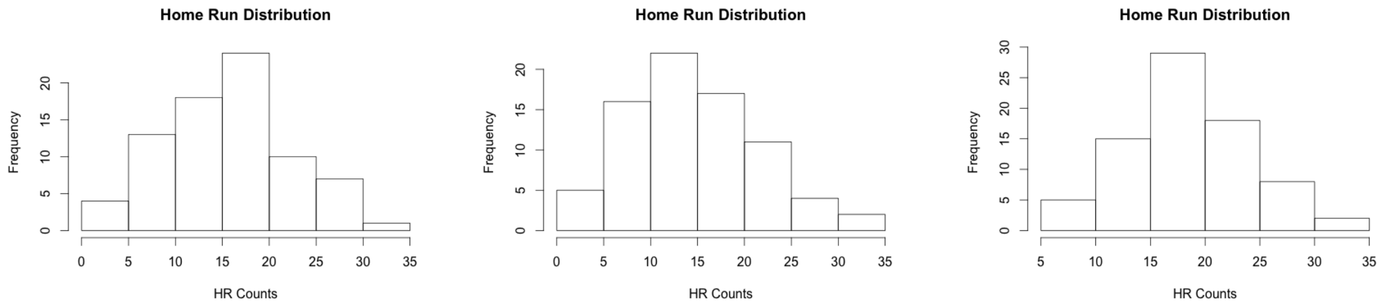

Figure 4: Home Run distributions in 2015, two simulated datasets and one from the true data.

As seen in Figure 4, it is difficult to tell the difference between the simulated and the real data - the true data is on the right. This speaks to the validity of the model, and we proceed with the inference.

4.3 Results

I am most interested in how the coefficients affect home run rates in these pitchers, so I will focus my analysis on the posterior estimates for β. Figure 5 depicts the posterior distributions for the coefficients.

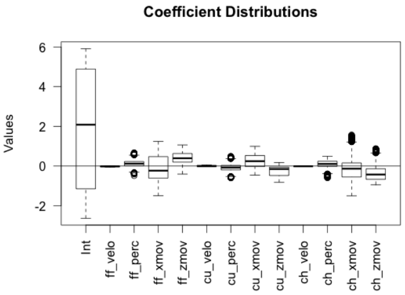

Figure 5

Similarly to frequentist models, we often aim to see when the coefficient values differ significantly from zero - in the Bayesian setting we examine where zero lies in the posterior quantiles. Figure 5 shows that most coefficients have posterior distributions that are close to or center around zero. However, the are some notable tendencies. Somewhat surprisingly, four-seam z-movement and curveball z-movement have positive coefficients, indicating that conditioning on other coefficients these variables are correlated with higher home-run rates. Curveball z-movement and change-up z-movement are conditionally correlated with lower home run rates, more so than other PITCHf/x information.

Though these results are admittedly underwhelming, they do point to an interesting hypothesis - home runs are statistically difficult to predict given PITCHf/x information. We saw in the previous section that the stadium is quite important to predicting home run to fly ball ratios, but including home ballparks as a fixed effect would be an inadequate proxy (as half of the teams' games are played on the road). Including percentage of pitches inside and percentage of pitches in the top of the zone lead to models with worse fits.

5 Conclusion

In Section 3 I estimated the probability of hitting a home run given a fly ball, using logistic regression. With no evidence of lack of fit and reasonable predicted probabilities, I found that the largest effects on HR/FB probabilities come from stadium effects and whether or not the pitch was inside. Though these results are intuitive, I was able to estimate the magnitude of these effects in a statistically principled way. In a different approach to estimating the effect of quality of pitches, I applied a Bayesian hierarchical model in Poisson regression to attempt to estimate average home run rate solely from PITCHf/x and pitch type interaction data. A preliminary model for fastball/curveball/changeup starting pitchers indicated that while there are intuitive trends, PITCHf/x data alone does not predict home run rates with much power.

It is possible that there are other types of pitchers for whom pitch movement, velocity, and percent usage affect their proneness to home runs, though a more extensive, comparative analysis is required.There are also other PITCHf/x variables that could possibly be constructed - a metric for pitch command, some interaction between count and pitch type, a measure of variance of pitch quality, etc. Another extension could be to consider a fixed effect for stadium, though low counts would probably require a zero-inflated Poisson model rather than (or perhaps in addition to) an overdispersion model.

These results suggest that the reasons that pitchers give up home runs lie more on a pitch-by-pitch basis rather than an average level, though they do not rule out the possibility that there are pitcher-level qualities that lead to elevated home run rates.

6 Appendix

6.1 Logistic Regression Coefficients

	Estimate	Std. Error	Significance
(Int)	0.587	0.328	0.073
CU	**-0.252**	**0.060**	**0.000**
EP	-0.498	0.444	0.262
FA	0.446	0.296	0.132
FC	-0.039	0.060	0.518
FF	-0.060	0.051	0.236
FO	-0.090	0.477	0.851
FS	0.140	0.094	0.137
FT	-0.044	0.056	0.430
KC	-0.193	0.104	0.064
KN	**-0.475**	**0.154**	**0.002**
SC	0.177	1.102	0.873
SI	-0.058	0.059	0.323
SL	-0.063	0.045	0.164
ATL	-0.007	0.090	0.939
BAL	**0.360**	**0.085**	**0.000**
BOS	0.036	0.088	0.687
CHC	**0.285**	**0.087**	**0.001**
CIN	**0.321**	**0.091**	**0.000**
CLE	0.101	0.090	0.262
COL	**0.335**	**0.089**	**0.000**
CWS	**0.187**	**0.088**	**0.034**
DET	0.056	0.087	0.520
HOU	0.301	0.086	0.000
KC	**-0.237**	**0.091**	**0.009**
LAA	0.039	0.091	0.668
LAD	**0.217**	**0.089**	**0.015**
MIA	**-0.449**	**0.105**	**0.000**
MIL	**0.462**	**0.086**	**0.000**
MIN	-0.089	0.090	0.323
NYM	-0.041	0.089	0.649
NYY	**0.290**	**0.085**	**0.001**
OAK	-0.167	0.093	0.073
PHI	0.126	0.091	0.166
PIT	0.117	0.091	0.200
SD	0.085	0.092	0.353
SEA	**0.304**	**0.088**	**0.001**
SF	**-0.301**	**0.097**	**0.002**
STL	**-0.243**	**0.096**	**0.012**
TB	-0.044	0.090	0.627
TEX	-0.021	0.092	0.819
TOR	**0.302**	**0.086**	**0.000**
WSH	-0.065	0.096	0.499
same_hand	**-0.060**	**0.023**	**0.009**
man_on	-0.043	0.022	0.055
velo	**-0.026**	**0.004**	**0.000**
tot_mov	**-0.153**	**0.022**	**0.000**
top_zone	**-0.073**	**0.027**	**0.006**
inside	**0.491**	**0.022**	**0.000**

Table 1: Table of coefficients for Logistic Regression results. Rows that are bold are for coefficients that are significant at the 0.05 level

6.2 Logistic Regression Background

The observations are individual pitches, with the outcome being a one or zero depending one whether or not a home run was allowed. In Ordinary Least Squares, we treat our outcome Y as coming from a *Normal* (i.e. Gaussian) distribution with mean μ; we let $\mu = x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p = \mathbf{X}\beta$ and we try to say whether each $\beta_1 \ldots \beta_p$ differ significantly from 0. A Normal distribution is inappropriate here because our outcome is a binary 1 or 0, rather than any number from $-\infty$ to ∞. Instead we say each observation Y (home run) follows a *Bernoulli* distribution, with mean π that is exactly the probability of getting a 1. Similar to Ordinary Least Squares, we model the mean with $\mathbf{X}\beta$, but we cannot do it directly because π is restricted to $(0, 1)$. Thus, we make use of a *link function* - logistic regression entails using the logit function $logit(p) = log\left(\frac{p}{1-p}\right)$. This function is useful because it is monotone increasing and has support on the entire real line. Thus, we set $logit(\pi) = \mathbf{X}\beta$ and proceed with our inference.

The logit link function (i.e. logistic regression) is preferred over other link functions for binary data largely because of the interpretability of the covariates. Note that the odds of event E happening is $\frac{P(E)}{1-P(E)}$ - it works out that in logistic regression if the coefficient for column x_j is β_j, then with a unit increase in x_j the odds of a home run being hit is multiplied by e^{β_j} (equivalently, a unit increase in x_j corresponds to an increase in the *log odds* by β_j). R's basic output is in log odds rather than odds - if the log odds are close to 0 (or the odds are close to 1) then the variable does not have a large effect on the outcome. If x_j is a category (like ballpark) then β_j is the ratio of the log odds of x_j to that of x_1 (or whatever variable is the reference level) - this is known as the *log odds ratio.*

6.3 Overview on Bayesian Data Analysis

In contrast to frequentist methods, which use only information given from the data, Bayesian methods use prior information about parameters when making inference. Even when little or nothing is known about the data, Bayesian methods have been shown to have good predictive properties, especially when there is less data.

The first step in a Bayesian analysis is model specification. The Poisson distribution is a natural choice for count data, and requires one parameter (the average count) to explain the distribution. Similar to logistic regression, Poisson regression requires a link function to connect the mean to $\mathbf{X}\beta$. The canonical choice for a link is the log function. Let Y be the number of home runs allowed, and p_i be the number of pitches thrown. p_i is an offset that corrects for different amount of pitches thrown.

$$Y_i \sim Poisson(\mu_i)$$
$$log(\mu_i) = log(p_i) + \mathbf{X_i}\beta$$

This specifies the probability distribution of Y given β and $\mathbf{X}$, which we denote with $p(Y|\mathbf{X}, \beta)$. Unlike frequentist methods, in Bayesian models we assume that each β_j is also random (rather than fixed but unkown). Thus, we specify a *prior* distribution for β. In order to not over-specify the model and to let the data guide the inference, we set the variance of the prior distribution to be large (here 100).

$$\beta_j \sim Normal(0, 100)$$

This gives us the marginal probability of β, denoted $p(\beta)$. Our inference will come from the *posterior* distribution of β, which is the distribution of β given Y and $\mathbf{X}$, or $p(\beta|Y, \mathbf{X})$. In most most models, $p(\beta|Y, \mathbf{X})$ is not available analytically, so we need to make use of Markov Chain Monte Carlo, which allows us to sample independently from the posterior distribution of the parameters.