Final Report

Maya Doitch

Ruben Jimenez

Alea Seifert

December 13, 2024

Introduction

Baseball is a captivating sport that is a big part of American culture. Besides its ability to keep many entertained, baseball is also rich in data with many opportunities to conduct statistical analyses. Each play contains information that can be interpreted and used within a multilevel data modeling structure. Our motivation for choosing this data set was our interest in the sport, as well as the analytical opportunities and our curiosity in uncovering unique trends that influence the trajectory of the game.

One factor of baseball games we found intriguing was variability in hit distance. We were interested in exploring this further as well as the potential variables that influences a batter's hit distance. We have seen prior research focusing on factors like bat speed and bat velocity, but have seen fewer studies examining the impact of multilevel factors. To explore baseball games in a multilevel manner, we examined the batter's dominant arm as our level one variable and the specific attributes of a pitch such as angle, speed, and type as our level 2 variables. We also accounted for player to player variability by including batter as a random effect in our model. We hypothesized that individual differences in skill can yield significantly different outcomes.

Ultimately, our goal was to determine whether these factors, random effects, and interactions attribute to the outcome of a home run, or a ball that is hit a far distance. We hypothesize that pitch angle and speed will significantly impact how far a ball travels. By examining this we hope to contribute to the field of baseball analytics and gain valuable insights to inform coaching and player development.

Data and Methods

The data comes from Baseball Savant which records various aspects of plate appearances in MLB games. The data recorded games over the 2024 regular

season Major League Baseball (MLB), from the beginning of April through June. This documentation was available at **Baseball Savant** and can be downloaded at **Share Point.** The data was collected from a Statcast system which includes various cameras that are dedicated to tracking quantitative data from plate appearances with respect to both the pitcher and the batter in addition to data on the game states. In the dataset, there are over 90 variables about all aspects of an MLB game, but through our analysis with our goal of using variables that are directly observable from a batter's role, we reduced our model to include the launch angle of the ball off the swing, the launch speed of the ball off the swing, the batter's preferred stance side, and whether a pitch was a type fastball or not. These variables are displayed in Figure 1.

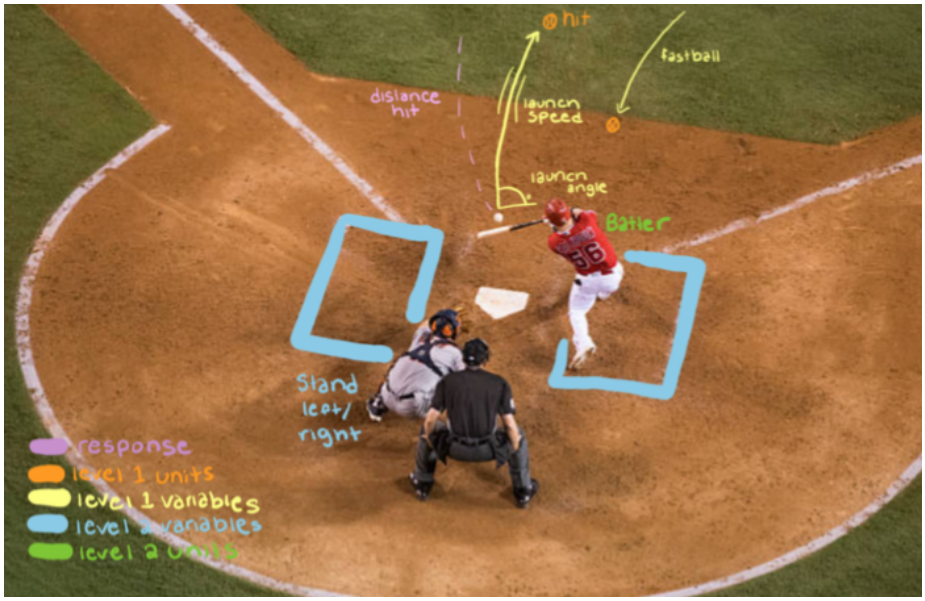


Figure 1: This figure illustrates the various variables of interest and how they are observed on the playing field. The differing colors depict what level the variables are measured at.

Initially, the data set contained 346,250 observations. However, we filtered the data to include only observations of balls hit into play, removing instances where the hit was categorized as a strike or foul. Additionally, we excluded hits with a recorded distance of 0 feet since this indicates that the ball did not travel any meaningful distance. Hits with launch angles less than 0 were also removed, as these represent balls hit directly into the ground, which would not result in meaningful travel distance. Including these observations produced non-linearity into the data, which could negatively impact the ability to meaningfully interpret the model.

We also grand-mean centered all quantitative variables were in the final model.

While the original data set included a variable for pitch type, analysis revealed that only fastball-related categories (four-seam, two-seam, and cutter) were statistically significant after dummy coding. Consequently, we encoded these categories into a single indicator variable, is_fastball, where a value of 1 indicates a fastball and 0 indicates all other pitch types. Additionally, we encoded batter stance side to indicator coding. Our final data set includes 41,610 hit observations over the variables of interest seen in table 1.

Table 1: Variable Descriptions within the final dataset. Includes information on how the variable is measures, the data type, and the role within the multilevel structure.

Name	Variable Role	Type	Values (units)
Hit distance	Response (by hit)	quantitative	Feet
Fastball	Level 1 predictor	categorical	Yes (1) / No (0)
Launch angle	Level 1 predictor	quantitative	Degrees
Launch speed	Level 1 predictor	quantitative	Mph (Mile per Hour)
Stand	Level 2 predictor	categorical	Right (1) / Left (0)
Batter	Level 2 variable	Character, a Unique ID	IDs for each Batter

In our analysis, we fit a total of 11 models, starting with a null model in which batters were treated as random intercepts at level 2. Initially, this approach assumed that variation in hit distance could be explained by differences between batters. As a result, we proceeded with a multilevel model, which accounts for the hierarchical structure of the data, with hits nested within each batter which can be seen in figure 2.

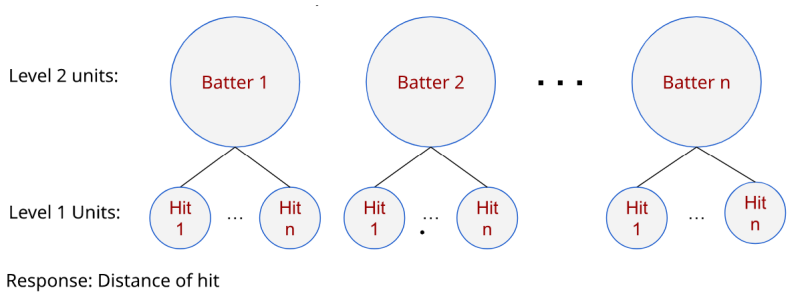


Figure 2: This image depicts the multilevel structure of the data with hits nested within each batter

Results

In our initial analysis, we viewed various terms and found that launch angle and launch speed appeared to have some relationship with hit distance. In figure

3 it appears that there is a weak to moderate negative relationship between launch angle and hit distance (corr = -0.307). The matrix plot also shows a moderate positive relationship between launch speed and hit distance (corr = 0.444). Additionally, the plot showed a possible non linear relationship between hit distance and launch angle, so we decided to include a squared term in our later model.

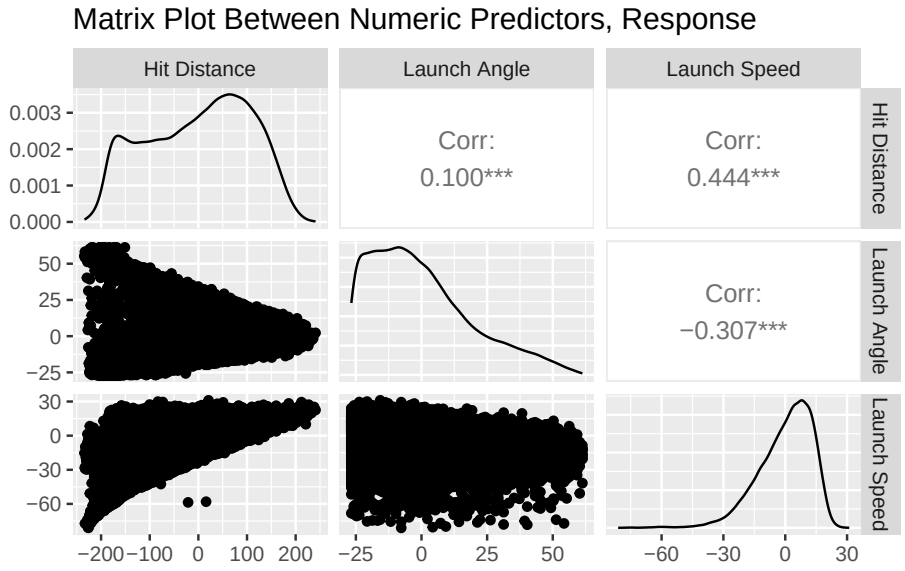


Figure 3: Matrix plot of the quantitative features in the data including launch angle and launch speed

We also viewed relationships between categorical variables including whether the pitch was a fastball or not and which side the batter is standing and hit distance. We see that the average hit distance is greater for fastball than non-fastballs. It also shows that the average distance hit is larger for a hit where the batter stands on the right side.

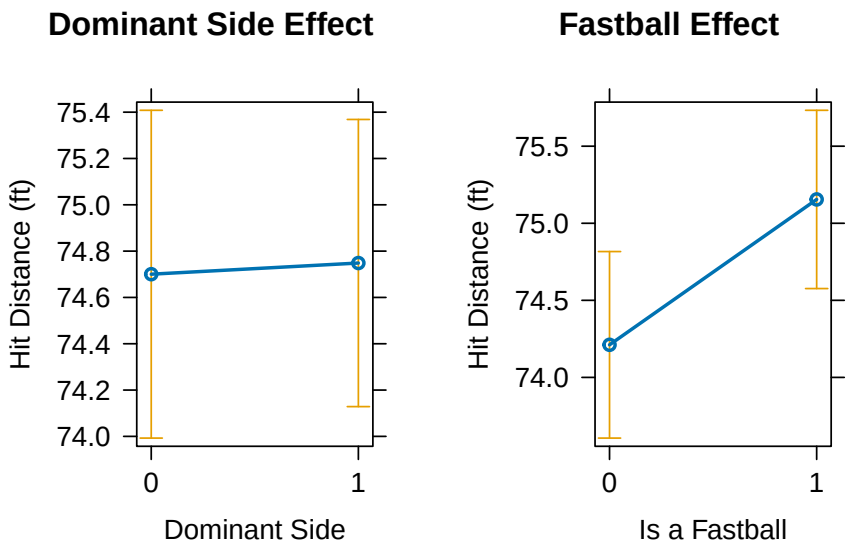


Figure 4: Effect plots of batter standing position and if the pitch was a fastball or not

The next step, was building the model. We first ran an anova to find the significance of the level 2 grouping variable batter. We found that batter is a statistically significant predictor hit distance (df numerator = 551, df denominator = 41058, F-statistic = 1.7012, p-value = <0.001).

Figure 5 depicts a random sample of the thousands of batters in the dataset to observe the within batter and between batter variability. The box plots appears to be fairly similarly distributed within each batter. There also does appear to be some differences in mean hit distance between the batters.

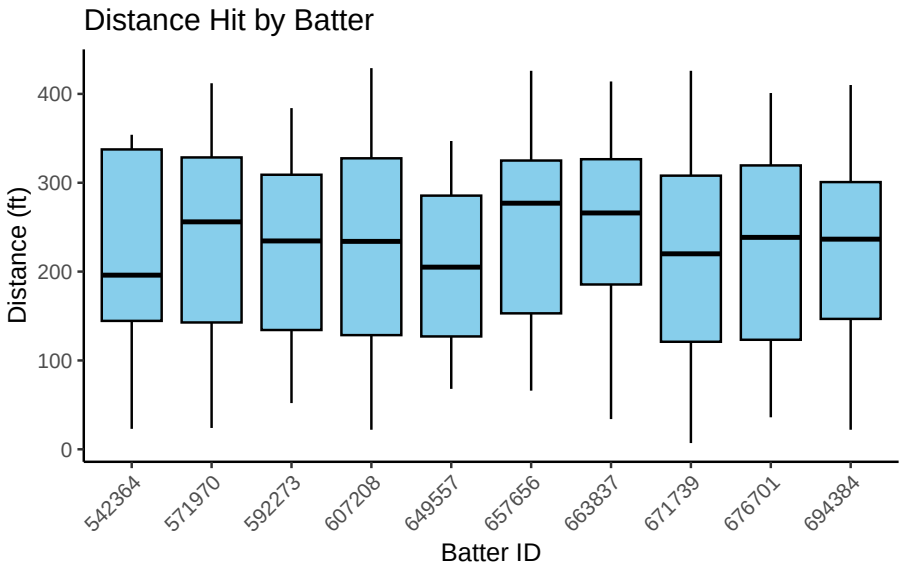


Figure 5: A random sample of 10 Batters in the dataset to observe difference in distance of hit variability

We then created a null model where batter is a random intercept.

$$\tau_0^2 :$$

The batter to batter variance in average hit distance is 102.8.

$$\sigma^2 :$$

The variance in average hit distance for each batter is 10929.5.

$$\beta_0 :$$

The average hit distance across all batters is 234.5283.

$$ICC : \frac{102.8}{102.8 + 10929.5} = 0.00932$$

The correlation between two hits by the same batter is .009. Only 0.00932% of the variation in hit distance is explained by between batter variation rather than within batters. This is not substantial. The deviance is 104.544 feet. The null model also has an AIC and BIC value of 505300. We will still proceed the analysis process with batters as random slopes.

For model selection, we chose the model that had the highest conditional R^2 value while still being easily interpretable and not overfit to the data, as the goal is to determine what the strongest predictors of large hit distances is and to be able to generalize it across players in the MLB.

We are 95% confident that variation in the random launch angle slopes is between 9.418 and 15.366.

Fixed Effects Interaction

Within the fixed effects we found that there was a significant interaction between squared launch angle and launch speed.

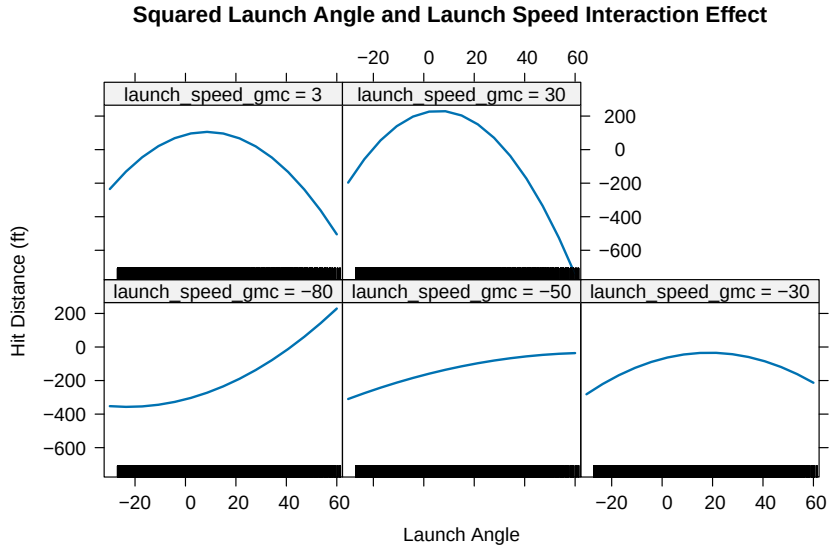


Figure 6: The launch angle and launch speed interaction over various values of launch speed

Random Effects

Additionally, we found a that including launch angle as a random slope significantly improved the model.

Model11 unexplained level 1 variation (σ^2): 911.0484

Null model unexplained level 1 variation: 17985.5

(17985.5 - 911.0484)/17985.5 = .95

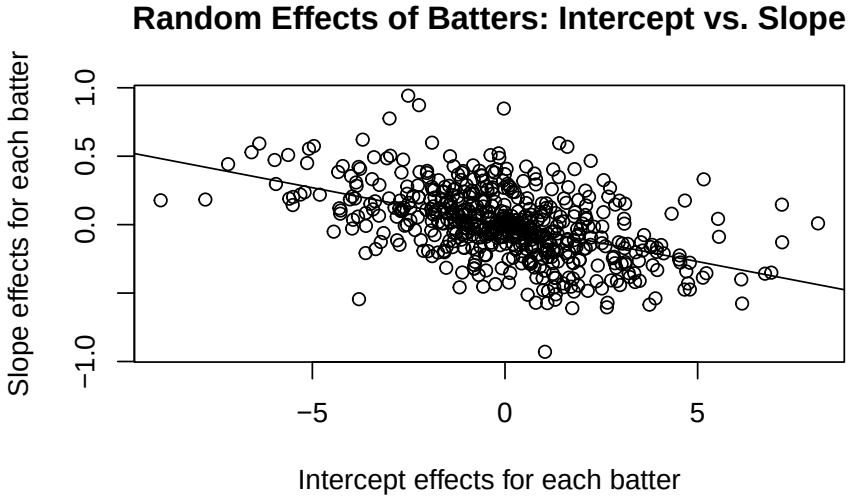
95% of unexplained variation at level 1 has now been explained.

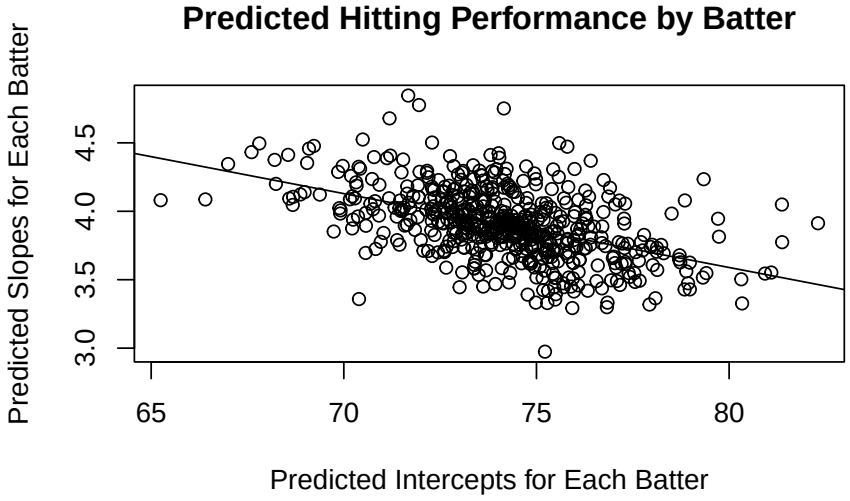
Model11 unexplained level 2 variation (τ_0^2/intercept variation): 12.2167

Null model unexplained level 2 variation: 292.9

(292.9 - 12.2167)/292.9 = .96

→ 96% of unexplained variation at level 2 has now been explained.





Final Model

Our final model was given by

Table 2: Strength of model fit metrics

AIC	AICc	BIC	R-squared conditional	R-squared marginal	ICC	RMSE	Sigma
402715.3	402715.3	402810.3	0.918	0.913	0.053	29.973	30.184

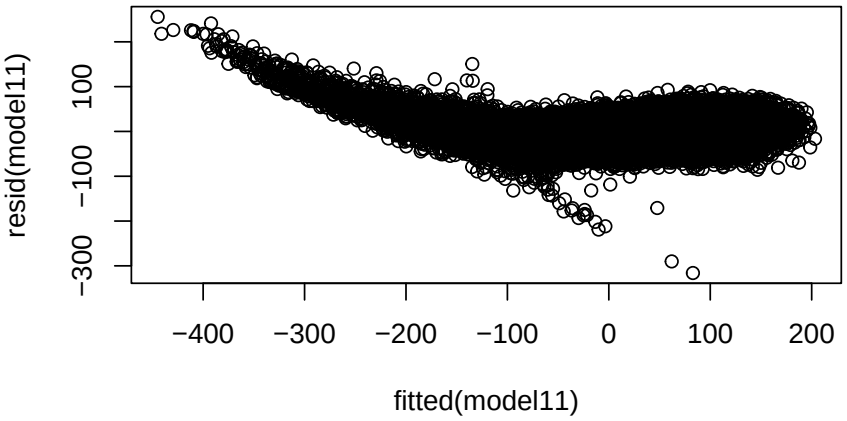
$$\begin{aligned}
\text{hit } \widehat{\text{distance}}_{ij} &= \beta_{0i} + \beta_{1i}(\text{launch angle}_{ij}) + \beta_{2i}(\text{launch angle}_{ij}^2) && (1)\\
&+ \beta_3(\text{launch speed}_{ij}) + \beta_4(\text{stand}_{ij}) && (2)\\
&+ \beta_5(\text{fastball}_{ij}) + \beta_6(\text{launch angle} \cdot \text{launch speed})_{ij} && (3)\\
&+ \epsilon_{ij} && (4)
\end{aligned}$$

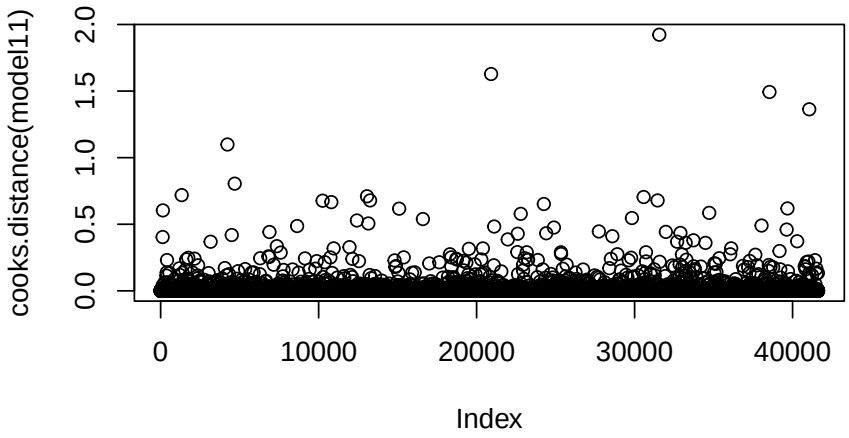
Interpretation of all the variables

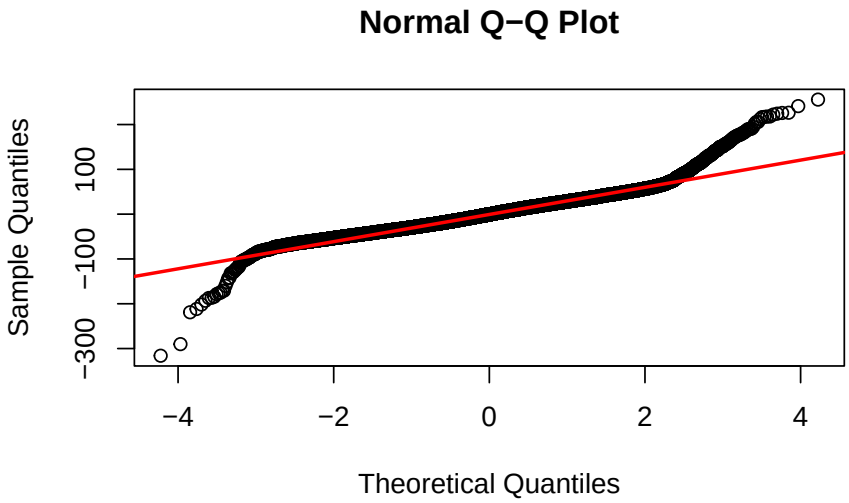
1. Fixed Intercept: When launch angle, launch angle squared, launch speed are at their average, and ball pitch is not a fastball with the batter standing on the left side, the predicted average hit distance is 309.7 feet.
2. Launch Angle Fixed Effect: A one degree increase in launch angle increases the predicted hit distance by 3.86 feet, with launch angle squared, launch speed, batter stand side, and fastball satus fixed.
3. Squared Launch Angle Fixed Effect: The effect of increasing launch angle decreases at higher values hit distance.
4. A one mile per hour increase in launch speed increases the predicted hit distance by 4.825 feet, with launch angle, launch angle squared, launch speed, batter stand side, and fastball status fixed.
5. The predicted hit distance for a batter standing on the right side is 0.7278 above the average hit distance, for a fixed launch angle, launch angle squared, launch speed, and fastball status.
6. The predicted hit distance for a fastball is 0.917 above the average hit distance, for a fixed launch angle, launch angle squared, launch speed, and batter standing side.
7. For each additional mile per hour increase in launch speed, the effect of the quadratic launch angle on hit distance decreases by 0.00384, for batter standing side and fastball status fixed.
8. Random Intercept: The batter to batter variation in the average hit distance is 3.517 feet.

9. The effect of launch angle on hit distance between batters has a variance of 1.056 feet^2.
10. The effect of launch angle squared on hit distance between batters has a varinace of 0.00254 feet^2.
11. The variation in the average distance hit among hits within the same batter is 878.5 feet^2.
12. (0.66 * 1.02778 * 1.87542 = 1.272163) For batters with higher average hit distance, the effect of the launch angle on hit distance tends to be more positive.
13. (-0.88 * 0.01595 * 1.87542 = -0.0263234) For batters with higher average hit distance, the effect of the launch angle squared on hit distance tends to be more negative.
14. (-0.63 * 0.01595 * 1.02778 = -0.01032765) Within batters where the launch angle squared has a more positive effect on hit distance, the effect of having a higher launch on hit distance tends to be more negative.

Model Assumptions:








```
# A tibble: 5 x 16
  type  hit_distance_sc pitch_type dist_from_cen launch_angle launch_speed
  <fct>           <dbl> <fct>              <dbl>        <dbl>        <dbl>
1 X                  49 FF                 3.15            86         89.1
2 X                  40 FC                 2.83            84         94
3 X                  34 FF                 3.28            85         91.9
4 X                   2 SL                 0.843           82         13.3
5 X                  67 ST                 3.11            89         79
```

```
# i 10 more variables: game_type <fct>, batter <fct>, stand <fct>,
#   p_throws <fct>, dist_from_cen_gmc <dbl>, hit_distance_gmc <dbl>,
#   launch_angle_gmc <dbl>, launch_speed_gmc <dbl>, is_fastball <fct>,
#   standR <fct>
```

These observations we found to be outliers and potentially influential.

To assess our final model, model 11, we made three graphs. The residual vs fitted plot tested the model's equal variance and linearity condition, the QQ plot tested the model's normality condition, and the Cook's distance plot tested the model's influential observations.

After examining the QQ plot, we concluded that the normality condition is okay. The majority of the residuals sit along the diagonal line, indicating that the bulk of the data is normal. However, deviations at the tails indicate that the residuals may not be perfectly distributed. This could be attributed to the large dataset we have or outliers. We will consider this assumption met, even though we may decide to still proceed with caution.

Next, we observed the residual vs fitted plot to assess the equal variance and linearity condition. The plot reveals a curved pattern, suggesting non-linearity. Due to this, we must proceed with caution as we cannot say that our linearity condition has been met. The funneling or heteroscedasticity, where the residuals get more spread out as the fitted values get larger indicate that there is unequal variance. Unfortunately, due to this we decided to proceed with caution.

Lastly, we observed the Cook's distance plot to test whether there are influential observations in the data. There are five points that appear to have Cook's distance value greater than one, indicating potentially influential points. The observations that correspond to these points are 4229, 20918, 31564, 38538, and 41055. These points have high leverage, large residuals, or both- making them influential.

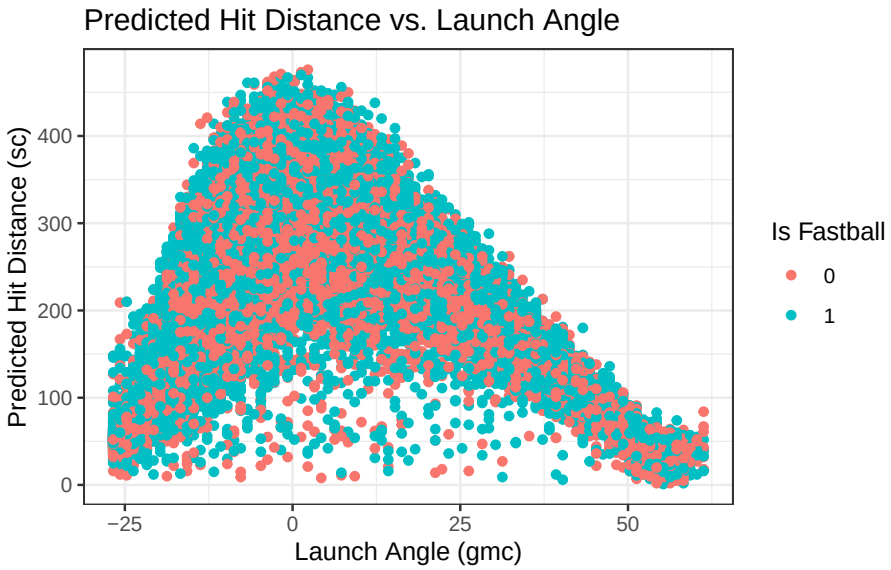
Discussion

Using a multilevel linear regression model fit with considerations toward launch angle, launch speed, batter stance, pitcher throwing arm, and whether a pitch was a fastball or not, where batters were considered random units sampled from a larger population of batters, we were able to explain 92% of variation in hit distances. Our analyses found that the strongest predictor of hit distance for MLB hitters is the launch speed of the ball from contact with the bat. This is followed closely by the vertical launch angle of the ball from contact with the bat, while the next strongest predictors, batter's stance and whether the pitch type is a fastball or not, while statistically significant, were not practically significant. Using random intercepts for batters is a strong point in our analysis, as we are able to extrapolate the observed effects to a larger population of batters in the MLB, including present players and future batters who may not be in the league yet.

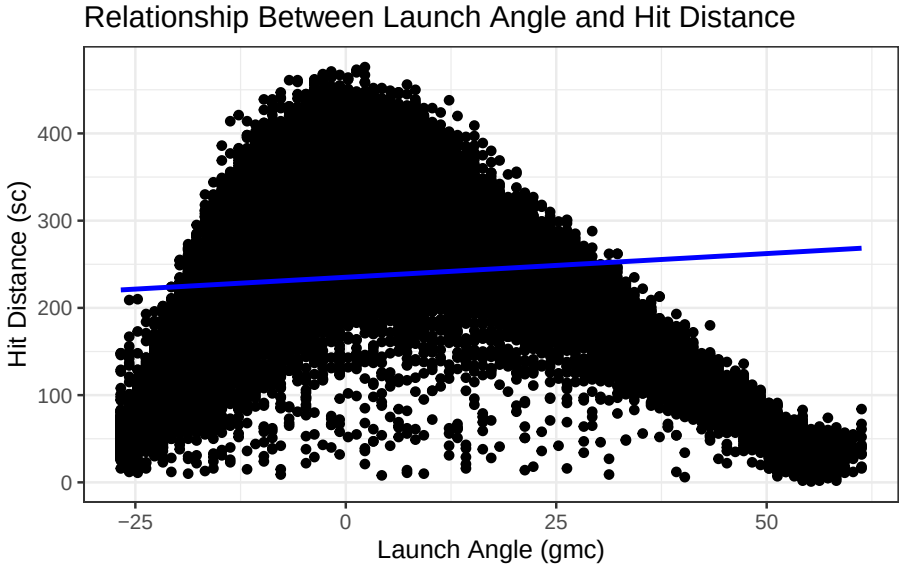
Our scope of inference is limited to hits above a zero degree vertical launch angle, as we did no analysis on data below that, so the results do not extrapolate there. Further research into ball behavior when balls are launched toward the ground may need to be explored in a similar isolated study.

Appendix

Tables and Figures Not Included



Annotation: This scatterplot shows the relationship between launch angle and hit distance, with points colored by whether the pitch was a fastball. The data reveals a parabolic trend, where hit distance increases with launch angle up to peak (around 20-30 degrees) and decreases after that. Fastballs (in teal) and non-fastballs (in red) appear to have similar distributions, which may mean that pitch type has a minimal impact on the relationship between launch angle and hit distance.



Annotation: This plot visualizes the relationship between launch angle and hit distance using scatter points and a fitted linear trend line. The trend line, added highlights the general direction of the relationship without showing confidence intervals.

Statistical Modeling

To figure out the best model we did some hypothesis testing along the way. All of our variables, aside from standR (the batter's stance), had a t-value that was larger than 2. This confirmed that all but one of our fixed explanatory variables explained significant variation in hit distance. This supports the inclusion of these variables in our final model. Due to standR's small t-value we decided to fit a model without stand to see if it was valuable to keep standR in our final model. We ran an Anova test comparing our final model, Model 11, to a model without standR. Although the likelihood ratio test indicates that adding standR to the model does not significantly improve the model fit (p=0.9146), we decided to retain it for several reasons. First, including standR does not worsen the model, as we saw by the identical AIC and BIC values between the two models. Second, standR is conceptually important as it captures potentially meaningful differences in performance between left- and right-handed batters, which we found important to use as a Level 2 variable. By retaining standR, the model remains robust and interpretable for contexts where stance might play a role in understanding hit distance.