# Mile High Pitching: An Analysis of the Relationship Between Altitude and Pitching Success

Jack Karisch and Hadyn Craft

April 29, 2016

## 1 Abstract

There is a belief in baseball that player performance, especially for pitchers, is affected by altitude. This belief shapes the way teams and players prepare for games at high elevation. It springs from the fact that the thinner air found at higher elevations acts differently on the ball than the denser air at sea level. This project seeks to explore these altered characteristics' effects on pitcher performance at high elevation, primarily using logistic regression.The primary finding is that altitude does have a significant effect on a pitcher's performance, and that performance at elevation can be predicted at an accuracy of about 63% to 69%.

## 2 Introduction

The physical effects of altitude on the trajectory of baseballs are documented and accepted as fact. The ball experiences less drag at higher elevations because of the thin air. This causes the ball to travel faster and further. By the same token, the ball also has less to "grab onto" when moving through the air, so rotation has a lower effect on its trajectory.[1] These effects impact pitcher performance directly in two ways. Speed is perhaps the most important single attribute a Major League Baseball (MLB) pitcher can have. A fast pitch reaches the plate in less time than a slow one, giving a batter less time to react. The slight boost in speed from lower drag should therefore help pitchers. However, another important weapon in a pitcher's arsenal is the amount of break (deviation from a straight line) he can put on his pitches, and the additional speed granted by reduced drag is offset by less break due to spin. This analysis examines the effect of the thinner air on pitching performance by attempting to answer two key questions, which are:

1. Is there a significant difference in performance at high altitude and low altitude?

2. If so, can we predict a pitcher's performance at high altitude using statistical models?

Pitchers were selected as the position of interest because altitude is believed to affect them more strongly and because the format of the data lent itself to pitch analysis. It is worth noting, however, that there is an inverse relationship between pitcher performance and batter performance, so some conclusions about batting can be drawn from an analysis of pitching.

The pitcher has a few options when choosing how to throw a particular pitch. He can throw it in such a way that it travels to the plate as quickly as possible. This is called a fastball. The pitcher can also alter his grip and release to put spin on the ball, leading to some amount of break in a direction of his choosing. Curveballs, sinkers and sliders generally fall in this category. Curveballs are the slowest and break the most, sliders are between curveballs and fastballs in terms of speed and break, and sinkers are almost as fast as fastballs and break relatively little. Finally, the pitcher can pretend that he is throwing a fastball, but throw the ball much slower than expected, hoping to trick the batter into swinging early. This is called a changeup. In the National League during 2013 and 2014, 54% of pitches thrown were fastballs, 15% were sliders, 11% were sinkers, 10% were curveballs, and 10% were changeups.

For this project, any pitch thrown at the Colorado Rockies' Coors Field in Denver, Colorado was considered thrown at high altitude, because theirs is the only MLB ballpark that is located at a significantly higher altitude than the rest (5,211 ft., with the next highest, the Arizona Diamondbacks' Chase Field, at 1,059 ft. All the other stadiums are near or under 1,000 ft.)[2].

# 3 Data Collection

The data for this project came from multiple sources. Most of it is from the PitchFx database, maintained by Sportvision with the cooperation of the MLB. The full dataset contains comprehensive trajectory information for almost every MLB pitch since around 2008. It includes, among others, the following variables: name and ID number of the pitcher and batter, the handedness of the pitcher and batter, the simple result of the pitch, the start and end speed, the strike zone of the batter, the position within the strike zone when the pitch crossed the plate, the characteristics of the break, and the type of pitch thrown. In all, there are 81 different variables in the PitchFx data as it was used. The project uses PitchFx data from all the games that the Rockies participated in during the 2013 and 2014 regular seasons (the Rockies did not make the post season either year).

The PitchFx data was merged with data from Sean Lahman's Baseball Database, which includes yearly statistics for every team and player in the MLB. Lahman's yearly pitcher statistics for 2013 and 2014 were merged with the PitchFx data by pitcher ID number, primarily to add earned run average (ERA) to the dataset. After merging and cleaning, the total number of distinct pitches thrown in games involving the Rockies in 2013 and 2014 was 69,218.

To this hybrid dataset a range of additional variables was added, each calculated from different PitchFx values to be more useful. Examples of these additional variables include count, vertical and horizontal distance from the center of the strike zone, and a variable measuring the angle of break specific to the handedness of the batter. After adding these metrics, the dataset contained 96 variables in total.

# 4 Study Design and Methodology

The first step in measuring pitcher performance was to create a binary variable for which a good outcome of a given pitch counts as a success, while a neutral or bad outcome counts as a failure. To that end, a variable was made from the pitch-by-pitch data that counted as a success any pitch that resulted in a strike (e.g. swinging strike, called strike, or a foul when there are not two strikes in the count) or an out, and anything that resulted in a ball, hit by pitch or hit in play with runs scored or no outs as a failure.

During this project, logistic regression was employed often. This operates much like multiple linear regression, except the dependent variable is a binary factor. The model predicts the log of the odds that a given observation was a success:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

for $p$ = probability of successful outcome, $\beta_0, ..., \beta_n$ = model coefficients, $X_1, ..., X_n$ = dependent variables, and $\epsilon$ = logistically distributed error term. Conversion to an easily interpretable probability is straightforward:

$$p = \frac{1}{1 + \exp{(\hat{\beta_0} + \hat{\beta_1} X_1 + ... + \hat{\beta_n} X_n)}}$$

for $\hat{\beta_0}, ..., \hat{\beta_n}$ = fitted coefficients.

## 4.1 Determining Difference in Performance at High and Low Elevation

To evaluate whether there is a significant difference between pitcher performance at high altitude and low altitude, the data was first subset into only games the Rockies played in, to obtain roughly equally sized subsets of high and low elevation pitches. Any pitch thrown by a Rockies pitcher was then removed to keep them from dominating the results, as they took up roughly half of the dataset. After subsetting and removal the dataset size was 45,243 observations. Using this data, the outcome variable described above (hereafter called "positive outcome") was used as the dependent variable of a logistic regression with a binary variable indicating whether the pitch was thrown at Coors field. This model was run with and without pitcher ERA, to control for pitcher skill: $y = \beta_0 + \beta_1 X_1 + \epsilon$ and $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, for $y$ = positive outcome, $\beta_0$ = intercept term, $X_1$ = Coors, $X_2$ = ERA, and $\epsilon$ = error term.

### 4.2 Predicting Performance at High Elevation

In order to build a predictive model for pitcher performance at high elevation, a simple subset of all pitches thrown at Coors field was created. Throughout, positive outcome was used as the dependent variable. Because positive outcome was a binary variable, logistic regression was employed. The final set of independent variables included in the full model totaled 22. See section 8.1 (Appendix) for the full list of variables.

The data was then subset by pitch type, between fastballs, curveballs, sinkers, sliders and changeups, and the same model, excluding the pitch type variable, was run on each of these subsets. This was done because of the range of different attributes that make each pitch type successful. For example, while speed is very important for a fastball, it might have much lower impact on a curveball. Separating the types allows the researcher to glean more information from the model, and hopefully make more accurate predictions.

In an attempt to improve on the full models, backward and forward stepwise variable selection was performed on each of them. Backward selection starts with the full model and iteratively evaluates whether the removal of predictor variables substantially hurts the fit of the model. It uses a metric called Akaike information criterion (AIC) to compare successive options; the model with the lower AIC is preferred. Forward selection starts with an empty model and iteratively evaluates the effect of adding variables using AIC.

Because there were some substantial correlations in the predictor variables, ridge regression was applied to the full models. This technique includes a penalty term (the trailing term in the equation below) in the least squares calculation which shrinks the coefficients as $\lambda$ increases. Ridge regression is the minimization of:

$$\sum_i (y_i - \beta_0 + \beta_1 X_{1i} + ... + \beta_n X_{ni}) + \lambda \sum_j \beta_j^2$$

for the observations $i$ and coefficients $j$. This decreases the mean squared error (MSE) of the model, at the cost of added bias in the coefficients. The decrease in MSE could be worth it if the multicollinearity of the predictors were substantially increasing error. The optimal value for $\lambda$ was found using 10-fold cross-validation.

## 5 Results

### 5.1 Determining Difference in Performance at High and Low Elevation

Both the first model, which only contained the binary variable indicating whether or not the pitch was thrown at Coors Field, and the second model, which controlled for pitcher skill using ERA, resulted in a p-value of ˜0 for the Coors variable. Using the transformation shown above, the probability of

success for pitches thrown at Coors field was found to be 55.2%, while the probability of success for pitches thrown at all other fields was found to be 58.7%.

### 5.2 Predicting Performance at High Elevation

The main results from the different predictive models are shown in Table 1. The full model, forward selection and backward selection show the McFadden pseudo $R^2$ value for these models. This value can be used to compare the fit of two models, although it is not calculated in the same way as conventional $R^2$ in linear regression. Instead, it shows how much more variability the full model explains than the null model. A higher value means the full model explains more variability. The forward and backward selection sections of the table also feature the number of variables remaining in the model after selection. The actual list of variables for each pitch type and model can be found in Section 8.2 (Appendix). For convenience, Table 2 shows the most commonly occurring variables in the forward and backward stepwise models. The Ridge section of Table 1 displays the $\lambda$ value found by cross-validation for each pitch type. Figure 1 shows the classification accuracy of each model. This was done by setting a threshold on the outputs of the model at 0.5, which yielded the best results. Anything over the threshold value was set to 1, while anything less than or equal to the value was set to 0. The thresholded values were then compared to the actual positive outcome values and the percentage correct was taken as the classification accuracy of the model.

Table 1: Results from Logistic Regression Models

| | **Full Model** | **Forward Selection** | | **Backward Selection** | | **Ridge** |
|---|---|---|---|---|---|---|
| | **McFadden** | **# of Vars.** | **McFadden** | **# of Vars.** | **McFadden** | $\lambda$ |
| **Fastball** | 0.07 | 15 | 0.06 | 20 | 0.07 | 0.01 |
| **Curveball** | 0.13 | 8 | 0.08 | 8 | 0.08 | 0.26 |
| **Sinker** | 0.09 | 8 | 0.06 | 8 | 0.06 | 0.04 |
| **Slider** | 0.11 | 11 | 0.09 | 13 | 0.09 | 0.04 |
| **Changeup** | 0.15 | 9 | 0.09 | 11 | 0.09 | 0.14 |
| **All Pitch Types** | 0.08 | 16 | 0.07 | 16 | 0.07 | 0.01 |

## 6 Discussion and Conclusions

### 6.1 Determining Difference in Performance at High and Low Elevation

The p-values indicate very high statistical significance for the influence of Coors as a factor of positive outcome, with and without ERA included in the model. An easier interpretation of the output of the model is that there is a reduction in success rate of 3.5 percentage points for pitches thrown at Coors field.

Table 2: Most Common Variables in Stepwise Selection Models

| **Variable** | **Num. of Models With Variable** |
|---|---|
| Horizontal location | 12/12 |
| Outs | 12/12 |
| Bounced | 12/12 |
| In scoring position | 12/12 |
| Break angle | 11/12 |
| Vertical location | 10/12 |
| Count | 10/12 |
| ERA | 8/12 |
| Speed | 8/12 |

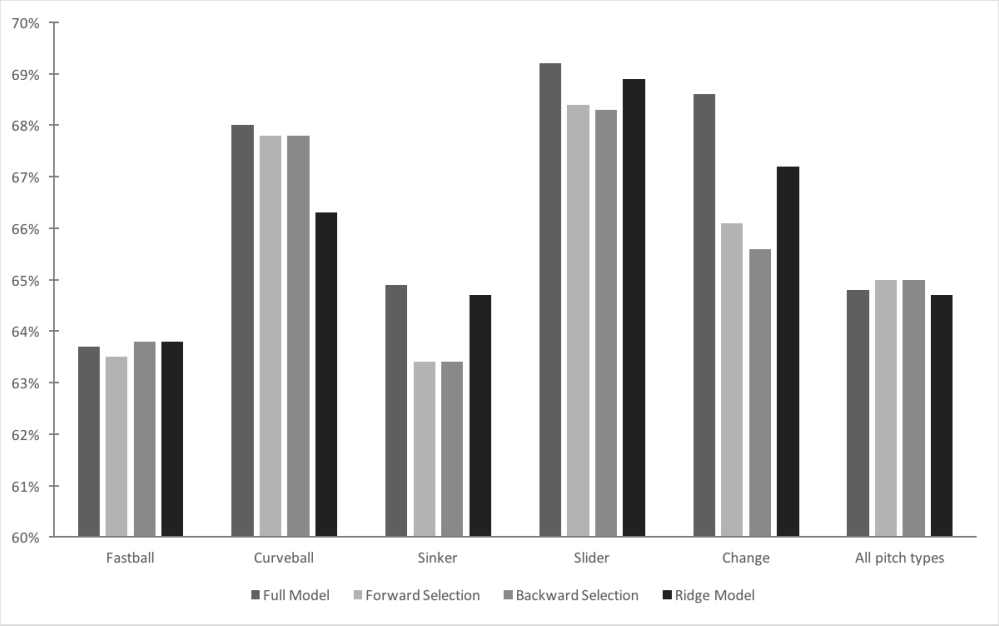


Figure 1: Classification Accuracy of Models Grouped by Pitch Type

According to these results, the belief that pitchers do worse at higher altitudes is accurate. Despite the extra speed afforded especially to the fastball, which is thrown far more frequently than any other pitch, pitchers actually struggle more. The culprit is probably the straighter (and therefore more predictable) path of the ball, caused by a reduction in the effectiveness of spin to influence the trajectory. This suggests that perhaps break is more important than sheer speed, or maybe that it is a pitcher's ability to throw a variety of pitches that makes him effective, and the higher speeds and reduced break found at Coors

field rob him of this ability. These ideas could be a good starting point for further research on this topic.

## 6.2 Predicting Performance at High Elevation

The predictive models offer further insight into the nature of pitching at altitude. The first model run featured pitch types as a variable and used the entire dataset. The pitch type variable was found to be statistically significant for all 5 types, indicating that they all have an effect on pitcher performance. In an attempt to improve on accuracy and interpretation, the data was therefore subset by pitch type, as mentioned above.

In general, the models varied in terms of statistically significant variables depending on the pitch type. This was captured by the stepwise selection, which was able to cut most of the variables out of the model for the slower pitch types that feature more break. One possible reason that so many variables were cut is that break is very important, and these models were able to explain much more of the variability relying only on break variables. The forward selection curveball model kept break angle and the vertical deviation from a straight line in the model, as well as horizontal location at the plate, three measures that can give an idea of break. The forward selection fastball model kept 15 variables to the curveball's 8, and these included a wide range of variable types. As expected, speed is very significant, but the model also contains release point variables, ERA, handedness of pitcher, strike zone location and count. This could indicate that, even at high altitude, break is more important than speed.

The McFadden values also support this idea. For the forward and backward models, one can see that even with fewer variables, the curveball, slider and changeup models explain more of the variability over the null model than the fastball model.

As a way to compare the predictive effectiveness of all the models, the classification accuracy of each was calculated. As Figure 1 shows, the different regression models tended to sit around the same accuracy for each pitch type. Across all models and types, accuracies varied from about 63% to about 69%. The main takeaway of the plot is that, except in two cases, neither stepwise selection nor ridge regression was able to improve on the accuracy of the full model. Even when beaten, the full model is only just edged out. This suggests that perhaps the multicollinearities in the full models do not have as bad an effect as one might suspect.

The results for the predictive models were not as conclusive as those for evaluating the effect of elevation. Although the predictive accuracy of the model with all pitch types included can be improved upon by subsetting the data by pitch type, one only gains about 5 percentage points at best. The bottom line is that predictive models can be built for pitcher performance at high elevation, and they can predict outcomes with about 63%-69% accuracy. Additionally, the stepwise models offer insight into the importance of the different attributes of pitch trajectory and game situations. This could be an interesting way to extend this research in the future.

# 7 References


1. Nathan, Alan M. "The Physics of Baseball Alan M. Nathan University of Illinois." Baseball At High Altitude. University of Illinois, n.d. Web. Feb. 2016

2. Weber, Roger. "Ballpark Elevation Above Sea Level" Baseball Judgments. Web. Accessed Apr. 28 2016.

3. Marchi, Max, and Jim Albert. Analyzing Baseball Data with R. N.p.: ChapmanHall, 2013. Print.

4. We would like to thank Duncan Wadsworth, our mentor on this project


# 8 Appendix

## 8.1 Full List of Variables in Models

1. Number of outs

2. Handedness of batter

3. Handedness of pitcher

4. Speed at release

5. Horizontal movement of pitch compared to a straight line from release point to the location the pitch landed in the strike zone

6. Vertical component of pitch compared to a straight line

7. The distance from the plate that the maximum break length occurred

8. The spin rate of the ball

9. The count

10. The batter's height

11. The angle of the break, measured from a bird's eye view and specific to the handedness of the batter (i.e. negative values signify the ball broke in toward the batter)

12. The previous pitch thrown in the at-bat; "None" if it is the first pitch of the at-bat

13. The vertical distance from the center of the strike zone that the ball ended up

14. The horizontal distance from the center of the strike zone, with negative value indicating that the pitch landed inside

15. Whether the pitch bounced before reaching the catcher
16. The height above the ground of the ball at release
17. The horizontal distance from the center of the pitcher's rubber of the ball at release, with a positive value indicating the same direction as the handedness of the pitcher
18. Whether there were runners on base
19. Whether there were runners in scoring position
20. The pitcher's ERA for that year
21. The ID number of the pitcher
22. The type of pitch

## 8.2 Variables Remaining After Stepwise Selection

Numbers below correspond to order in Part 8.1.

- Fastball, forward selection: 1, 3, 4, 5, 6, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20
- Fastball, backward selection: 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
- Curveball, forward selection: 1, 2, 4, 6, 11, 14, 15, 19
- Curveball, backward selection: 1, 2, 4, 6, 11, 14, 15, 19
- Sinker, forward selection: 1, 9, 10, 11, 13, 14, 15, 19
- Sinker, backward selection: 1, 9, 10, 11, 13, 14, 15, 19
- Slider, forward selection: 1, 2, 4, 9, 11, 12, 13, 14, 15, 19, 20
- Slider, backward selection: 1, 2, 4, 5, 8, 9, 11, 12, 13, 14, 15, 19, 20
- Changeup, forward selection: 1, 9, 12, 13, 14, 15, 16, 19, 20
- Changeup, backward selection: 1, 8, 9, 11, 12, 13, 14, 15, 16, 19, 20
- All pitch types, forward selection: 1, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14, 15, 16, 19, 20, 22
- All pitch types, backward selection: 1, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14, 15, 16, 19, 20, 22