Bayesian Ridge Regression for MLB Pitcher Performance

David Teuscher

December 9, 2021

1 Introduction

Pitchers play an integral role in the success of Major League Baseball (MLB) teams. One of the standard measures that has been used to evaluate a pitcher's performance is Earned Run Average (ERA), which is the number of runs a pitcher allows, excluding runs that scored due to errors. One of the shortcomings of ERA is that it is volatile and depends a lot on the quality of the fielders in the field. Previous research has shown that once the ball is put into play, the results are mostly random. As an attempt to isolate the individual performance of the pitcher, an alternative metric, Fielding Independent Pitching (FIP), was developed to assess pitcher performance. FIP accounts for the home runs a pitcher gives up, the number of walks and strikeouts, and the innings pitched, which are the events that a pitcher is able to control and is solely responsible for. Teams have used FIP to get a better idea of how a pitcher performs and contracts and salaries tend to reflect this. Pitchers with a poor ERA, but a better FIP often getting contracts that may be surprising to the casual fan, while other pitchers with a good ERA don't get paid as well because of a higher FIP suggesting potential poor performance in the future.

Since 2015, MLB has collected additional data about pitches that was not previously available, such as the spin rate, launch angle, and exit velocity of pitches. The goal of this analysis is to determine what factors a pitcher could focus on to improve his FIP, as well as illustrating penalization/regularization using the Bayesian paradigm.

2 Data

The data for this analysis was obtained from two different sources. The FIP measurements for all pitchers in 2016 was taken from Fangraphs using the baseballr package. Information about each pitcher was taken from Baseball Savant and included variables such as the average spin rate, break, and speed for fastballs, breaking, and offspeed pitches, the percentage of pitches a pitcher threw in the strike zone and out of the strike zone, the percentage of the types of hits a pitcher gave up (flyball, popup, ground ball, line drive, etc.), and the average launch angle of hit pitches. There were a total of 49 covariates that were obtained for each pitcher. Since relief pitchers and starting pitchers are generally different in how they are used and their performance, the data was filted to only included pitchers who threw at least 90 innings in 2016. Only right-handed pitchers were considered in this analysis as well since left and right handed pitchers likely have

a somewhat different approach to pitching to batters, so to avoid those differences, this analysis only focuses on the right handed pitchers, giving a total of 108 pitchers who were used in this analysis.

3 Model

The model for the data is presented below. The model is parameterized this way in order to penalize coefficients and shrink them towards zero depending on the prior distribution for σ_b^2.

$$y|\beta, \sigma^2, \sigma_b^2, \alpha \sim N(\alpha + X\beta, \sigma^2 I) \tag{1}$$

$$\beta \sim N(0, \sigma_b^2 I) \tag{2}$$

$$\alpha \sim N(4, 1.5) \tag{3}$$

$$\sigma^2 \sim IG(2, 20) \tag{4}$$

$$\sigma_b^2 \sim Gamma(40, 8) \tag{5}$$

In the model above, y is a vector of the FIP for the 108 players. The prior for β is selected to implement Ridge regression from the Bayesian paradigm. With this prior, the β coefficients will be shrunk towards zero and the amount of penalization is determined by the prior distribution for σ_b^2. The intercept shouldn't be shrunk like the other coefficients, so it is included in the model as a separate parameter, α. The average FIP is generally around 4 and can range as low as 2 and as high as 6 or 7 generally, so the prior chosen for α centered around 4 with a variance of 1.5 to reflect that belief. The prior for σ^2 was selected because it is conjugate and simplifies sampling, especially when not using a probabilistic programming language. The prior for σ_b^2 was selected so that some shrinkage would occur, but not to shrink the coefficients too much. A further analysis of the impact of the prior distribution is shown in Section 5.

4 MCMC Algorithm and Diagnostics

Posterior samples were obtained by an MCMC algorithm using Gibbs sampling and Gaussian random walk Metropolis-Hastings, as well as using Stan. The MCMC algorithm implemented is outlined below:

1. Set $r = y - \alpha$; $r \sim N(X\beta, \sigma^2)$
2. Sample β from the full conditional; $\beta \sim N((\frac{X'X}{\sigma^2} + \frac{1}{\sigma_b^2}I)^{-1}(X'r/\sigma^2), (\frac{X'X}{\sigma^2} + \frac{1}{\sigma_b^2}I)^{-1})$
3. Sample σ^2 from the full conditional; $\sigma^2 \sim IG(2 + \frac{n}{2}, 10 + \frac{(y-X\beta)'(y-X\beta}{2})$

4. Set $r = y - X\beta$; $r \sim N(\mathbf{1}\alpha, \sigma^2 I)$
5. Sample α from the full conditional; $\alpha \sim N((\frac{\mathbf{1}'\mathbf{1}}{\sigma^2} + \frac{1}{\sigma_b^2}I)^{-1}(\mathbf{1}'r/\sigma^2), \frac{\mathbf{1}'\mathbf{1}}{\sigma^2} + \frac{1}{\sigma_b^2}I)^{-1})$
6. Sample σ_b^2 using Gaussian random walk Metropolis Hastings
7. Repeat steps 1-6 for n iterations

The MCMC algorithm and the model fit using Stan converged to the same values for the parameters when taking 10000 draws. The trace plots and effective sample sizes were worse for the MCMC algorithm, so the posterior draws using Stan were used for inference because of the improved diagnostics. The inference made was the same using either parameters, but since the diagnostics are better for the draws from Stan, those draws are used. Table 1 includes some of the parameters with $\hat{R}$, effective sample size, and the Raftery-Lewis diagnostic calculed for the samples from Stan. There were two chains fit for the Stan model and they converge to the same place according to the $\hat{R}$ values and the effective sample sizes and Raftery-Lewis diagnostics look satisfactory.

Parameter	Rhat	Effective Sample Size	Raftery-Lewis
Sweet Spot Percent	1.0	7756	1.010
Poorly Under Percent	1.0	7722	0.986
Meatball Percent	1.0	7762	1.010
Pull Percent	1.0	7684	1.010
Fastball Average Break	1.0	7739	1.020
Offspeed Pitch Percentage	1.0	8374	1.020
σ^2	1.0	6254	0.986
σ_b^2	1.0	4759	1.000
Intercept	1.0	8580	0.995

Table 1: Convergence diagnostics for sampling using Stan

Table 2 includes the effective sample size for some of the parameters from the model fit using the MCMC algorithm. The effective sample size was a lot smaller for σ_b^2, so the posterior draws from Stan were used for the analysis. Regardless of the difference in effective sample size, the draws from both methods converged to the same place.

Parameter	Effective Sample Size
Sweet Spot Percent	10000
Poorly Under Percent	10000
Meatball Percent	10000
Pull Percent	10000
Fastball Average Break	10000
Offspeed Pitch Percentage	10000
σ^2	4551.736
σ_b^2	1395.28
Intercept	9501.712

Table 2: Convergence diagnostics for sampling using MCMC algorithm

As an additional diagnostic check for convergence, the trace plots for a few of the parameters are shown in Figure 1. The trace plots for these parameters look satisfactory and there are no concerns about convergence

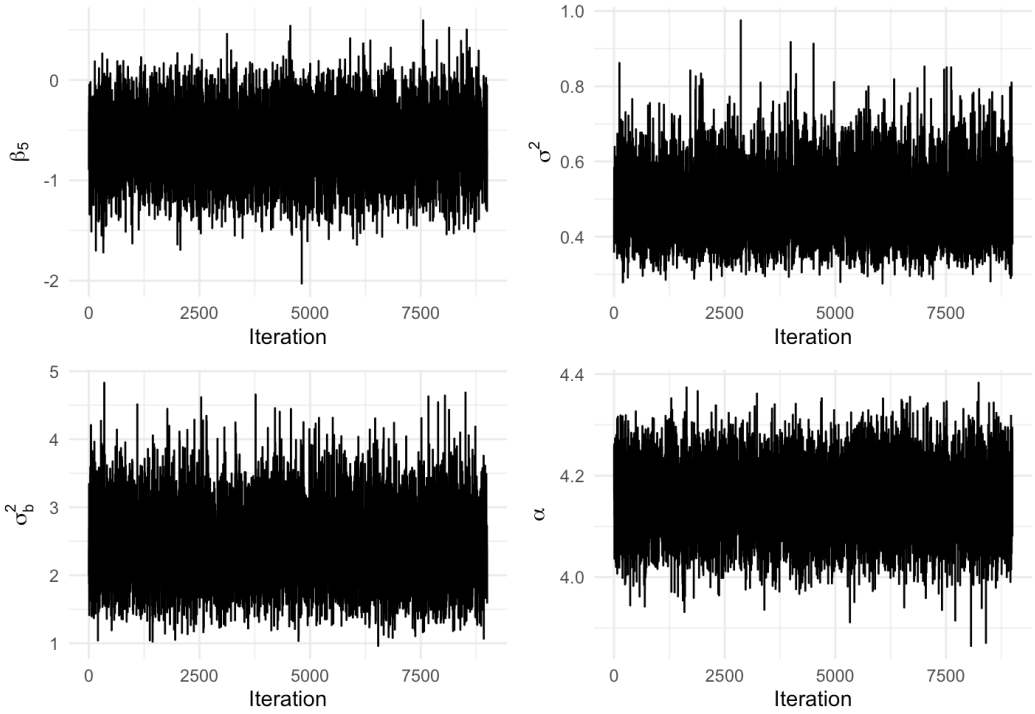

Figure 1: Trace plots for a β coefficient, α, σ_b^2, and σ^2

that arise from these trace plots. Since the trace plots look good and the diagnostics for samples from Stan are acceptable, the samples can be used as draws from the posterior distributions of these parameters for inferential purposes.

5 Sensitivity Analysis

A sensitivity analysis was performed to explore the impact of the choice of prior distribution on the results. Since σ_b^2 controls the penalization and shrinkage, three different priors for these parameter were examined.

$$\sigma_b^2 \sim Gamma(40, 8) \tag{6}$$

$$\sigma_b^2 \sim Gamma(36, 12) \tag{7}$$

$$\sigma_b^2 \sim Gamma(1, 1) \tag{8}$$

The first prior is the prior used for the model, which allows for a decent amount of variance for β. The second prior shrinks β more than the first prior, but still doesn't strongly penalize them. The third prior

imposes a large penalty on β, shrinking a lot of the values close to zero. The following figures show the 95% credible intervals for the 49 coefficients.

Figure 2 shows that a lot of coefficients are shrunk towards 0, but a number of the coefficients have a strong effect and a number of them don't contain 0 in the credible interval.

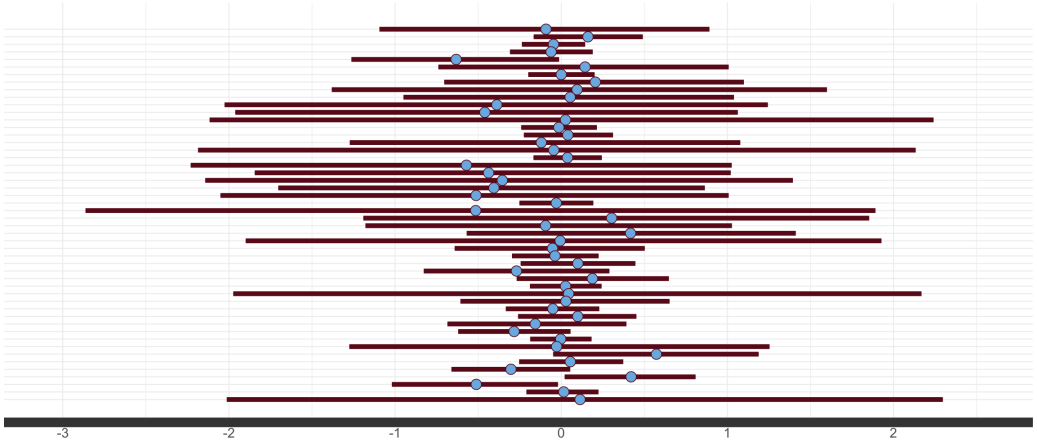

Figure 2: 95% posterior credible intervals for β for $Gamma(40, 8)$

Figure 3 shows similar intervals to Figure 2, but the estimates have begun to be pulled in more towards zero. For example, Figure 2 has a handful of coefficients that are less than -.5, while Figure 3 shows that only one coefficients is less than -.5 for the second prior distribution.

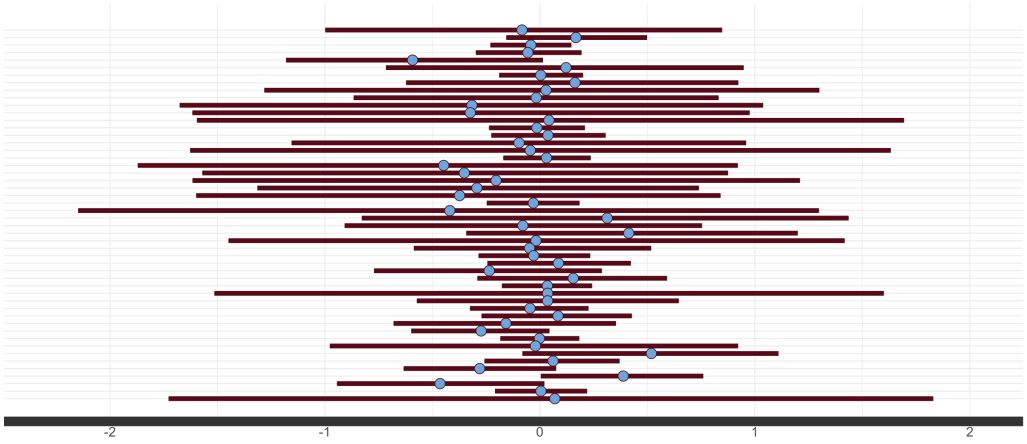

Figure 3: 95% posterior credible intervals for β for $Gamma(36, 12)$

Figure 4 shows that all of the coefficients have been pulled in towards zero because of the heavy penalization on the coefficients. All of the coefficient estimates are between -.1 and .1 and they all include 0 in the 95% credible interval, showing that the coefficients have effectively been shrunk towards zero.

The sensitivity analysis illustrates the impact of prior distributions on the results of an analysis, especially

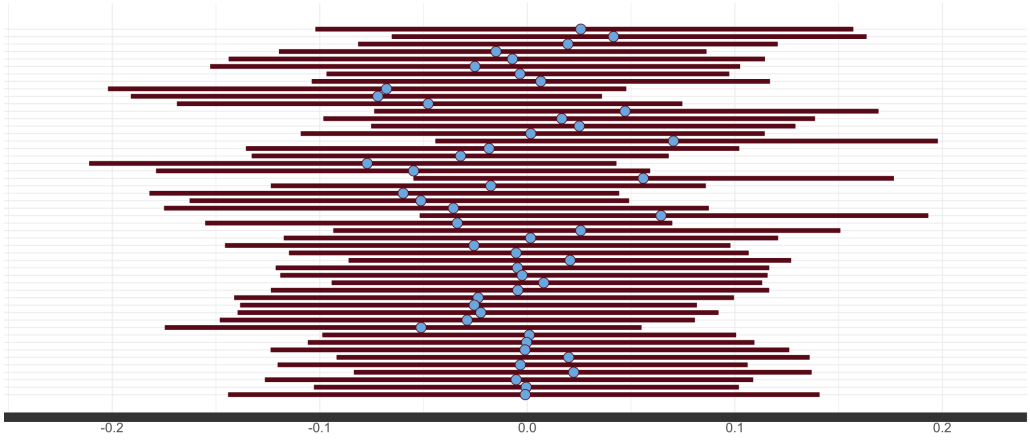

Figure 4: 95% posterior credible intervals for β for $Gamma(1, 1)$

in the context of Ridge regression where you are shrinking values towards 0. The prior distribution on σ_b^2 will impact how much the coefficients are shrunk towards 0.

6 Frequentist Analysis

As a comparison to the Bayesian Ridge regression, ridge regression was implemented for this dataset from the Frequentist paradigm. The best λ value was determined via cross-validation and then 500 bootstrap samples were generated and the coefficients for the 500 bootstrap samples were obtained. Confidence intervals for the bootstrap samples were calculated and there were 16 coefficients that did not include 0 in the bootstrapped confidence intervals.

The results from the Frequentist analysis differed greatly from the Bayesian results, where only 3 of the coefficients didn't contain zero in the 95% posterior credible interval. This illustrates that Bayesian and Frequentist methods often will not produce the result, which is often due to the information that is incorporated by prior distributions.

7 Results and Conclusion

The goals of this analysis were to show how Bayesian Ridge regression can be implemented and determine which factors a pitcher could work on to improve their FIP. The sensitivity analysis in Section 5 showed how Bayesian Ridge regression works and how the prior distribution for σ_b^2 determines the amount of shrinkage for the coefficients. Table 3 presents the coefficients fromt eh model that didn't contain zero in the 95% credible interval.

The poorly under percent is the percentage of hit balls were the batter got under the pitch a lot. It isn't surprising that as this percentage increases, a pitcher's FIP would decrease. Home runs are one of the components for calculating FIP and if a pitcher has a lot of their fly balls be ones where the batter is poorly under instead of hitting the ball on the sweet spot, they will prevent a lot of home runs from being hit.

Parameter	Mean	2.5%	50%	97.5%
Poorly Under Percent	-0.635	-1.263	-0.633	-0.013
Offspeed average vertical break	0.421	0.020	0.421	0.809
Offspeed average break	-0.513	-1.020	-0.511	-0.019

Table 3: Significant coefficients from the Bayesian model

The offspeed average vertical break has a positive coefficient, which means that as it increases, a pitcher's FIP is expected to increase. This variable is recorded where break down towards the ground is recorded as a negative number. For example, if a pitch drops 3 inches on average, this would be recorded as a -3. This means that as there is less vertical break on a pitcher's offspeed pitches, his FIP will increase. This is a reasonable conclusion since offspeed pitches that don't break are generally easier to hit since they are slower and don't move much. The average offspeed break, in the horizontal and vertical direction, has a negative coefficient, so as an offspeed pitch breaks more, a pitcher's FIP will decrease.

As a result of these findings, the most effective thing for a right handed starting pitcher to decrease his FIP would be develop an effective offspeed pitch. Most starting pitchers throw a fastball and then usually some sort of breaking ball, but many pitchers do not have an effective offspeed pitch. The effective pitchers, who have a low FIP, likely have an effective third or fourth pitch, which is likely to be an offspeed pitch. Pitchers who are average or below average likely have a no offspeed pitch or their offspeed pitches are not very good and are hit hard. As a result, the development of an effective offspeed pitch seems to be a reasonable approach for a pitcher who wants to lower his FIP.