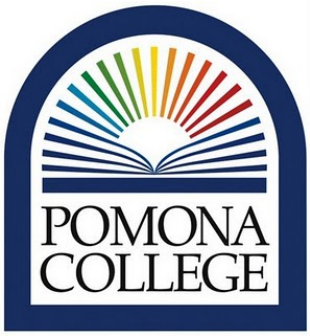
Pomona College



Senior Thesis in Mathematics

Bayesian Statistics and Baseball

Author:
Guy Stevens

Advisor:
Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 5, 2013

Contents

1 Introduction

The projection of future performance in professional baseball has long been a serious question of interest to those inside and outside the sport. It has been heavily researched in the academic community and among quantitative analysts within the baseball community. Past academic research has taken a number of distinct approaches to making predictions for hitters, but research on pitchers is a bit more limited. The study of projections often provides valuable insight into aspects of performance used in modeling.

One projection engine, created by Tom Tango, is called MARCEL; it uses a weighted average of recent performance and regresses it to the mean. It uses simple, publicly-available data and, in general, performs quite well as a prediction tool for both hitters and pitchers (Tango 2004). However, this approach leaves much to be desired. Tango explains that this model is the simplest possible model that should be accepted; while it stacks up well against some basic systems, it is meant as a simple threshold for considering a model to be effective. Baseball Prospectus' projection tool, nicknamed PECOTA, is a huge improvement on the MARCEL model; unfortunately, it uses proprietary data and significant effort to update and manage (Silver 2003). The limitations of publicly available data make building a system based purely on freely available statistics difficult. Varying career-to-date lengths of individual players and the limited information contained within full-season summary statistics pose problems to the construction of an automated prediction model. However, Bayesian techniques have successfully been used in the modeling of home runs for individual hitters (Jensen et al., 2009). A similar approach can be implemented for the modeling of pitching performance in terms of strikeouts and walks, as sabermetricians have determined these to be the elements of pitching over which pitchers have the most control.

One of the important components of any career performance model is the career trajectory. Very young players tend to improve year-to-year, while aging veterans tend to decline. This is very difficult to model, but higher-order regression models have been successful in approximating career age trajectories (Albert 2002). Understanding the effects of age on performance is key to successfully projecting performance.

Our research consists of two major pieces. The first goal of the project is to develop a model of the aging process in Major League pitchers in terms of strikeouts and walks. The second part is studying the distribution of the parameters within the model. A Bayesian model involving the aging curves introduced by Albert would provide additional information on the value of aging patterns in projecting performance. Before creating a model of individual performance, a good place to start would be looking at overall trends in the population of pitchers as a whole.

2 Variables and Data

Our data comes from a database called the Baseball DataBank. It contains full-season statistical totals for every player-season, extending back to the mid-1900's. It also includes information on player ages, physical attributes, and other miscellaneous facts. Information is gathered from this database using SQL (Structured Query Language), and an excerpt of the code used for this project is provided in Appendix A.

The data being used is seasonal totals for every pitcher-season in Major League Baseball from 1993 to 2010. I chose to use only starting pitchers, as they throw many more innings than relief pitchers do and thus provide much more reliable data, with less variability in season-to-season performance. Additionally, comparing starting pitchers to relief pitchers directly is unfair. For example, pitchers can throw harder out of the bullpen because their outings are shorter, so most former starters see an uptick in strikeouts when they move to the bullpen. Therefore, pitcher-seasons are only included if a pitcher started at least 15 games, and appeared in relief in fewer than 5 games. Additionally, I only included pitchers with eight or more seasons that meet the qualifications. Otherwise, there would be an influx of mediocre pitchers in their late-20's, as this is when pitchers who are on the cusp are generally given a season to prove they belong. They would greatly affect the results without actually providing useful information to the study of pitcher aging. Limiting the data this way gives me a total of 765 pitcher-seasons that are attributed to 74 different pitchers.

One issue with this structure is that these are not independent data. Since each pitcher appears at least

eight times in the data set, these are not 765 randomly sampled seasons over the time interval. This is a problem that will persist throughout the analysis.

For each pitcher-season, data is included for a number of variables. The hand the pitcher uses to throw (handedness) and his age are both characteristics that are used in the model. The first is a categorical variable, while the second is numerical over a defined range of integers (the youngest pitcher-season in the data set is 20 years old, while the oldest is 47 years old). The performance variables included in the model are total strikeouts (K), total walks (BB), and total batters faced (BFP). K and BB are the outcomes of interest and will be modeled in separate models, while BFP indicates how many opportunities the pitcher had to strike out or walk a batter.

Section 3 will detail the methodology used to conduct this analysis. Its subsections will explain the different pieces of the process behind our work. Some will be simplified explanations, while others will be detailed mathematical breakdowns or proofs. The different parts will fit together by the end of the chapter; in Section 4, we move on to the results of our analysis before concluding and evaluating potential future research or improvements in Section 5.

3 Methodology

3.1 Logistic Regression

Logistic regression is a generalized linear regression model used in fitting continuous or discrete explanatory variables to a binary response variable. The simple logistic regression model is

$$ln(\frac{\theta}{1-\theta}) = \beta_0 + \beta_1 x \tag{1}$$

where θ is the probability of a positive response and x represents the explanatory variable(s). The binary response variable takes a value of 1 for a positive response and 0 otherwise. We can call this variable Z with $Z \sim Bernoulli(\theta)$ so $\theta = E[Z]$.

The right side of equation (1) can take any real value while θ is constrained between 0 and 1, making θ useful as a probability. The left side is the natural logarithm of the odds ratio, which is the ratio of a probability to its complement. Logistic regression is a widely used tool for regression analysis resulting in estimation of the probability of success. Thus, we will not present the mathematical support for its use.

The logistic regression model assumes an underlying Binomial distribution of the response variable, which is very convenient because the Beta prior distribution is conjugate to the Binomial likelihood. They form the Beta-Binomial family, often used for a known number of trials in a Binomial process when the success probability is unknown and to be estimated. The utility of this link will be more apparent in Section 3.3 and beyond.

For every batter faced, the pitcher has the opportunity to strike him out. We only observe the number of batters faced and the total number of strikeouts in a season, but not the true probability of striking out each hitter. By treating each matchup as having some probability θ of a strikeout, the Binomial setting allows us to describe the likelihood of θ.

3.2 Bayesian Statistics

Bayesian statistics is a field of statistics rooted in Bayes' Theorem which, in its simplest form, is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2}$$

As shown in this formula, Bayes' Rule uses a conditional distribution $P(B|A)$ to construct a conditional distribution with the same random variables but playing opposite roles, $P(A|B)$.

Two important features of Bayesian statistics include the use of prior distributions and the use of likelihoods in the modeling of all uncertainty. *Prior distributions* are specified for every parameter in the model.

They are chosen to reflect the modeler's beliefs about the distribution of the parameter in question. If the modeler would prefer his prior beliefs not affect his results, he can choose a non-informative prior distribution. This is done by choosing such a high variance that the distribution is essentially flat over the support of the parameter. The modeler need only choose a distribution with the same support as the possible parameter values. Often, prior distributions are chosen so that they will combine with the data in the form of a conjugate family. If the data is assumed to come from a certain distribution, and a *conjugate prior* is chosen for the parameter, the subsequent steps are much simpler because the posterior distribution is in the same family as the prior.

Bayesian modeling techniques use data to update prior distributions to create *posterior distributions*. As more data is used in the model, the posterior will be updated to reflect these values. Posterior distributions provide more informative results than simple point estimates. Additionally, the results of Bayesian inference are more interpretable than those of classical inference, which relies on often misinterpreted tools such as p-values.

3.3 Hierarchical Bayesian Logistic Regression

A hierarchical model uses a series of smaller models to specify the prior or sampling distribution of the parameters (Christensen et al., 2011). In this case, the logistic regression coefficients must be given prior distributions while the response variable is assigned a likelihood function. These are separate levels of a multilevel model that we will now specify.

3.3.1 Model Specifications

We are creating two separate models: one for strikeouts and one for walks. To do so, we must specify a distribution on each variable. A given player i in year j has strikeout total K_{ij} and walk total W_{ij}, with each modeled as a Binomial variable:

$$K_{ij} \sim Binomial(X_{ij}, \theta_{ij}) \tag{3}$$

$$W_{ij} \sim Binomial(X_{ij}, \gamma_{ij}) \tag{4}$$

where X_{ij} is the number of Batters Faced for pitcher i in year j; θ_{ij} is a player- and year-specific strikeout rate; and γ_{ij} is a player- and year-specific walk rate. The number of batters faced (BFP) represents the number of opportunities for the pitcher to strike out or walk a hitter. The model assumes that each trial– that is, each batter faced– is equally likely to end in a strikeout, just as each trial is equally likely to end in a walk. Such assumptions do not hold on a per-batter basis, but they have been validated in previous literature for full-season hitting totals. It seems reasonable to extend the analogous assumption to pitching totals, as the drawbacks of the assumption are equivalent.

To model θ_{ij} and γ_{ij}, which are unobserved but related to the observed K_{ij} and W_{ij} as parameters in the Binomial distributions specified in (3) and (4), a logistic regression model is appropriate. For simplicity, for the remainder of Section 3 we will elucidate only the first model, in terms of θ_{ij} and K_{ij}, but a similar model will be constructed for γ_{ij} and W_{ij}. The rates of interest are functions of several non-performance variables: pitcher i's **handedness** in year j (H_{ij}) and pitcher i's **age** in year j (A_{ij}).

$$log(\frac{\theta_{ij}}{1-\theta_{ij}}) = \lambda + f(A_{ij}) \tag{5}$$

Rather than being a single parameter, λ is the sum of products of parameters and dummy variables by hand:

$$\lambda = \psi_R h_m + \psi_L h_n \tag{6}$$

where ψ_R and ψ_L are parameters and

$$h_m = 1 - h_n = \begin{cases} 1 & m = H_{ij} \\ 0 & n = H_{ij} \end{cases} \tag{7}$$

The function $f(A_{ij})$ defines a model of career trajectory using a player's age in that season. We expect aging to be a nonlinear process, and have chosen a cubic function. The other parameters, based on handedness, serve as the intercept for this function, so

$$f(A_{ij}) = \phi_1 A_{ij} + \phi_2 A_{ij}^2 + \phi_3 A_{ij}^3 \tag{8}$$

The handedness variables should provide interesting results in this model. We are interested to see how right-handed and left-handed pitchers differ as a whole, and any significant intercept difference between the two would be a meaningful result. Still, the aging curve is truly the key piece of the logistic regression model. After controlling for handedness differences, we hope to study how pitching performance is affected by age. The cubic model allows for improvements with experience and physical maturation while accounting for wear and tear or declining skills over time. The overall shape, as well as features such as maxima or minima, could provide insight into the aging process of Major League Baseball pitchers as a whole.

3.3.2 Parameter Priors

Each of the parameters specified in the model must be assigned a prior distribution. The Normal distribution is easy to work with and, with a sufficiently high variance, can be non-informative. This is useful because it allows the model to stress the data in the process of constructing posterior distributions. With enough data and no obvious restrictions on the parameters in question, choosing a non-informative prior may provide the best results.

The handedness intercepts (ψ_R, ψ_L) will have the same Normal prior distribution

$$\psi_h \ \sim \ Normal(0,\ 1/\tau) \quad \forall \quad h = R, L$$

The coefficients of $f(\,\cdot\,)$ will also need prior distributions. In this case, they will each have the same Normal prior distribution

$$\phi_v \ \sim \ Normal(0,\ 1/\tau) \quad \forall \quad v \in \{1, 2, 3\} \tag{9}$$

where the values of v indicate the different coefficients in the cubic model.

As discussed earlier, the extensive dataset means non-informative priors will still lead to acceptable posterior distributions. The choice of a small τ (e.g. $\tau = .001$) will make these non-informative. It is worth noting that we have assigned all of the parameters the same prior distribution. Since the aim was to assign non-informative priors, and without any reason to constrict the support of any of them, this is a justifiable choice.

3.3.3 Joint Distributions

Now that the prior distribution of each parameter has been defined, we must define conditional distributions involving combinations of these distributions. For conciseness, we will use bold lettering to represent vectors of parameters. So, we define

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_R \\ \psi_L \end{pmatrix}$$

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}$$

The joint posterior distribution of the parameters is:

$$p(\boldsymbol{\psi}, \boldsymbol{\phi} | \boldsymbol{Y}) \propto \prod_{i,j} p(K_{ij} | X_{ij}, \theta_{ij})\, p(\theta_{ij} | H_{ij}, A_{ij}, \boldsymbol{\psi}, \boldsymbol{\phi})\, p(\boldsymbol{\psi}, \boldsymbol{\phi}) \tag{10}$$

where $\boldsymbol{Y}$ represents the observed data, i.e. $(\boldsymbol{K}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{A})$.

The first term in the right-hand side of (10) is the Binomial distribution for K_{ij}. The second term, the distribution of θ_{ij} conditioned on the covariates and parameter values, must be calculated. In this model,

$$p(\theta_{ij}|B_{ij}, H_{ij}, A_{ij}, \boldsymbol{b}, \boldsymbol{\psi}, \boldsymbol{\phi}) \sim Beta(\alpha_1, \alpha_2) \tag{11}$$

where α_1 and α_2 are determined by the data. The Beta and Binomial distributions belong to a conjugate family, which means that when sampling from a Binomial distribution to update a Beta prior, the posterior distribution will also be in the Beta family. The choice of the Beta distribution for θ is typical in this type of logistic regression model (Christensen et al., 2011). The last term is the product of the prior distributions on each parameter, which were defined in the preceding section. Because they are independent, their joint distribution is just the product of their marginal distributions.

It is important to note that the equivalence relation in (10) is not equality but proportionality. That means even if the right-hand side were not so complicated, we would still be unable to sample directly from it. We know only the kernel of the posterior distribution of the parameters, but without the normalizing integrating constant we cannot fully specify the distribution. To continue, we move into the realm of Markov Chain Monte Carlo (MCMC) simulation.

3.4 Metropolis Hastings and the Gibbs Sampler

The MCMC method we will use is a hybrid of the Metropolis-Hastings (M-H) algorithm and the Gibbs Sampler. Since our knowledge of the posterior distribution of the parameters is limited to the kernel, we have to use simulation to build the posterior distribution of the parameters.

The M-H algorithm/Gibbs sampling method requires iteratively sampling from the marginal distribution on each parameter in the model. For example, we begin with a starting value for each parameter. This starting value should not affect the results, as we will discard early samples as burn-in. The starting values may thus be chosen randomly, but the algorithm may converge faster for well-chosen starting values. Once these values are set, we would ideally sample the first parameter

$$p(\psi_1|\boldsymbol{\psi_{-1}}, \boldsymbol{\phi}, \boldsymbol{Y}) = \frac{p(\boldsymbol{\psi}, \boldsymbol{\phi}|\boldsymbol{Y})}{p(\boldsymbol{\psi_{-1}}, \boldsymbol{\phi}|\boldsymbol{Y})}. \tag{12}$$

To calculate the distribution in the denominator, we would integrate b_1 out of the distribution:

$$p(\boldsymbol{\psi_{-1}}, \boldsymbol{\phi}|\boldsymbol{Y}) = \int p(\boldsymbol{\psi}, \boldsymbol{\phi}|\boldsymbol{Y}) db_1. \tag{13}$$

However, the distribution on the right side of (13) is not always integrable. Furthermore, we do not even know the density to integrate, merely the kernel of the density. Fortunately, the algorithm we will use does not require us to actually carry out this integration.

Given a starting value for each parameter, we would sample from the distribution in (12) conditional on each of the current values of the parameters. After sampling a new value of the first parameter, that will then be the value of the parameter in the conditional distribution being sampled for the next parameter. This sampling procedure is carried out for each parameter once and then returns to the first parameter and begins again. This will result in a chain of values for each parameter. While we discard a number of the first samples as burn-in to remove the effect of our choice of starting value, the expectation is that the values will converge in distribution to the true posterior distribution with enough samples.

This type of sampling strategy is useful because we can sample from these distributions without fully specifying them; that is, we know only the kernel of the posterior and thus do not have the true distribution. However, our MCMC techniques allow our iterative samples to form a Markov Chain with a stationary distribution equal to the posterior distribution we are looking to construct. The next section will prove that this convergence occurs.

3.5 Proof of Convergence

Before using Metropolis-Hastings, we will prove that the result of the M-H algorithm converges to the desired distribution. This is a three-step process, although we will focus our efforts on the third step.

We begin with Theorem 1 in Tierney (1994), which states that if a Markov chain θ is π-irreducible, aperiodic, and Harris recurrent (i.e. ergodic), then θ converges to the unique stationary distribution π. Let g be the distribution of the next value in the Markov Chain (see (16) on the next page). Rephrasing Theorem 1, we know that an ergodic Markov Chain means g will converge to the stationary distribution π.

We must then show that the Metropolis-Hastings algorithm produces an ergodic Markov chain. First, we show that a Markov Chain with stationary distribution $\pi(\theta)$ is π-irreducible if, for any initial value, there is positive probability of eventually reaching any set A for which $\int_A \pi(\theta)d\theta > 0$. We must also show that the Markov Chain is aperiodic. A periodic chain is one that can only return to an initial set at regularly spaced iterations; otherwise, it is aperiodic. To show π-irreducibility and aperiodicity, we check that regardless of where the chain was started, it is possible to get to any set of interest A in one transition. That is, if $\int_A \pi(\theta)d\theta > 0$, then $\int_A q(\theta|\theta^1)d\theta > 0$ for all θ^1.

Having met this condition, we know that the stationary distribution is unique. Therefore, only $\pi(\theta)$ satisfies the condition

$$\pi(\theta) = \int q(\theta|\theta^*)\pi(\theta^*)d\theta^*.$$

We must now assess the recurrence of the set. Harris recurrence means that for every starting value $\theta^1 = \theta_*$ and any set A with positive probability under $\pi(\theta)$, the probability that A is revisited by the chain infinitely often is one. To confirm Harris recurrence, we must ensure both π-irreducibility and the condition that if $\int_A \pi(\theta)d\theta = 0$, then $\int_A q(\theta|\theta^1)d\theta = 0$ for all initial values θ^1. This means that the transition distribution must be absolutely continuous with respect to the stationary distribution; that is, when constructing the Markov Chain, the support of the transition density must not "shrink," thus preventing the chain from exploring regions that have positive probability under the stationary distribution π.

This leaves us with a single condition to check to establish ergodicity and therefore meet the conditions set forth in Tierney's proof: $\int_A \pi(\theta)d\theta = 0$ if and only if $\int_A q(\theta|\theta^1)d\theta = 0$ for all θ^1. Using the distribution g defined above, we need to have shown that the g produced by the M-H algorithm converges to the stationary distribution π. In this case, the condition is satisfied because the distributions have positive probability on the entire real line, so we will move on to the final piece of the process with the knowledge that our g will converge to π.

The third step is to show that π, the distribution produced by the Markov chain from the Metropolis-Hastings algorithm, is the posterior distribution p. This will rely on the Law of Total Probability and the definitions of Markov chains and distributions.

Define the Markov Chain in question as θ^r, with $r = 1, 2, \ldots$. Now, consider the chain having some current value $\theta^r = j$. We consider a value θ^* that is generated from a proposal density $h(i|j)$. The closer the density $h(i|j)$ is to the actual posterior distribution the better, but the only requirement is that $h(i|j)$ has the same support as the true posterior, $p(\theta)$. We then define a function

$$\alpha(i, j) = \min\{1, \frac{p(i)h(j|i)}{p(j)h(i|j)}\}$$

Essentially, the function α gives the acceptance probability of the proposed value. So, the subsequent value of the Markov Chain is chosen according to the result of a probabilistic event

$$\theta^{r+1} = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^*, j) \\ j & \text{with probability } 1 - \alpha(\theta^*, j) \end{cases} \tag{14}$$

To determine which value θ^{r+1} will take, we draw a value u^* from $U[0, 1]$. We can rewrite (14) as

$$\theta^{r+1} = \begin{cases} \theta^* & \text{if } u^* \leq \alpha(\theta^*, j) \\ j & \text{if } u^* > \alpha(\theta^*, j) \end{cases} \tag{15}$$

Now, let $g(k)$ be the density of θ^{r+1}. The goal of this proof is to show that for θ^r with density $p(j)$, $g(k) = p(k)$. When we undergo a step of the Metropolis-Hastings process, the distribution of the new value, θ^{r+1}, is equal to the distribution of the previous value, θ^r. Once θ reaches the stationary distribution, the subsequent iterations of the algorithm will act as samples from the posterior distribution $p(\theta)$. So, let's show that this claim is true. To begin, let

$$\delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} .$$

Next, by the definition of the Markov process, we can use the Law of Total Probability to define g. We must consider all possibilities $(\theta^*, \theta^r) = (i, j)$:

$$g(k) = \sum_i \sum_j \delta_{ik} \alpha(i,j) h(i|j) p(j) + \sum_i \sum_j \delta_{jk} [1 - \alpha(i,j)] h(i|j) p(j) \tag{16}$$

and, since the first term is nonzero if and only if $i = k$,

$$g(k) = \sum_j \alpha(k,j) h(k|j) p(j) + \sum_j \delta_{jk} \sum_i [1 - \alpha(i,j)] h(i|j) p(j). \tag{17}$$

This defines the distribution of the next term, $g(k)$, in terms of distributions we have more information about: the proposal density h and posterior p. In order to simplify we will look at each term of (17) individually. We begin with the second term:

$$\sum_j \delta_{jk} \sum_i [1 - \alpha(i,j)] h(i|j) p(j)$$

which, using the law of total probability, can be broken into two terms whose subscripts show that they are sums over disjoint sets:

$$= \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)=1\}} [1 - \alpha(i,j)] h(i|j) p(j) + \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} [1 - \alpha(i,j)] h(i|j) p(j).$$

We can see that the first term of this sum will always equal zero, since each term is multiplied by $[1 - \alpha(i,j)]$ and we are summing over all values of i such that $\alpha(i,j) = 1$. Using the definition of $\alpha(i,j)$, we substitute in for p and h leaving,

$$\begin{aligned} &= 0 + \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} [1 - \frac{p(i)h(j|i)}{p(j)h(i|j)}] h(i|j) p(j) \\ &= \sum_j \delta_{jk} \sum_{\{i:\alpha(i,j)\neq 1\}} [h(i|j)p(j) - p(i)h(j|i)]. \end{aligned}$$

We insert k in place of j, as the only positive terms are those such that $j = k$; in all other cases, $\delta_{jk} = 0$, making the entire term equal zero. Therefore, we have

$$= \sum_{\{i:\alpha(i,k)\neq 1\}} [h(i|k)p(k) - p(i)h(k|i)]$$

which, by replacing i with j for simplicity in subsequent steps, leaves

$$= \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)]. \tag{18}$$

Replacing the second term from the original sum in (17), we get

$$g(k) = \sum_j \alpha(k,j)h(k|j)p(j) + \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)]. \tag{19}$$

Now consider the first term of (19):

$$\sum_j \alpha(k,j)h(k|j)p(j) = \sum_{\{j:\alpha(k,j)=1\}} \alpha(k,j)h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} \alpha(k,j)h(k|j)p(j)$$

We can plug in the value of $\alpha(k,j)$ in each term, which gives us

$$\begin{aligned} &= \sum_{\{j:\alpha(k,j)=1\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} \frac{p(k)h(j|k)}{p(j)h(k|j)} h(k|j)p(j) \\ &= \sum_{\{j:\alpha(k,j)=1\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k). \end{aligned}$$

After plugging back into (19), we have

$$g(k) = \sum_{\{j:\alpha(k,j)=1\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k) + \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)]. \tag{20}$$

Note, by the definition of $\alpha(k,j)$,

$$\{j : \alpha(k,j) = 1\} = \{j : \alpha(j,k) \neq 1\} \cup \{j : p(k)h(j|k) = p(j)h(k|j)\} \tag{21}$$

If (21) does not seem obvious at first, we can examine how the value of α is affected by the ratio of p's and h's, which should reveal some answers. By switching the positions of j and k, the second term in α is the reciprocal of its value before the switch. If

$$\frac{p(k)h(j|k)}{p(j)h(k|j)} < 1$$

switching j and k makes the ratio of interest $\frac{p(j)h(k|j)}{p(k)h(j|k)}$, which must be greater than 1. Thus, if $\alpha(j,k) < 1$, then $\alpha(k,j) = 1$. Additionally, if

$$\frac{p(k)h(j|k)}{p(j)h(k|j)} = 1$$

switching j and k means the ratio of interest will still equal 1. Under these conditions,

$$\alpha(k,j) = \alpha(j,k) = 1$$

which explains the second component of (21). Now we can move back to the proof. We use (21) to get

$$\begin{aligned} g(k) = &\sum_{\{j:\alpha(j,k)\neq 1\}} h(k|j)p(j) + \sum_{\{j:p(k)h(j|k)=p(j)h(k|j)\}} h(k|j)p(j) + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k) \\ &+ \sum_{\{j:\alpha(j,k)\neq 1\}} [h(j|k)p(k) - p(j)h(k|j)]. \\ = &\sum_{\{j:\alpha(j,k)\neq 1\}} h(j|k)p(k) + \sum_{\{j:p(k)h(j|k)=p(j)h(k|j)\}} h(j|k)p(k) + \sum_{\{j:\alpha(k,j)\neq 1\}} p(k)h(j|k) \\ = &\sum_j h(j|k)p(k) \end{aligned}$$

We can pull $p(k)$ out of the sum as it is not a function of j, giving us

$$g(k) = \quad p(k) \sum_j h(j|k).$$

Because $h(j|k)$ is a well-defined conditional density function, its sum over all j must equal 1. Thus, we are left with

$$g(k) = \quad p(k) = \pi(k),$$

since we have already shown that g converges to π. □

We explained that the Metropolis-Hastings ratio produces an ergodic Markov Chain. Theorem 1 tells us that θ must therefore converge to the unique stationary distribution. Finally, we proved that the Markov Chain converges not just to the stationary distribution but also to the posterior distribution from which we set out to sample. Recall (10) from Section 3.3.3, the left-hand side of which is the posterior distribution we would ideally sample from directly. We have now proven that this method allows us to sample from p using the right-hand side of (10), which was both too complicated to sample from and lacked the integrating constant needed to be a fully specified distribution. Therein lies the magic of the Metropolis-Hastings algorithm and MCMC methodology.

4 Results

4.1 Model Output

Now, we move on to examining the empirical results of the analysis. The model is

$$log(\frac{\theta}{1-\theta}) = \lambda_R + \lambda_L + \phi_1(Age) + \phi_2(Age^2) + \phi_3(Age^3). \tag{22}$$

Figures 1 and 2 show the shape of the curves, for strikeouts and walks respectively, based on the mean estimate of each parameter in the model. The average strikeout rate is highest at age 20 and lowest at age 40, with a local minimum at 26 and a local maximum at 34. This indicates an overall downward trend, meaning that the youngest subset of pitchers strikes more hitters out on average than the oldest subset of pitchers. However, between ages 26 and 34, the curve is increasing. This suggests that the average Major League 26 year-old actually strikes out fewer hitters than the average Major League 34 year-old.

On the other hand, the mean-parameter walks curve is strictly decreasing from age 20 to 40. The downward slope suggests that the average pitcher at any age is better at preventing walks than the average pitcher at any younger age.

Additionally, it is worth noting that the right-handed intercept is higher than the left-handed intercept for strikeouts and lower for walks, suggesting that on average, same-aged right-handed pitchers strike out more hitters and walk fewer hitters than their left-handed counterparts. By looking at the parameter values over many iterations, we can build not only a distribution of each parameter, but also a distribution of their difference. We can use this distribution to evaluate the significance of the difference between right-handed and left-handed pitchers.

As shown in Figures 3 and 4, the distribution of the difference in intercept from both models is nicely bell-shaped. More importantly, neither one contains zero. That means that the probability of there being no difference in strikeout and walk rates between right- and left-handed pitchers is essentially zero. In both cases, all of the values with positive probability in the distribution of differences suggest that, on average, right-handed pitchers perform better. In a non-Bayesian statistics framework, this would correspond to an approximately zero p-value and clear statistical significance. The conclusion can be drawn that on average, right-handed pitchers outperform left-handed pitchers in both strikeouts and walks.

In the same vein, we can represent the distribution of entire aging curves by plotting many of the curves sampled from the posterior distribution. Figure 5 shows the distribution of left-handed strikeout rate aging

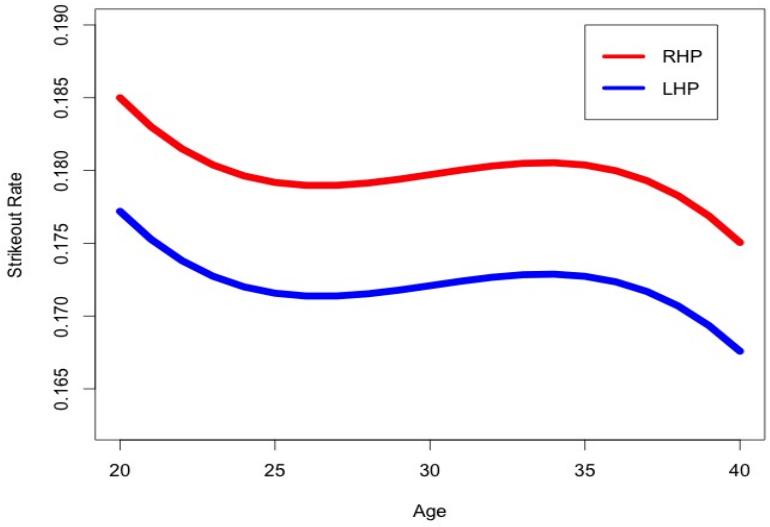


Figure 1: Cubic aging model of strikeout rate (K%) using parameter means

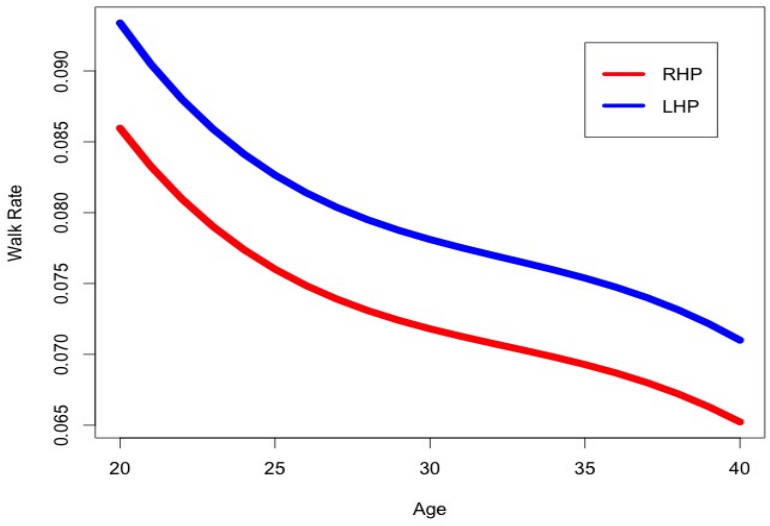


Figure 2: Cubic aging model of walk rate (BB%) using parameter means

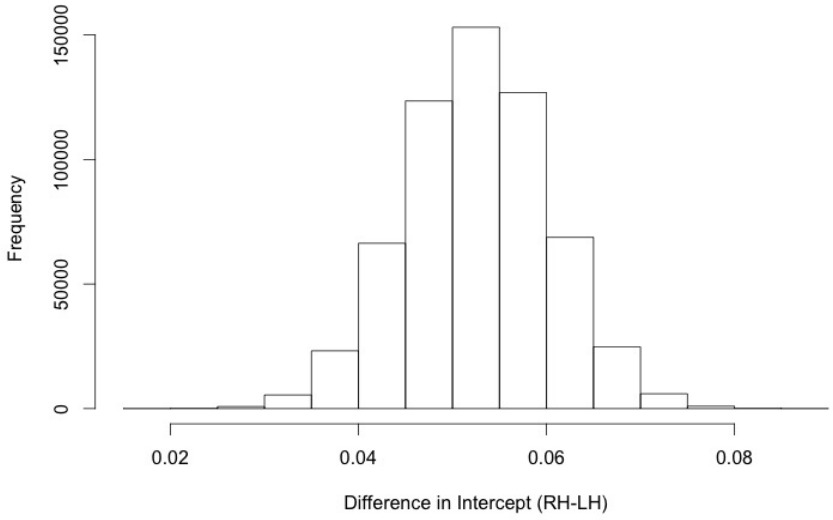


Figure 3: Histogram of difference between RH and LH intercept (Strikeouts)

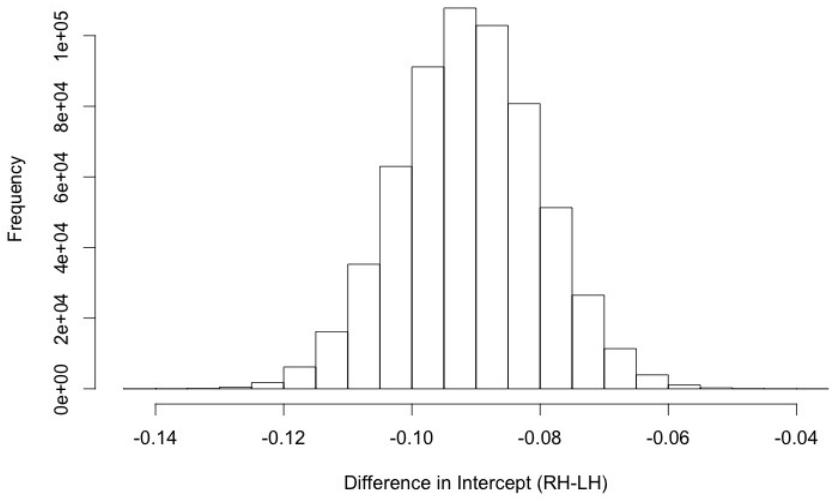


Figure 4: Histogram of difference between RH and LH intercept (Walks)

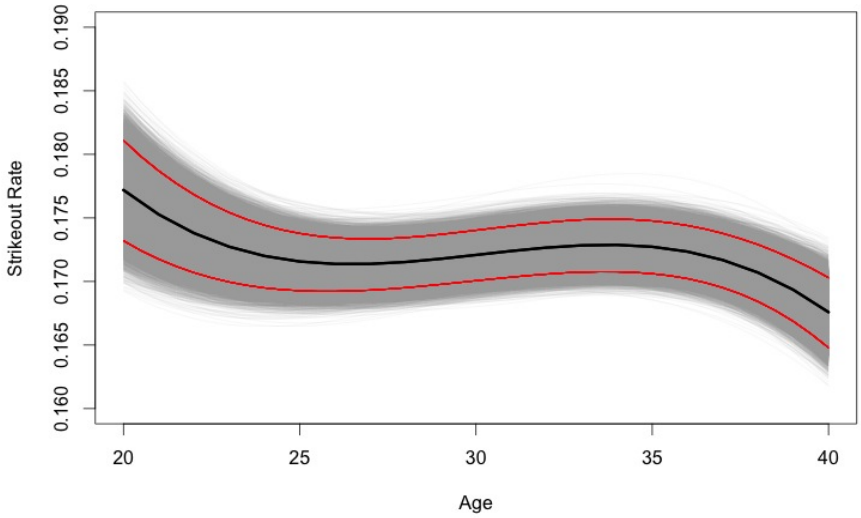


Figure 5: Distribution of LHP Aging Curves (K%)

curves as an example. The grey block is actually a set of hundreds of thousands of aging curves sampled from the posterior distribution, with only the most extreme curves visible as unique lines. The two red lines on the plot represent the 10^{th}-percentile and 90^{th}-percentile curves, while the black line in the middle shows the mean parameter curve. The black curve in Figure 5 is the same as the blue curve in Figure 1; this time, it is shown alongside other left-handed curves sampled from the distribution rather than its right-handed counterpart. These curves form a distribution of aging curves, much like a histogram would, just without making the details of the distribution's shape clear; rather, it shows the overall shape with respect to strikeout rates instead of aging function parameters that are difficult to interpret as standalone values.

4.2 Individual Pitchers

To get a different look, Figures 6 through 9 show how four pitchers saw their strikeout and walk rates change as they aged. These were all elite pitchers who are legitimate candidates for the Baseball Hall of Fame. Randy Johnson is arguably the best left-handed pitcher of all time, a power pitcher who routinely led the league in strikeouts. Though he did walk his fair share of hitters relative to other elite pitchers, his walk rate overall was nothing to scoff at. Greg Maddux was a control specialist, with one of the best career walk rates in history. His strikeout rate was unimpressive, but he still put together a fantastic and lengthy career. Mike Mussina did it all, regarded as a top-tier power and finesse pitcher at different points in his career. He may be the least accomplished of the group, but he had a long career and remarkably consistent results. Finally, Curt Schilling was a pure power pitcher who overpowered hitters with a full arsenal of pitches. He also managed to keep his walk rates down, such as when he paired the second-highest strikeout rate with a league-low walk rate in 2002 at age 35.

There are some similarities in the pitchers' trends and the model-estimated curves shown earlier, but there are also distinct differences. Their walk rates (in green) show a decreasing trend with age, similar to the trend seen in the model's results. Maddux is an exception, as his walk rate was among the best in the league every year, making it almost impossible for him to bring it any lower; even his worst walk rate, at age 27, was top-10 in the Major Leagues that season. On the other hand, Maddux's strikeout rate (in orange) decreased over the course of his career, which matches the model as well. However, the other three

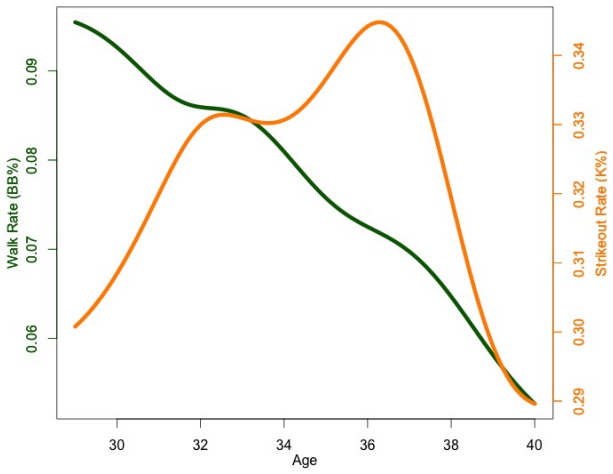


Figure 6: LHP Randy Johnson

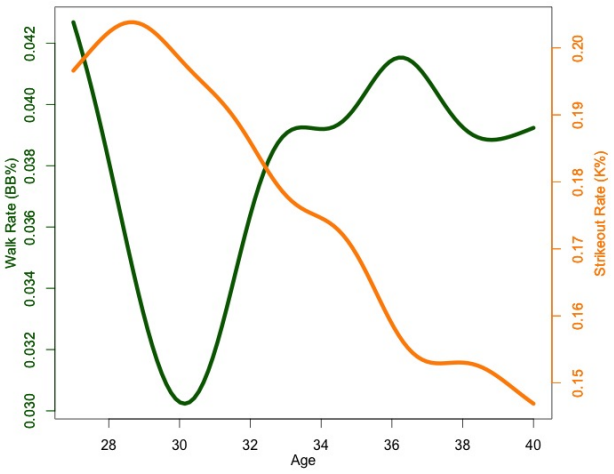


Figure 7: RHP Greg Maddux

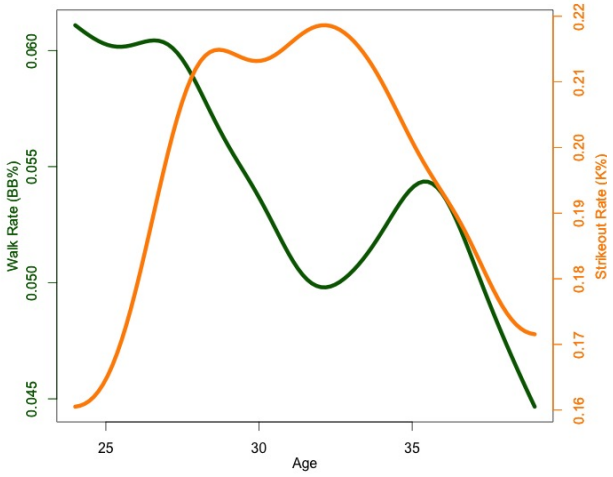


Figure 8: RHP Mike Mussina

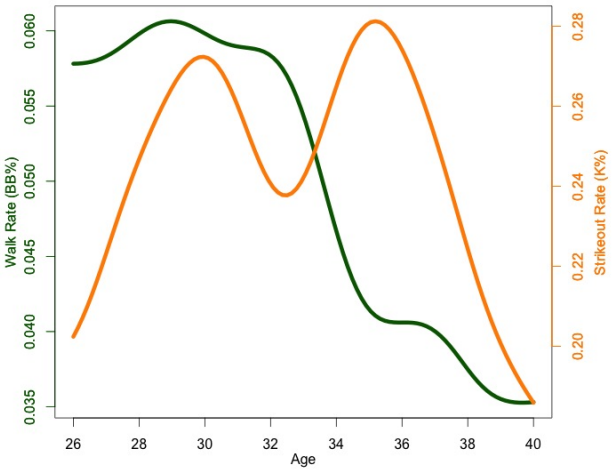


Figure 9: RHP Curt Schilling

pitchers saw early increases in strikeout rate, peaking in the middle of their career in the early- to mid-30's and declining steeply over their last few years. Each of these pitchers pitched until age 39 or beyond, but posted career lows in strikeouts immediately preceding retirement.

While these pitchers were included in the model, their curves do not necessarily mimic the model's estimated population aging curve. This is because the model does not aim to represent individual aging curves, but rather the average strikeout and walk rates for all pitchers at a given age. Furthermore, the individual pitchers discussed here have elite skill sets and do not necessarily represent the rest of the population. Greg Maddux had league-average strikeout rates, and saw his strikeouts trend with the general population. His elite tool was his control, which he used to maintain well-below-average walk rates. Johnson, Mussina, and Schilling set themselves apart with exceptional strikeout ability but improved their walk rates over their careers, just as the general population sees decreasing walk rates with age. These four pitchers were a class above their contemporaries, yet followed the league trend in either strikeouts or walks. It was their incredible ability in the other statistic that made them such dominant performers.

A key idea to keep in mind when evaluating these curves is that the model estimates averages by age, without taking into account specific players' career trajectory. When individual pitchers' show trends that deviate from this model, it is not an indictment of the model but rather a look at how unique pitchers can be. The four pitchers from Figures 6 through 9 show similarities and distinct differences to the league-average model, as do many other pitchers in the sample.

5 Discussion

These results are an interesting look into the type of analysis at a Bayesian modeler's disposal. There are some clear opportunities for further investigation. It would be interesting to see if the patterns were different by looking at pitcher-seasons as rookie season, sophomore season, third season, and so on rather than simply linking them to age. Pitchers reach the Major Leagues at different ages and improvement can come not only with age but also with Major League experience. Similarly, while pitchers gain experience, they also increase the wear and tear on their arms, which could be detrimental to performance. This is part of what makes the aging curve such an interesting topic of study.

Another huge opportunity to expand on this analysis would be to move away from the league-average estimation and aim for more specificity toward individual pitchers. Including player-specific parameters would vastly increase computation time and could lead to overfitting, but prior works have explored different methods of accounting for a player's past performance when estimating parameters. This could provide some very interesting results as well as significant value to a baseball analyst.

That the model gives league-wide averages rather than predictions of individual player performance prevent it from being a very powerful tool for evaluating and projecting pitchers. Regardless of the practical value of these specific results, the use of Bayesian Hierarchical Modeling and the Metropolis-Hastings algorithm provides an opportunity to look at the distributions of the parameters in a logistic regression model in a baseball setting. While there are clearly some interesting results from a baseball standpoint, the modeling process is where much of the most interesting research took place.

6 Appendix A: Code

Here is an excerpt of SQL (Structured Query Language) code used to pull the requisite data from the database of baseball performance statistics:

```
SELECT
a.*,
IF(b.birthMonth< 9,a.yearID−b.birthYear,a.yearID−b.birthyear−1) AS Age,
b.throws AS Hand
FROM Pitching a
JOIN Master b
ON a.playerID=b.playerID
WHERE a.yearID> 1992
AND a.GS> 15
AND a.G−a.GS< 5
ORDER BY playerid, yearid
```

References

Albert, Jim. "Smoothing Career Trajectories of Baseball Hitters," August 22, 2002.

Albert, Jim. "Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time" Journal of Quantitative Analysis in Sports 2.1, 2011.

Christensen, Ronald, Wesley Johnson, Adam Branscum, and Timothy E. Hanson. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton, FL: CRC, 2010. Print.

Jensen, Shane T.; McShane, Blakeley B.; and Wyner, Abraham J. (2009) "Hierarchical Bayesian Modeling of Hitting Performance in Baseball," Bayesian Analysis: Vol. 4: Num 4, 631-652.

Piette, James; McShane, Blakeley B.; Braunstein, Alexander; and Jensen, Shane T. (2010) "A Point-Mass Mixture Random Effects Model for Pitching Metrics," Journal of Quantitative Analysis in Sports: Vol. 6: Iss. 3, Article 8.

Silver, Nate. (2003). "Introducing PECOTA." Baseball Prospectus, 2003: 507514. 631, 641, 647

Tango, Tom. (2004). "Marcel The Monkey Forecasting System." Tangotiger.net, March 10, 2004. URL http://www.tangotiger.net/archives/stud0346.shtml 631, 632, 641, 647.

Tierney, Luke. (1994) "Markov Chains for Exploring Posterior Distributions," Annals of Statistics: Vol. 22, Number 4, 1701-1728.