2006-03-09

Modeling the Performance of a Baseball Player's Offensive Production

Michael Ross Smith
Brigham Young University - Provo

MODELING THE PERFORMANCE OF A BASEBALL PLAYER'S OFFENSIVE PRODUCTION

by

Michael R. Smith

A project submitted to the faculty of

Brigham Young University

In partial fulfillment of the requirements for the degree of

Masters of Statistics

Department of Statistics

Brigham Young University

April 2006

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

Of a master's project submitted by

Michael R. Smith

This master's project has been read by each member of the following graduate committee and majority vote has been found to be satisfactory.

Date

Scott Grimshaw, Chair

Date

Gilbert Fellingham

Date

Shane Reese

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the master's project of Michael R. Smith in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_________________ ______________________________________

Date

Scott Grimshaw
Chair, Graduate Committee

Accepted for the Department

____

G. Bruce Schaalje
Graduate Coordinator

Accepted for the College

____

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

MODELING THE PERFORMANCE OF A BASEBALL PLAYER'S OFFENSIVE PRODUCTION

Michael Smith

Department of Statistics

Masters of Science

This project addresses the problem of comparing the offensive abilities of players from different eras in Major League Baseball (MLB). We will study players from the perspective of an overall offensive summary statistic that is highly linked with scoring runs, or the Berry Value. We will build an additive model to estimate the innate ability of the player, the effect of the relative level of competition of each season, and the effect of age on performance using piecewise age curves. Using Hierarchical Bayes methodology with Gibbs sampling, we model each of these effects for each individual. The results of the Hierarchical Bayes model permit us to link players from different eras and to rank the players across the modern era of baseball (1900-2004) on the basis of their innate overall offensive ability. The top of the rankings, of which the top three were Babe Ruth, Lou Gehrig, and Stan Musial, include many Hall of Famers and some of the most productive offensive players in the history of the game. We also determine that trends in overall offensive ability in Major League Baseball exist based on different rule and cultural changes. Based on the model, MLB is currently at a high level of run production compared to the different levels of run production over the last century.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1.0

Introduction

The purpose of this project is to create a model to evaluate the offensive performance of MLB players. In the game of baseball, the offensive goal of a team is to score as many runs (R) as possible. This goal is accomplished by the batter facing a pitcher (an at-bat (AB)) and either getting a hit (H), a walk (BB), sacrifice (SF), error (E), out or a hit-by-pitch (HBP). Some hits are more valuable because they allow the hitter and the runners already on base to advance. A hit in which the batter makes it safely to first base is called a single (1B). If the ball is hit far enough, the batter can advance more bases. If the batter makes it to second base it is called a double (2B), and if the batter makes it to third base it is called a triple (3B). If the ball is hit over the outfield fence, this is called a home run (HR) and the batter gets to circle the bases. When a hitter makes contact with the ball but does not make it safely to first base, this is called an out (a groundout or fly out). Sometimes, a batter can be out at first but still advance the runners already on base or allow the runner on third to tag up and score (called a sacrifice hit (SH)) or the batter can strikeout but still make it to first if the catcher drops the ball.

These core events have been calculated differently from one era to the next. For example, let's consider at-bats. At-bats included walks in 1887, sacrifice hits from 1889 to 1893, and sacrifice flies at various times between 1930 and 1953. Before the official rule of the four-ball walk was adopted in 1889, a walk varied from 5 to 9 balls. Some statistics were not officially recognized until after 1900. The strikeout (K) as an out for a batter was not officially recognized as a baseball statistic until after 1912. Caught-stealing (CS) was not adopted officially in the majors until 1920. A fly ball that scores a runner, called a sacrifice fly (SF), became its own category apart from a sacrifice hit in 1954. The intentional walk (IBB) was differentiated from a walk starting in 1955 (Thorn, 2001).

From these core baseball statistics, the following formulae allow calculation of summary values:

$$BA = Batting\ Average = \frac{1B + 2B + 3B + HR}{AB}$$
$$OBP = On - Base\ Percentage = \frac{1B + 2B + 3B + HR + BB + HBP}{AB + BB + HBP + SF}$$
$$SLG = Slugging\ Percentage = \frac{1B + 2(2B) + 3(3B) + 4(HR)}{AB}$$

Many researchers have proposed ways to calculate a player's offensive contribution to team performance (Lindsey, 1963; Pankin, 1978; Bennet and Flueck, 1983; Berry, 2000). Batting average, home runs, and runs-batted-in have traditionally been used to measure a player's offensive value. These traditional statistics all have problems. A key problem in using runs-batted-in or RBIs to assess player value is the statistic's dependence on the strength of the team that surrounds the individual player. Runs and RBIs are a culmination of a series of events. If a player gets on base, he relies on his teammates to bring him in to score. Similarly, if a player gets a hit he is dependent on his teammates being on base so that he can drive them in. Home runs are dependent on the stadium or era in which a player is batting. For example, home runs are much more common today than they were in the 1960s or the 1980s. More home runs have been hit on average at some stadiums than others because of differing ballpark dimensions, climate and altitude conditions. The key problem with batting average is that even though it is only dependent on what an individual batter accomplishes, it does not take into account the value of getting on base with a walk and does not discriminate between the value of a single, double, triple, or a home run (Hoffman, 1989; Berry, 2000).

Because the goal of the game of baseball is to score as many runs as possible and prevent the other team from scoring, the number of runs a team scores is the key to the calculation of offensive value. The player value should be based on a player's contribution to the team total of runs scored. The number of runs scored at the team level is used because at the individual player level, runs scored is dependent on the other team members. There have been various measures created over the years based on analyzing runs scored at the team level (Lindsey, 1963; Pankin, 1978; Bennet and Flueck, 1983; Berry, 2000). These measures generally assign a value to each offensive event (such as a 1B, BB, or HR). These assigned values act as weights and the resulting measure is a weighted average of these values that is highly correlated with runs scored on a team level. In this thesis, we use the Berry Value because it is most highly correlated to runs scored at the team level. We propose a methodology that will compare players' offensive production for all players from 1900 to 1998.

This comparison is complicated by the fact that ballplayers are from different time periods. The previously mentioned measures are static for the effect of each offensive event on runs scored. A more complete model would account for these effects varying over time. The era in which a ballplayer plays should be a factor in determining his player value.

To create comparable player value system, we must determine how much of a player's performance is due to the ballpark he is playing in, the quality of competition he is playing against, the quality of his teammates, the equipment he uses or the coaches he plays for. Further, we must consider how much of his performance is due to his own ability level. In this paper, we will be focusing on the impact of season (or era) and age. Generally, a player's performance tends to increase until they reach a certain age and then begins to decline. Also, because the game of baseball has changed so much over the years, it is important to be aware of the impact of seasonal differences. Other factors may also be important, but most, like equipment and coaches, would be difficult to analyze because of lack of records or confounding elements. Berry, Reese, and Larkey (1999) look into the impact of different ballparks in addition to age and era effects. The objective in this thesis is to construct a system from which we can simultaneously compare players of different eras and ages and gain the ability to assess them on equal grounds.

Our approach to assessing offensive production is framed within a Bayesian Hierarchical Model (BHM) using Markov Chain Monte Carlo (MCMC) (Gilks, Richardson, and Spiegelhalter, 1996) as the computational tool. We present these methodologies and discuss why they are appropriate and beneficial to building a player performance model in Chapter 2 in which we review current approaches and methods used to calculate the offensive productivity of a player. Additionally, we look into the benefits and problems within each of these formulations. We then review the statistical foundations of BHM. In Chapter 3, a simple BHM is discussed in detail, and an MCMC model is presented and shows the importance of modeling a player’s aging curve. In Chapter 4, a BHM is estimated for the offensive productivity of players that models aging and the effect of different eras. By putting players on a comparable scale, we then rank players by their offensive value.

Chapter 2.0

Literature Review

Section 2.1 Baseball Statistical Measures

We looked at how the offensive value or worth of a player has been modeled in the past, and found that most models of player performance were formulated using the basic hitting statistics (At-bats, Hits, Doubles, Triples, Home Runs, Walks, etc.) and their relationship to runs scored. Researchers used these basic hitting statistics in a variety of ways.

A paper on modeling player performance could not be complete without the mention of Bill James (Lewis, 2003). James sparked a revolution in the way that baseball statistics were viewed with his *1977 Baseball Abstract*. Starting with fielding in 1977 and moving to hitting in 1979, James' main focus was the gross miscalculation of a player's value based on then current baseball statistics and the need to create new methods that better mapped a player's value. In his *1979 Baseball Abstract*, James wrote: "*A hitter should be measured by his success in that which he is trying to do, and what he is trying to do is create runs."* The new measure that he created was what he called the "Runs Created" formula:

$$TB = 1B + 2(2B) + 3(3B) + 4(HR)$$
$$Runs\ Created = ((H + BB)TB)/(AB + BB)$$

This function is the basis of many of the offensive production functions in the rest of this chapter.

Schultz (1995) used structural equation modeling to examine the stability of individual baseball player performance. This type of modeling allowed Schultz to create two latent constructs (POWER and AVERAGE), and analyze the year-to-year stability of these constructs using a combination of factor analysis and regression. His power construct consisted of HR, SLG, and RBI. His average construct consisted of BA, OBP, and R. He concluded that power is very consistent over time but average is less consistent.

Bennett and Flueck (1994) used Player Game Percentage (PGP) to measure the value of each player to his team in the 1992 and 1993 World Series. PGP estimated player value based on contribution to a team's victory, measured by the degree to which a player increases or decreases his team's probability of victory. A player's PGP for a game is the sum of these probability changes for each play that a player participates in. Each probability change is divided evenly between the hitter and the pitcher. On plays like stolen bases, the probability change is divided evenly between the catcher and the runner. Using PGP, Bennett was able to select a most valuable and least valuable player for the 1992 and 1993 World Series. One problem with PGP was that it overstated the defensive contribution of the pitcher and catcher. The pitcher should not be held defensively responsible for all offensive contributions of the other team.

Covers and Keilers (1977) used Markov Chains, sought to find a good index of a player's offensive effectiveness. Their data was derived from the box score data of all games for a particular player. OERA (Offensive Expected Runs Per Game) was based on the total runs scored by a line-up of a given player, which can be computed both empirically from the data and through a probability algorithm. Empirically, the value is computed by starting at the beginning of a player's career. A record is created, describing what happened during each at bat (out, single, double, triple, etc.) of that player's career. Then, a game was simulated as if each of the player's career at-bats were consecutive plate appearances in the same game for the same team. Game rules were as follows: 1) sacrifices were not counted, 2) error was considered as an out, 3) no runners advanced on an out, and 4) singles and doubles would both advance runners two or three bases. Based on these rules, the number of runs this player would score in a nine-inning game was computed. Computationally, the probability of six different at-bat occurrences (outs, walks, singles, doubles, triples, and home runs) was calculated from a player's data. The following functions were used to calculate OERA:

s = one of the 24 possible states of the game for the 8 possible runners on base states and the 3 out states

H = the hit type, either a 0 for an out, a B for a walk, or 1,2,3, or 4 for a single, double, triple, or home run

s' = the new state of the game after a particular hit type

$R(H,s)$ = the runs scored by a hit at a particular state of the game:

$$R(s) = \sum_H p_H R(H,s)$$

$$p(s'|s) = \sum_H f(H,s) = s' p_H$$

$$E(s) = \sum_H p_H (E(f(H,s)) + R(H,s) = \sum_{s'} p(s'|s)E(s') + R(s)$$

$E(s)$ was a Markov Chain calculating the expected runs in an inning beginning with state (s). Using the theory of ergodic Markov chains, if batting events were independent and identically distributed random variables, then the simulated OERA approached nine $E(1)$, the expected runs scored in an inning starting at the first state ($s=1$), with the probability of one. Covers and Keilers used this function to rank hitters.

Bukiet, Harold, and Palacios (1997) used Markov Chains to evaluate the performance of teams and the influence of a player on team performance. The root of their analysis was to create optimal batting orders for all teams. They set up a 25X25 transition matrix for each player, which included one row and column for each of the twenty-four potential states of the game (eight possible base-runner combinations multiplied by the three out combinations) plus one row and one column for three outs. The entries in this matrix were the probabilities that a player would change the current state of the game to any other game state in a single plate appearance. Using 0 as the beginning state of the game, *n* as the current state of the game, *U* as a 21X25 matrix with rows equal to the number of runs and columns equal to the current state, and p_k as the transition probability leading to the scoring of *k* runs, the following function was used:

$$\begin{aligned} U_{n+1}(rowj) = U_n(rowj)p_0 + U_n(rowj-1)p_1 + U_n(rowj-2)p_2 \\ + U_n(rowj-3)p_3 + U_n(rowj-4)p_4 \end{aligned}$$

This matrix would be iterated until the last column was reached using a set of rules to govern the transition that occurs at each hit type. The scoring index (U) of nine starting players was used to find the optimal batting order. This could also be used to evaluate trades by 1) switching player order and scoring indices, and 2) calculating runs scored.

Lindsey (1963) created a dataset based on over 400 baseball games occurring in the 1959 and 1960 baseball seasons. This data included the number of outs and the base runner position during each at-bat in those 400 games. By tabulating the number of runs scored at each base-out combination, he found out when it would be advisable (when it would decrease or increase the probability of scoring a run) to intentionally walk a batter, to attempt a double play, attempt a sacrifice, or attempt to steal a base. From the tabulation of runs-scored at each base-out combination, he was able to create a measure of batting efficiency. Using the notation base runner position (B), runner on first (1), runner on second (2), runner on third (3), runner on first and second (12), runner on first and third (13), runner on second and third (23), runners on first, second, and third (F), number of outs in the inning (T), number of times that base-out scenario occurred (N(T,B)), probability that runs were scored in that base runner-out scenario (P(r|T,B)), and the expected number of runs scored in a particular base runner-out scenario (E(T,B)), the results are summarized in Table 2.1.

Table 2.1: Lindsey's approach to calculating value for each offensive event by creating a table of probability for each runner-out scenario.

B	T	N(T,B)	P(0|T,B)	P(1|T,B)	P(2|T,B)	P(>2|T,B)	E(T,B)
0	0	6561	0.747	0.136	0.068	0.049	0.461
0	1	4664	0.855	0.085	0.039	0.021	0.243
0	2	3710	0.933	0.042	0.018	0.007	0.102
1	0	1728	0.604	0.166	0.127	0.103	0.813
1	1	2063	0.734	0.124	0.092	0.05	0.498
1	2	2119	0.886	0.045	0.048	0.021	0.219
2	0	294	0.381	0.344	0.129	0.146	1.194
2	1	657	0.61	0.224	0.104	0.062	0.671
2	2	779	0.788	0.158	0.038	0.016	0.297
3	0	67	0.12	0.64	0.11	0.13	1.39
3	1	202	0.307	0.529	0.104	0.06	0.98
3	2	327	0.738	0.208	0.03	0.024	0.355
12	0	367	0.395	0.22	0.131	0.254	1.471
12	1	700	0.571	0.163	0.119	0.147	0.939
12	2	896	0.791	0.1	0.061	0.048	0.403
13	0	119	0.13	0.41	0.18	0.28	1.94
13	1	305	0.367	0.4	0.105	0.128	1.115
13	2	419	0.717	0.167	0.045	0.071	0.532
23	0	73	0.18	0.25	0.26	0.31	1.96
23	1	176	0.27	0.24	0.28	0.21	1.56
23	2	211	0.668	0.095	0.17	0.067	0.687
F	0	92	0.17	0.27	0.17	0.39	2.254
F	1	215	0.31	0.242	0.186	0.262	1.632
F	2	283	0.645	0.114	0.11	0.131	0.861

From the values included in this table, the value of a type of hit for a particular base runner-out scenario was calculated by taking the number of runs scored in the situation plus the increase in expected runs (E(T,B') – E(T,B)) where E(T,B') was the new runner scenario after the hit using the following assumptions:

1. runners always scored from second and third on any type of hit
2. runners scored from first on a triple
3. runners went from first to third on half of the doubles and scored from first on the other half.

These hit values were multiplied by the percentage of times that the base runner-out scenario occurred (N(T,B)/$\Sigma_{T,B}$N(T,B)) and summed over all base runner-out scenarios. From performing these calculations, Lindsey created coefficients for each hit type from which he built the following measure for offensive ability:

$$Lindsey\ Value = 1B + 1.97(2B) + 2.56(3B) + 3.42(HR)$$

For each individual batter, his value would be the sum of the products of each of these hit type values times the number of that particular hit type divided by the number of at-bats. This would give each batter a “runs per at-bat” value to measure hitting effectiveness. There were some problems with this measure: first, the formula only took into account the value of each type of hit; second, the value of a walk, sacrifice fly, or out were not included; third, Lindsey concluded that it would be difficult to generalize results outside the small range of years of data that he modeled on.

Pankin (1978) attempted to find a statistical formulation which 1) indicated how well a player performs offensively, 2) was independent of teammate performance, and 3) accounted for differing degrees of opportunity. Following Lindsey’s lead, Pankin created an offensive performance average (OPA) with the idea that each base runner-out combination had a different number of expected runs. Pankin rounded the Lindsey values for each hit type and used Lindsey’s approach of summing the differences in expected runs scored to calculate a value for BB and SB. The following is Pankin’s OPA:

$$OPA = \frac{1B + 2(2B) + 2.5(3B) + 3.5HR + 0.8(BB + HBP) + 0.5SB}{AB + BB + HBP}.$$

Pankin also adjusted this formula to include the negative influence of an out as follows:

$$OPA_3 = OPA - 0.65(AB - H)/(AB + BB + HBP) .$$

He then compared his function to runs scored by team from 1965 to 1975. OPA had a very high correlation with runs scored and was significantly better than batting average and slugging percentage. One problem with Pankin's function was that because he only compared it to data gathered from 1965 to 1975, it was not known if his function would perform well over a long time period or over different eras of the game. Pankin's function was also very ad hoc and not based on an analysis of the runs scored data. Finally, Pankin's function did not account for other negative impact variables of baseball, such as being caught stealing or striking out.

Bennett and Flueck (1983) evaluated the accuracy of 10 baseball offensive performance models and created their own model based on 1969 to 1976 major league baseball team statistics. They included the traditional baseball evaluation functions (batting average, on-base percentage, and slugging percentage) and the functions that have been introduced above in their evaluation. In their regression analysis, they included all offensive baseball statistics and used a combination of adjusted R^2 and Mallow's C_p to select the number of variables. They came up with the following Expected Runs Produced (ERP) model from their regression:

$$\begin{aligned} ERP = {} & .499(1B) + .728(2B) + 1.265(3B) + 1.449HR + .353BB + .362HBP \\ & + .126SB + .394SF - .395GIDP - .085OUT \end{aligned}$$

In their analysis, Bennett and Flueck (1983) found that batting average least correlated with runs scored ($R^2 = .644$) and slugging percentage and on-base percentage correlated significantly better than batting average but were still two of the least correlated with runs scored ($R^2 = .822$ and $R^2 = .824$ respectively). Lindsey's and Pankin's functions performed quite well ($R^2 = .928$ and $R^2 = .937$ respectively), while Bennett and Flueck's function fit runs scored best ($R^2 = .950$). The weakness of this approach was that because it was purely driven by data within a specific time period, the function was not flexible to different eras. Further, extraneous effects such as ballpark effects and different competition levels were not taken into account.

Berry (2000) also studied the relative worth of various hitting statistics. In addition to the shortcomings listed in the introduction, Berry also found shortcomings in slugging percentage and on-base percentage. Slugging percentage over-valued extra base hits and did not account for walks. On-base percentage accounted for walks but did not account for the extra value of extra base hits. Berry also performed a regression analysis to which he could compare the current traditional baseball function. He used the 1990 to 1998 season team statistics with runs scored being a function of 1B, 2B, 3B, HR, SB, CS, BB, K, and Outs. The Berry measure was:

$$\begin{aligned}Berry\ Value &= 0.34BB + 0.49(1B) + 0.72(2B) + 1.14(3B) + 1.51HR \\ &+ 0.26SB - 0.14CS - 0.10(OUT + K)\end{aligned}$$

Although not taken into account in his model, Berry realized that these coefficients would change over time. He combined team data over five year periods starting in

1900 and calculated the coefficients of each term in the above function, except for strikeouts and stolen bases. He found that the coefficients for walks and outs were the most stable over time, while triples and home runs were the least stable.

Berry, Reese, and Larkey (1999) compared the performances of athletes from different eras in baseball, hockey, and golf. In order to make this comparison, they made a bridge of player comparisons (older players from one era competing against younger players in another era) so that there was a link between different eras. Berry, et al used the 1998 season as a benchmark and created a bridge by comparing performances of current players and past players whose careers overlapped. Because players played against each other at different ages, individual-level aging curves were added to estimate the change in player performance as they aged. The basis of their model was a decade-specific Hierarchical Bayes Model where they built in an indicator for seasonal effects, ballpark effects, and the aging function. The goals for the Hierarchical Bayes Model were to discern the effect of aging, to discover the effect of era, to find out if the "hot hand" effect exists, and to characterize the talent of each player. They analyzed batting average and home runs to rank baseball player performance. Unlike some past researchers, the model included data on every non-pitcher who batted in Major League Baseball from 1900 to 1998, including year-of-birth, home ballpark, at-bats, hits, and home runs. Their models used batting average (hits per at-bat) and home run average (home runs per at-bat) as their offensive production measures. In the dataset of 7031 players, x_{ij} was the number of hits for the ith player in his jth season, h_{ij} was the number of home runs hit by the ith player in his jth season, m_{ij} was the number of at-bats, a_{ij} was the player's age in that season, y_{ij} was year of play (1900-1998), and t_{ij} was player's home ballpark for that season. Looking at just the batting average model and using π^a_{ij} as the probability of getting a hit for the ith player in his jth season,

$$x_{ij} \sim Binomial(m_{ij}, \pi^a_{ij})$$

where

$$\log(\frac{\pi^a_{ij}}{1-\pi^a_{ij}}) = \theta^a_i + \delta^a_{y_{ij}} + \xi^a_{t_{ij}} + f^a_i(a_{ij})$$

and

$$\begin{aligned} &\theta^a_i \sim Normal(\mu^a_\theta(d_i), (\sigma^a_\theta)^2(d_i)) \\ &\mu^a_\theta(d_i) \sim Normal(m^a, (s^a)^2) \\ &(\sigma^a_\theta)^2(d_i) \sim InverseGamma(a^a, b^a) \end{aligned}$$

where θ^a_i was the decade-specific conditionally-independent hierarchical model, $\delta^a_{y_{ij}}$ was the season indicator, $\xi^a_{t_{ij}}$ was the home ballpark indicator, $f^a_i(a_{ij})$ was the aging function, m^a was the known mean of the prior mean distribution, s^a was the known standard deviation of the prior mean distribution, a^a was the known alpha parameter of the prior standard deviation distribution, and b^a was the known beta parameter of the prior standard deviation distribution. Berry, Reese, and Larkey's paper provided a strong basis on which to build a statistical model for comparing baseball player performance.

Section 2.2 Bayesian Models

In this paper, we are attempting to model players' performances from different years, eras, and competition levels. Baseball is a game of constant change. Therefore, the data collected from year to year is similar to gathering data from different studies. Draper, Gaver, Goel, Greenhouse, Hedges, Morris, Tucker, and Waternaux (1992) suggested that the best way to get estimates from heterogeneous sources of data is to use a BHM.

The foundation of Bayesian statistics is the belief that unknown parameters are random variables, not fixed values. These parameters can be approximated by imposing a distribution on them, based on our prior knowledge of the parameter. When we do not know much about the parameter, we can impose a relatively flat distribution (such as the uniform distribution) on the parameter over a range of acceptable values. Combining the likelihood from the data and the prior distribution, Bayes Theorem can be used to create the posterior distribution. Bayes' Theorem is:

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{\int P(X \mid \theta)P(\theta)d\theta} = \frac{P(X \mid \theta)P(\theta)}{P(X)}$$

where $P(\theta \mid X)$ is the posterior distribution of the parameters, θ, given the data, X, $P(X \mid \theta)$ is the likelihood of the data given the parameters, and $P(\theta)$ is the prior distribution of the parameters**.** If the prior is conjugate with respect to the likelihood, a closed form expression of the posterior distribution is available. A distribution class of prior distributions is said to be conjugate if the posterior distribution is in the same class of distributions for all choices of x. For example, if the likelihood is normal, then the conjugate prior in the normal family results in a normal posterior distribution. If the prior distribution is beta and the likelihood is binomial, then the posterior will be beta with updated parameters. (Lee, 1997)

Efron and Morris (1972) presented a example of a Bayesian model and the estimates that result from that model. In their paper, Efron and Morris looked into the properties of Bayes estimators, using the following Bayesian model as an example. If we assume that the variance is known, we have the following likelihood and prior distributions:

$$X_j \sim Normal(\theta, \sigma^2)$$
$$\theta \sim Normal(\mu, \tau^2)$$

From this model, the following is the posterior distribution,

$$\theta \mid X \sim Normal(\frac{(1/\tau^2)\mu + (n/\sigma^2)\overline{x}}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2})$$

where $\overline{x} = \sum_{j=1}^{n} x_j / n$.

According to Lindley and Smith (1972), Bayes' estimates performed better than the frequentist least squares estimates when estimating the expected value for general fixed effect linear models because the mean squared error for the Bayes estimate was smaller than the mean squared error for the least squares estimate. This conclusion holds only if the prior distributions are exchangeable (the prior distribution does not change for all i) and we are looking at estimates over multiple samples (n>4).

Bayes models have been used in the past to model baseball data. Sobel (1993) used World Series batting averages from 1948-1955 as the data to make comparisons between parameters from different populations. He created parameters to compare ranks based on maximum likelihood and Bayes estimates. By analyzing the first 45 at-bats and the final batting averages in the 1970 season for a group of hitters, Efron and Morris (1975) were able to compare Bayes estimators created by Stein to traditional maximum likelihood estimation methods. For each batter, Efron and Morris illustrated that the Bayes estimator outperformed the maximum likelihood estimate.

Section 2.2.1 Application of Bayesian Statistics to Baseball Problem

In 2001 Barry Bonds set a new major league home run record with 73 home runs in one season. Considering Barry Bonds' home run rate based just on the 2001 baseball season may bias the analyst's perception of Barry Bonds' true home run rate is: additional historical information would be beneficial to a complete analysis. Because of the ability to add additional information using a prior distribution, this case lends itself to the use of Bayesian data analysis.

In setting up this problem, we have data from all 159 games that Barry Bonds played in the 2001 season. For each game, we have the number of home runs he hit and the number of his plate appearances in that game. At each plate appearance, Bonds had a specific probability of hitting a home run. The probability of Bonds hitting a home run in a particular plate appearance was assigned the value θ, and the probability of that he would not hit a home run was (1- θ). He either hit a home run (y=1) or didn't (y=0). Each plate appearance has a Bernoulli distribution, P(y) = θ^{y} $(1-\theta)^{(1-y)}$, where y can be either 1 or 0. If we look at each plate appearance as an independent and identically distributed Bernoulli distribution, the number of home runs Bonds hits in a season is a Binomial distribution with the two parameters being the total number of at-bats in the season and Bonds' home run rate. In choosing the likelihood for the data, we need to focus on whether the binomial distribution is appropriate. The home run rate has a probability between 0 and 1. Therefore, we would expect the distribution of home run rate to not be symmetrical at the low and high ends. This is caused by the distribution of rates with high or low probability hitting the bounds of 0 and 1. Of the distributions in the exponential family, the Binomial distribution fits these properties the best. In using this distribution, we are assuming that each plate appearance is an independent and identically distributed event. For the purposes of this example, we are not taking into account the time of season, different pitchers Bonds faces, different ballparks he plays in, or the different teams he is playing. Those variables are addressed in the larger scope of this thesis project.

To ease the computation of the posterior distribution moments, it is helpful to choose a prior such that the posterior distributions stay in a closed exponential form. A prior that has this property is called a conjugate prior. In order to formulate the conjugate prior distribution, we first need to recognize the features of the distribution of y (number of home runs Barry Bonds hits in a season).

$$P(y \mid \theta) = \frac{N!}{y!(N-y)!}(\theta^{y}(1-\theta)^{N-y})$$

We can break this down into a function of y, a function of θ, and a combined function of y and θ in the exponential family as follows:

$$P(y \mid \theta) = f(y)g(\theta)\exp(\phi(\theta)^{T}u(y)$$

Using the distribution of y, $f(y) = \dfrac{N!}{y!(N-y)!}$, $g(\theta) = \theta^{y}$, $\phi(\theta) = log(1-\theta)$, *and* $u(y) = N\text{-}y$.

From this form, we can now create the conjugate prior distribution which is of the following form:

$$\exp(\phi(\theta)^{T} u(y) = \theta^{y}(1-\theta)^{N-y}$$

This function looks very similar to a Beta distribution with density

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Therefore, the Beta distribution is the conjugate prior distribution for a Binomial likelihood and the posterior distribution will be in the Beta family.

The question is which Beta distribution should we use in this problem? One approach to eliciting the prior distribution parameters is to look at the moments of the Beta distribution. The first moment (or expected value) of the Beta distribution is equal to a/(a+b). Therefore, we used an approximation of Barry Bonds' home run rate as our expected value. The second moment (or the variance) of the Beta distribution is related to the size of a and b. The larger the size of a and b, the more confidence we have in the expected value of Bonds' home run rate and the smaller the variance of the Beta distribution. For example, if the approximated expected value of Barry Bonds home run rate is one out of every 10 at-bats, or .1, but we are not very confident in this value, we could use a = 1 and b = 9. If the approximated expected value of Barry Bonds' home run rate is about 15 percent and we are very confident in this, we could use a = 150 and b = 850.

The first thing to do to calculate the posterior distribution is to set up the joint distribution of Y and θ, which is the product of the prior distribution of Θ and the likelihood of Y| θ . If the prior distribution is Beta(1,9), then

$$
\begin{aligned}
p(y,\theta) &= p(\theta)p(y\,|\,\theta) \\
&= \frac{\Gamma(10)(1-\theta)^{8}\dfrac{N!}{y!(N-y)!}\theta^{y}(1-\theta)^{N-y}}{\Gamma(1)\Gamma(9)}
\end{aligned}
$$

The posterior distribution of Barry Bonds' home run rate is the distribution of θ given the data y. This is equal to the joint distribution of y and θ divided by the marginal distribution of y or the joint distribution of y and θ divided by the joint distribution of y and θ integrated with respect to θ. That is,

$$
\begin{aligned}
P(\theta\,|\,y) = \frac{P(y,\theta)}{P(y)} &= \frac{P(y,\theta)}{\int P(y,\theta)d\theta} \\
&= \frac{\dfrac{\Gamma(10)(1-\theta)^{8}}{\Gamma(1)\Gamma(9)}\dfrac{N!}{y!(N-y)!}\theta^{y}(1-\theta)^{N-y}}{\int\dfrac{\Gamma(10)(1-\theta)^{8}}{\Gamma(1)\Gamma(9)}\dfrac{N!}{y!(N-y)!}\theta^{y}(1-\theta)^{N-y}d\theta} \\
&= \frac{\Gamma(y+1+N-y+9)\theta^{y}(1-\theta)^{N-y+8}}{\Gamma(y+1)\Gamma(N-y+9)}
\end{aligned}
$$

Applying the results of the data (y = 73 and N-y = 403), we get a posterior distribution of Beta(74,412).

We can also do the same calculations to compute the posterior distribution using a more informative prior distribution of Beta(150,850). Then,

$$p(y,\theta) = p(\theta)p(y \mid \theta) \; = \frac{\Gamma(1000)\theta^{149}(1-\theta)^{849}(N,y)\theta^{y}(1-\theta)^{N-y}}{\Gamma(150)\Gamma(850)}$$

$$p(\theta \mid y) = \frac{\Gamma(y+150+N-y+850)\theta^{y+149}(1-\theta)^{N-y+849}}{\Gamma(y+150)\Gamma((N-y)+850)}$$

For the more informative prior, we get a new posterior distribution of Beta(223,1253). These Beta Distributions have the summary statistics in Table 2.2.

Table 2.2: Beta Distribution (74,412) and Beta Distribution (223,1253). Summary Statistics used to compare the impact of different priors on the posterior distributions.

Mean	0.1523	0.15108
Variance	0.0003	0.00009
Standard Deviation	0.0163	0.00932
Mode	0.1508	0.15061
Median	0.1518	0.15093

By comparing these two different posterior distributions, we can come to some conclusions based on the data and the choice of prior distribution. First, the more informative the prior distribution, the stronger the prior distribution's impact will be on the center of the posterior distribution. Second, the more informative the prior distribution, the smaller the variance is in the posterior distribution. The inverse is also true. The weaker the prior, the closer the posterior distribution is to the data (mean = .15336).

Section 2.2.2 Hierarchical Bayes Models

The Hierarchical Bayes Model is called "hierarchical" because it has multiple levels of connected prior distributions. The parameters of the prior distribution have their own prior distributions, known as hyperpriors (Lee, 1997). Using hyperpriors allows for the use of information from multiple levels of observational units and allows for the exchangeability of parameters in different but related studies. For example, if we were to perform a study on whether aspirin is an appropriate choice of treatment for post-heart attack patients, we would want to find out if there is a difference between the death rate in the aspirin group and the death rate in the placebo group. Multiple replications of this study would be performed at different places and times. We would want to combine the information gathered from each of these study locations. This is called a multi-station clinical trial. Knowing that the studies are different, we could not aggregate the data directly. To get estimates with heterogeneous studies, we would need to account for the differences while modeling the similarities.

The hierarchical model has two stages, the individual stage and the general stage. The first or individual stage of this hierarchical model would be one that states that the number of people who died when taking aspirin in a particular study (x_i) is Binomially distributed with a certain death rate proportion (π_i) and the number of people in the study (n_i) as the parameters. That is,

$$x_i \sim Bin(n_i, \pi_i)$$

with π_i having

$$\pi_i \sim Beta(\alpha, \beta).$$

The general stage of this hierarchical model would be that the prior distribution parameters of the individual study death rate have a prior distribution (or a hyperprior of the original distribution). Thus, we could assume that the α and β parameters have distributions as follows:

$$\alpha \sim Gamma(a_\alpha, b_\alpha) \text{ and}$$

the β parameter with

$$\beta \sim Gamma(a_\beta, b_\beta)$$

where a_α, b_α, a_β,and b_β are all specified constants.

In this hierarchical model, we assume that each study comes from a population of studies which has a certain distribution, instead of being independent. The estimates for a particular study borrow strength from the other studies. Each particular study helps to estimate its own parameters and also the general parameters. This is because each of the other studies gives us more information about the particular study for which we want to find estimates. In the simplest case (a linear fixed effect model), the estimates from a hierarchical model are a weighted average between the individual study values and the overall mean of the combined studies. (Draper et al, 1992)

Hierarchical Bayes models have been used to analyze different aspects of the game of baseball. Using player data from the 1988 to 1992 seasons, Albert (1994) used Hierarchical Bayes models to see which situational variables (home vs. away, grass vs. turf, etc.) explained a significant amount of variation in a player's hitting. The model he proposed was for h_{ij}, the number of hits by player i in the jth situation, and o_{ij}, the number of outs by player i in the jth situation:

$$\begin{aligned} y_{ij} &= \log(\frac{h_{ij}}{o_{ij}}) \\ y_{ij} &\sim Normal(\mu_{ij}, \sigma_{ij}^2) \\ \mu_{ij} &= \mu_i + \alpha_{ij} \end{aligned}$$

where

$$\alpha_{ij} \sim Students\ t(\mu_{\alpha}, \sigma_{\alpha}^{2}, \upsilon)$$
$$\sigma_{\alpha}^{2} \sim Inverse\ Gamma(k, \theta)$$

and μ_{α} and μ_{i} have a uniform distribution. Looking at the posterior means from this Hierarchical model, the "ahead in the count vs. two strikes in the count" effect is the largest and most significant with the median of the batting average differences of 123 points (on a realistic scale of 0 to 425 points).

Chapter 3.0

The Methodology and Approach of This Project

Section 3.1 Validating Berry Value over 1900 - 1998

The purpose of this chapter is to fit a simple BHM for the Berry measure of offensive capacity of a major league baseball player. By building a model with just one player, we investigate the properties of the Berry measure and verify the presence of an age effect. Before pursuing the model, it is interesting to compare the Berry measure to five other summary statistics of a baseball player's offensive productivity. In building our BHM model, we want to use the offensive productivity summary statistic that is most highly correlated with the amount of runs scored. Using the individual player data from 1900-1998, we collapse the data by year and team so that we have team totals for each year from 1900 to 1998. Each of the six summary statistics is also calculated for each team from 1900 to 1998. The resulting correlations of the functions on runs scored are in Table 3.1.

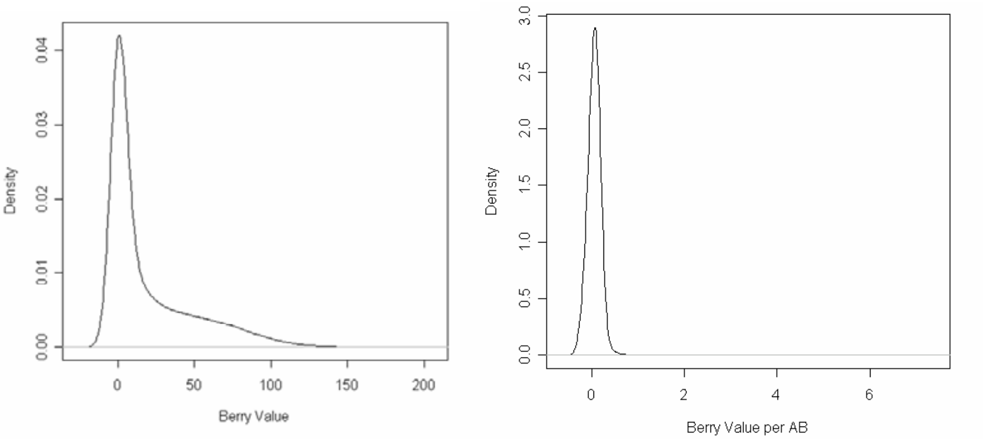
Table 3.1: Six Key Offensive Productive Summary Statistics' Correlations with Runs Scored on a team level from 1900 to 1998 which will help us determine the best offensive production summary statistic.

	Correlation
OPS	85.62%
Lindsey	87.67%
OPA	87.78%
OPA3	85.52%
ERP	90.38%
Berry	96.14%

Table 3.1 shows that Berry's function is the most highly correlated with runs scored at the team level (r = 96.14%).

In order to identify a reasonable distribution of the Berry measure, consider kernel density estimators on the Berry measure values calculated at an individual level for all of the players over each year.

Figure 3.1: Berry's Measure and Berry's Measure per at bat Density Curve for all players over all seasons. The Berry measure is very right skewed and varies between roughly -25 and 150.



From Figure 3.1, the Berry Value is very right skewed because it is highly correlated with the number of at-bats a player received. The more at-bats a player receives, the more runs he may produce. Another perspective is to standardize the Berry Value to a "per at bat" value as also seen in Figure 3.1. It should be noted that the Berry measure can result in a negative value when an individual player creates significantly more outs than he does hits. This result means that the player has a negative impact on his team scoring runs. From this point forward, we will focus on the raw Berry Value. The first reason for using the raw Berry Value is that the number of at-bats is a very important indicator of a player's value. The more games an above average player plays, the more the team benefits. The second reason for using raw Berry Values is that the number of games played is also clearly a function of age. Generally the older a player gets, the more his injuries and declining skills impact playing time.

Section 3.2 Hierarchical Bayes Model using the Berry Value

As stated above, the Berry Value is the baseball measure most highly correlated with runs scored at the team level. Also seen above, the distribution of the Berry Value is right skewed and the values tend to fall between -20 and 150 with the mean being around 40. We assume that within each season and on an individual level, the distribution of Berry Value is approximately Normal. In this example, we choose one player's Berry Values to analyze. Mickey Mantle was chosen for the model because of his long career (18 years) and his highly recognized name. We model his Berry Values as having a different mean per season. We assume that Mickey Mantle's distribution of mean Berry Values for each season are Normal distributions with parameters θ and σ_θ^2. These parameters also have their own prior distributions. Following are the distributions of the likelihood and prior distributions with *i* indicating the individual season of Mantle's Berry Values, using the convention Normal(μ,τ) where μ is the mean and τ is the precision (the inverse of the variance):

$$y_i \mid \mu_{i,}\ \sigma^2 \sim Normal(\mu_{i,}\ \frac{1}{\sigma^2})$$

$$\mu_i \mid \theta,\ \sigma_\theta^2 \sim Normal(\theta,\ \frac{1}{\sigma_\theta^2})$$

$$\sigma^2 \sim Inverse\ Gamma(2,100)$$

$$\theta \sim Normal(80,.01)$$

$$\sigma_\theta^2 \sim \mathit{Inverse\ Gamma(2,}100\mathit{)}$$

These prior distributions were created based on the assumptions of the general distribution of standardized Berry Values for all players. We also assumed that means generally follow a normal distribution and variances generally follow an inverse gamma distribution. It is important to note that these are conjugate prior distributions if we look at the marginal posterior distributions. Now with the prior and likelihood distributions of 18 seasons of Berry Values, we build the unnormalized joint posterior distribution of the data and parameters.

$$p(y_i,\mu_i,\sigma^2,\theta,\sigma_\theta^2) \propto (\sigma_\theta^2\sigma^2)^{-12}\exp(-\frac{(\theta-80)^2}{200}-\frac{100}{\sigma_\theta^2}-\frac{100}{\sigma^2}-\frac{\sum_{i=1}^{18}(\mu_i-\theta)^2}{2\sigma_\theta^2}-\frac{\sum_{i=1}^{18}(y_i-\mu_i)^2}{2\sigma^2})$$

Using the unnormalized joint distribution, we calculate the conditional distributions for each of the parameters (at least to a constant of proportionality). We used WinBUGS software to run our BHM, which uses Gibbs Sampling to iteratively sample each conditional distribution. If the conditional is in a closed form, WinBUGS recognizes the conjugate specifications and directly samples from the conditional. If it is not in closed form, WinBUGS uses adaptive rejection sampling to update the conditionals. WinBUGS allows the user to specify the likelihood of the data and each prior distribution, then it calculates the conditionals. Even though WinBUGS calculates this for us, it is important to understand the conditional distributions. The conditional distribution is identical to the joint distribution except that it only includes the terms dependent on that one parameter. The following are a couple of the conditional distributions:

$$p(\theta \mid y_i, \mu_i, \sigma^2, \sigma_\theta^2) \propto \exp(-\frac{(\theta-80)^2}{200} - \frac{\sum_{i=1}^{18}(\mu_i-\theta)^2}{2\sigma_\theta^2})$$

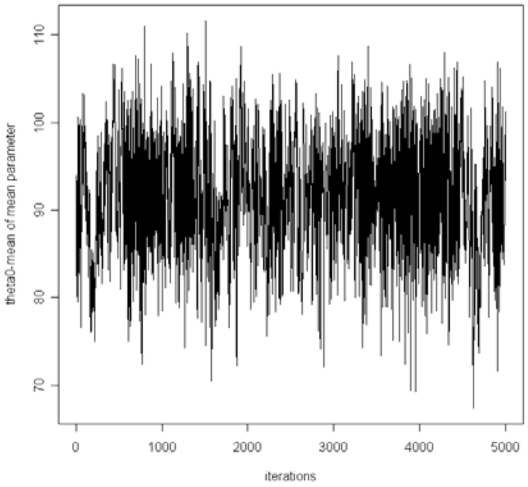
$$p(\sigma_\theta^2 \mid y_i, \mu_i, \sigma^2, \theta) \propto (\sigma_\theta^2)^{-12} \exp(-\frac{100}{\sigma_\theta^2} - \frac{\sum_{i=1}^{18}(\mu_i-\theta)^2}{2\sigma_\theta^2})$$

As seen above, each conditional distribution has multiple parameters. Therefore, it is impossible to estimate the distribution of each parameter without knowing the values of the other parameters. The Gibbs Sampler allows us to get estimates using a Markov Chain simulation. In the Gibbs Sampler, we must first choose starting values and candidate distributions for the conditional distributions that are not in a closed form. For WinBUGS, the user has the option to specify the starting values. If WinBUGS needs to use the Metropolis-Hastings algorithm within the Gibbs Sampler, it will choose its own candidate distributions. The starting values for all of the parameters are based on the means or expected values of their prior distributions. For example, we chose a starting value of θ and each μ_i of 80 because that is the mean of their prior distributions. After enough iterations (in this case we used 100,000), we are able to converge on the true joint posterior distribution.

The complicating factor of the Gibbs Sampler in this situation is the Hierarchical Bayes element. First, we must loop through each conditional distribution for the prior of the standard deviation. Then we must also loop through the conditional distribution for each season mean. As a result, we have 18 times more iterations than a model with a generic overall mean. After running all iterations, we have posterior distributions for each parameter to analyze.

Having run all of the iterations, we determine whether or not we converged onto the correct distributions for the parameters. It is common in MCMC algorithms for the parameters to take a certain number of iterations to stabilize, especially when we have to approximate distributions using Metropolis-Hastings. A trace plot is an important tool used to check the stability and convergence of parameters. The trace plot tracks the accepted samples throughout the iterations. Figure 3.2 details the trace plot of the last 5000 iterations of the mean of the season parameters in the Mickey Mantle problem. Several criteria exist for examining the parameter trace plots that help determine convergence. When the slope of the plot is relatively flat, and no obvious trends are present, then the parameters have converged.

Figure 3.2: Trace Plot of mean of the season parameter from the Mickey Mantle problem which shows a constant and converged value of around 90.



This trace plot shows that while there is a large amount of variance in the value of this parameter, it seems to be centered around a value: there are no obvious trends or slope. We conclude that the parameters have converged. However, we should keep in mind the high variability of these estimates.

After running the Hierarchical Bayes Gibbs Sampler that is shown in Appendix A, we can make some conclusions on the posterior distributions of the main parameters (the μ_is and σ^2). Table 3.2 includes some moments from these distributions.

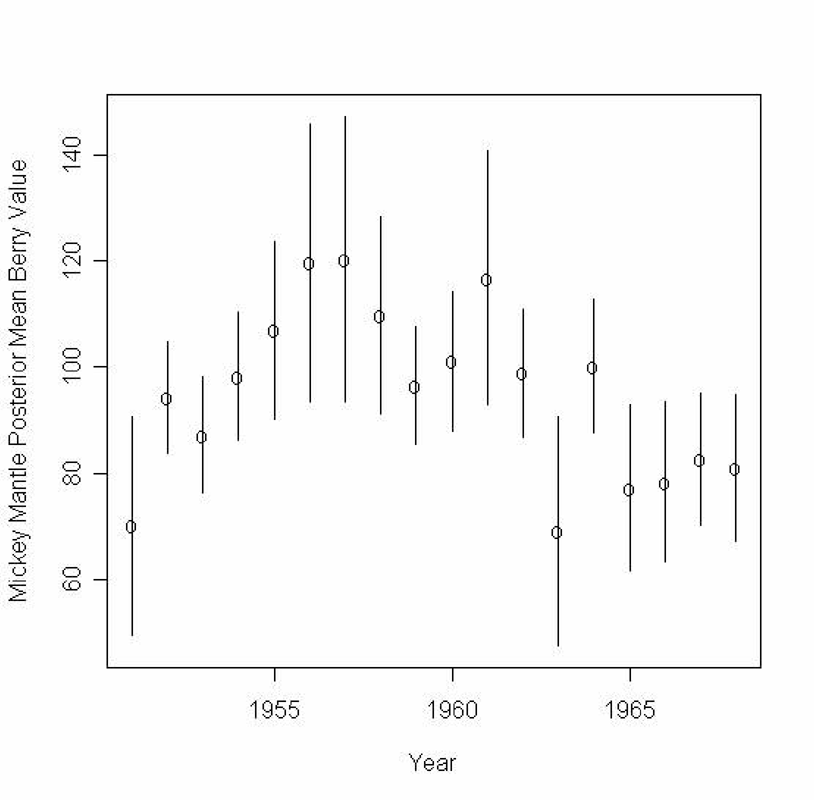
Table 3.2: Summary Statistics from Mickey Mantle problem showing the mean and standard deviation of all parameters which shows a general age effect and a large overall variance.

Variable	Mean	Std. Deviation
mu[1]	70.09	20.57
mu[2]	94.34	10.59
mu[3]	87.16	10.91
mu[4]	98.28	12.00
mu[5]	106.9	16.79
mu[6]	119.7	26.19
mu[7]	120.4	26.86
mu[8]	109.8	18.62
mu[9]	96.52	11.07
mu[10]	101.1	13.22
mu[11]	116.8	23.94
mu[12]	98.91	12.09
mu[13]	69.11	21.69
mu[14]	100.2	12.58
mu[15]	77.19	15.67
mu[16]	78.37	15.07
mu[17]	82.75	12.46
mu[18]	81.02	13.84
1/sigma	0.009012	0.01246
1/sigma0	0.00935	0.0121
theta	91.61	6.26

Figure 3.3 plots the posterior mean estimates of μ_i and the standard deviation of these means from Table 3.2. Season does have a significant impact on the Berry Values for Mickey Mantle. It seems that Mickey Mantle's Berry Values significantly increased over the beginning of his career, stabilized in the middle, and then decreased steadily until his retirement. The rate of increase in the beginning of

his career is steeper than the rate of decrease later in his career. There are some years that seem to be outliers. For example, Mantle averaged between 120 and 140 games per season, but in season 13 Mantle was injured with a broken foot and only played in 65 games. Apart from this outlier, the general trends from the Mickey Mantle model show a quadratic kind of effect. This validates the use of a quadratic age curve for the overall model. We will use a piecewise quadratic function for the age effect in our overall model, to take into account the different slopes of maturation and decline. According to the table and chart, we see that there is plenty of variability in the data. The overall variance is close to 100. The variances of the season means between 80 and 100 are much smaller than those outside of that range.

Figure 3.3: Chart of Mickey Mantle's Berry Values Posterior Means by Season with standard deviation included which shows a visual of the quadratic age effect and the large variability at extreme values.



Chapter 4.0

Project Setup and Results

The goal of this project is to create a Hierarchical Bayes model that will permit the linking of players from different eras, which enables us to calculate each player's innate value. We follow the paradigm of Berry, Reese, and Larkey (1999) and use Hierarchical Bayes models with season and age effects to define a player's overall offensive value based on Berry Value, a summary statistic highly correlated with run scoring. We have shown that a Hierarchical Bayes model on one individual player can be built and the year-to-year effects can be investigated. With one player, the year to year effect confounds the player's changing performance over their career and the differences in the game over different eras. This section fits a Hierarchical Bayes model with a season effect and a function for the physical age of a player. We will apply this model to all who played Major League Baseball between 1900 and 2004. In Section 4.1, we define the model and explain the joint posterior and conditional distributions. We also discuss the years in which dramatic changes have occurred in the game of baseball. In Section 4.2, we discuss methods of data cleaning and preparation. In Section 4.3, we determine the validity of the distributions. In Section 4.4, we discuss model results, particularly the season and age effects. We will discuss whether the impact of these changes can be seen in the seasonal effects of the model. We then look more closely at the innate player value. By calculating the innate player value, we can compare a player from past years like Babe Ruth, to a current player, like Barry Bonds, in the same context.

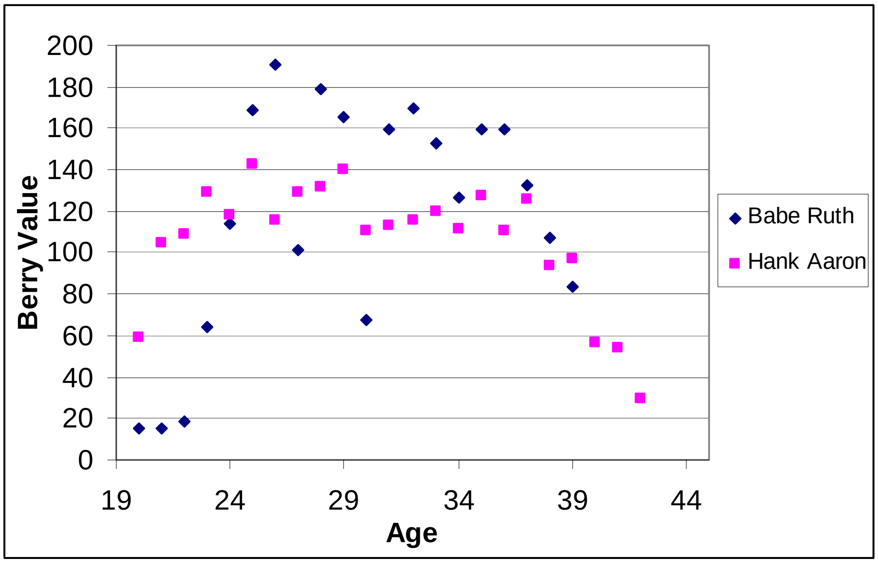
Section 4.1: Setting Up the Problem

The Berry Value is the summary statistic most highly correlated with runs scored at the team level. It is also the best value used to measure a baseball player's offensive performance. The following is a discussion of the Berry Value including some descriptive statistics. The higher the Berry Value, the better the player is overall. The minimum Berry Value is –9 and the maximum is 213, with the average around 44. Only about one percent of Berry Values are negative. Most current superstars and inductees to the Hall of Fame have average career Berry Values around 100. For example, some career mean Berry Values are Babe Ruth (117), Barry Bonds (133), Hank Aaron (105), Ted Williams (119), Lou Gehrig (140), and Tony Gwynn (80).

We make two assumptions in building our model. First, the Berry Value for a player in one year does not affect the Berry Value of the same player in the next year. The outcomes are independent. Second, the effects in the model do not have any impact on each other. There are no interactions.

We model each player's Berry Values in two different parts. The first is a player level age function. Based on the analysis of Mickey Mantle's career is Section 2.3 and further supported by Berry Values by age for Hank Aaron and Babe Ruth (which seem to be typical of most players) in Figure 4.1, we notice that players have a period of maturation where their value is increasing at the beginning of their careers. Then a player's value seems to peak and stays relatively constant for a period. At the end of a player's career, his Berry Value hits a period of decline. It can also be noted from the above two charts that the period of maturation and period of decline for each player can be different in size and magnitude.

Figure 4.1: Berry Value vs. Age Plot for Hank Aaron and Babe Ruth which shows the general age effect from the raw Berry Values.



Because of the increasing nature of the period of maturation and the decreasing nature of the period of decline, the Berry Value can be modeled by a quadratic function in the most simplistic form. One drawback of using a quadratic function is that the rate and interval of the period of maturation will be symmetric to the rate and interval of the period of decline. A piecewise continuous quadratic function is proposed to better approximate the aging curve.

Relying on Berry, et al. (1999) conclusions and some basic regression analysis on the overall data, the peak of the age distribution is found to be between 27 and 30 years. We have chosen 29 years as the knot for our piecewise age effect. The age function of our model therefore is: $\beta_{0i} + \beta_{1i}(a_{ij}) + \beta_{2i}(a_{ij})^2 + \beta_{3i}z_{ij}(a_{ij} - 29)^2$

where $z_{ij} = 1$ if age ≥ 29 and $z_{ij} = 0$ otherwise. It should be noted that at age 29 the age function is continuous and has constant slope.

The second part of our model considers the effect for each season denoted by δ_j. The δ_j captures the effect due to changes in rules, quality of competition, and other differences found over time. A good approximation for the time placement of the "era" effect is to use the years of major rule changes in Major League Baseball. Over the years, Major League Baseball has made rule changes that affect the way the game is played. Certain events were recorded differently and the competitive balance of the game itself was altered.

In the early 1900s, Major League Baseball made various adjustments in how the games are recorded. For example, in 1903, the foul strike rule was adopted. Before 1903, when a batter hit a foul ball, it did not count as a strike. The foul strike gave hitters a slight disadvantage, because if they hit a foul it brought them closer to three strikes and an out. In 1907, the sacrifice fly rule was adopted; when a batter hit a fly ball out, but it drove in a run, it was not counted as an at-bat. The sacrifice fly gave hitters a slight advantage because the out was not charged as an at-bat. In 1920, many rule changes occurred. The RBI became an official statistic. A home run was given to a batter in the bottom of the ninth if the player who scored the winning run was on base. The spitball and other freak deliveries were also outlawed since they gave the pitchers a distinct advantage. The most monumental change in Major League Baseball occurred on April 15, 1947, the date Jackie Robinson broke the color barrier. The inclusion of African-Americans in Major League Baseball dramatically increased the talent level in the game.   In 1959, Major League Baseball set minimum ballpark boundary regulations, affecting the hitters because it set a minimum distance for a batted ball to be counted as a home run. In 1969, hitters gained a huge advantage when Major League Baseball dropped the height of the pitcher's mound by 5 inches and shrank the strike zone. During most of the 60s, umpires called a large strike zone (from the top of the shoulders to the bottom of the knees). In 1969, this strike zone was rescinded and the traditional strike zone (between the armpits and top of the knee) was reinstated. This rule change put the pitchers on a more even plane with the hitters and a pitched ball did not have the extra advantage of gravity when it was pitched. The shrunken strike zone also gave the hitters an advantage because they could be more selective. In 1972, American League hitters gained an added benefit when its commissioner adopted the designated hitter rule. The hitting ability of a team drastically increased when the pitcher was replaced by a designated hitter in the lineup. The probability of a hitter getting an RBI or scoring a run increases when the hitting ability of the teammates surrounding that hitter increases. In 2001, the hitters were put back at a disadvantage when the strike zone was altered vertically. Over time, the strike zone had become smaller. The umpires were asked to call the traditional strike zone, which meant to call more high strikes and inside strikes. This increased the area into which a pitcher can aim and pitch a strike. The timing of these rule changes gives us chronological breaking points by which we can separate our era or time period functions. The severity of these rule changes demonstrates that an era effect is indeed necessary to model the changes these rules bring to the game.

We assume that the distributions of coefficients for the age function and overall season effect have their own normal distributions with distinct hyperparameters. These hyperparameters also have unique prior distributions. The following are the distributions of the likelihood and prior distributions where y_{ij} is the Berry Value for the ith player playing in the jth season; using the convention Normal(μ,τ) where μ is the mean and τ is the precision (the inverse of the variance)

$$y_{ij} \mid \beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i}, \delta_j, \sigma^2 \sim \mathrm{Normal}(\beta_{0i} + \beta_{1i}(a_{ij}) + \beta_{2i}(a_{ij})^2 + \beta_{3i}\, z_{ij}(a_{ij} - 29)^2 + \delta_j, \frac{1}{\sigma^2})$$

$$\beta_{0i} \mid \theta_{\beta 0}, \sigma^2_{\theta\beta 0} \sim Normal(\theta_{\beta 0}, \frac{1}{\sigma^2_{\theta\beta 0}})$$

$$\beta_{1i} \mid \theta_{\beta 1}, \sigma^2_{\theta\beta 1} \sim Normal(\theta_{\beta 1}, \frac{1}{\sigma^2_{\theta\beta 1}})$$

$$\beta_{2i} \mid \theta_{\beta 2}, \sigma^2_{\theta\beta 2} \sim Normal(\theta_{\beta 2}, \frac{1}{\sigma^2_{\theta\beta 2}})$$

$$\beta_{3i} \mid \theta_{\beta 3}, \sigma^2_{\theta\beta 3} \sim Normal(\theta_{\beta 3}, \frac{1}{\sigma^2_{\theta\beta 3}})$$

$$\delta_j \mid \theta_{\delta j}, \sigma^2_{\delta} \sim Normal(\theta_{\delta}, \frac{1}{\sigma^2_{\delta}})$$

$$\sigma^2 \sim Inverse\ Gamma(2,\ 25)$$

$$\theta_{\beta 0} \sim Normal(0,.01)$$

$$\sigma^2_{\theta\beta 0} \sim Inverse\ Gamma(2,\ 100)$$

$$\theta_{\beta 1} \sim Normal(0,.02)$$

$$\sigma^2_{\theta\beta 1} \sim Inverse\ Gamma(2,\ 64)$$

$$\theta_{\beta 2} \sim Normal(0,.1)$$

$$\sigma^2_{\theta\beta 2} \sim Inverse\ Gamma(2,\ 9)$$

$$\theta_{\beta 3} \sim Normal(0,.25)$$

$$\sigma^2_{\theta\beta 3} \sim Gamma(2,\ 4)$$

$$\theta_{\delta} \sim Normal(0,.02)$$

$$\sigma^2_{\delta} \sim Gamma(2,\ 36)$$

These prior distributions were not created based on any assumptions about the general distribution of Berry Values for all players. We can make assumptions about the directionality of the mean parameters for the prior distributions, based on the need for our individual age curves to peak and be concave down. We do not have enough evidence however, to make any conclusions on the magnitude of this directionality. All of our prior distributions of the mean parameters will therefore have a zero mean.

Now that we have the prior and likelihood distributions, we create the unnormalized joint posterior distribution by multiplying the likelihood and all prior distributions (assuming independence). The non-normalized joint posterior distribution follows where n denotes the number of players in the model (5393), m denotes the total seasons in the model (105), and m_i denotes the number of seasons that each individual player plays:

$$p(y_{ij},\beta_{0i},\beta_{1i},\beta_{2i},\beta_{3i},\delta_j,\theta_{\beta 0},\sigma^2_{\theta\beta 0},\theta_{\beta 1},\sigma^2_{\theta\beta 1},\theta_{\beta 2},\sigma^2_{\theta\beta 2},\theta_{\beta 3},\sigma^2_{\theta\beta 3},\theta_\delta,\sigma^2_{\theta\delta},\sigma^2)$$

$$\propto \sigma_{\theta\beta 0}^{-n-6}\sigma_{\theta\beta 1}^{-n-6}\sigma_{\theta\beta 2}^{-n-6}\sigma_{\theta\beta 3}^{-n-6}\sigma_{\theta\delta}^{-m-6}\sigma^{(-\sum_{i=1}^{n} m_i)-6}$$

$$\exp(-\frac{25}{\sigma^2}-\frac{36}{\sigma^2_{\theta\delta}}-\frac{100}{\sigma^2_{\theta\beta 0}}-\frac{64}{\sigma^2_{\theta\beta 1}}-\frac{9}{\sigma^2_{\theta\beta 2}}-\frac{4}{\sigma^2_{\theta\beta 3}}-\frac{\theta^2_{\beta 0}}{200}-\frac{\theta^2_{\beta 1}}{100}-\frac{\theta^2_{\beta 2}}{20}-\frac{\theta^2_{\beta 3}}{8}-\frac{\theta^2_{\delta}}{100})$$

$$\exp(-\frac{1}{2\sigma^2_{\theta\delta}}\sum_{j=1}^{m}(\delta_j-\theta_\delta)^2-\frac{1}{2\sigma^2_{\theta\beta 3}}\sum_{i=1}^{n}(\beta_{3i}-\theta_{\beta 3})^2-\frac{1}{2\sigma^2_{\theta\beta 2}}\sum_{i=1}^{n}(\beta_{2i}-\theta_{\beta 2})^2-\frac{1}{2\sigma^2_{\theta\beta 1}}\sum_{i=1}^{n}(\beta_{1i}-\theta_{\beta 1})^2$$

$$\exp(-\frac{1}{2\sigma^2_{\theta\beta 0}}\sum_{i=1}^{n}(\beta_{0i}-\theta_{\beta 0})^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\beta_{0i}-\beta_{1i}(a_{ij})-\beta_{2i}(a_{ij})^2-\beta_{3i}(z_{ij})(a_{ij}-29)^2)$$

From the unnormalized joint posterior distribution, we can calculate the conditional distributions for each of the parameters (at least to a constant of proportionality). The following is an example of a conditional distribution:

$$p(\theta_{\beta 0}\mid y_{ij},\beta_{0i},\beta_{1i},\beta_{2i},\beta_{3i},\delta_j,\sigma^2_{\theta\beta 0},\theta_{\beta 1},\sigma^2_{\theta\beta 1},\theta_{\beta 2},\sigma^2_{\theta\beta 2},\theta_{\beta 3},\sigma^2_{\theta\beta 3},\theta_\delta,\sigma^2_{\theta\delta},\sigma^2)$$

$$\propto \exp(-\frac{\theta^2_{\beta 0}}{200}-\frac{\sum_{i=1}^{n}(\beta_{0i}-\theta_{\beta 0})^2}{2\sigma^2_{\theta\beta 0}})$$

Because the conditional distributions contain multiple parameters, we use the Gibbs Sampler or Adaptive Rejection Sampling within Gibbs Sampling to get estimates using a Markov Chain simulation using WinBUGS. In the either sampling algorithm, we have the option of choosing starting values for the iterative process. The starting values for all of the parameters were based on the means or expected values of their prior distributions. The better the starting values, the quicker the model will converge.

Section 4.2: Data Collection and Cleaning

The data came from **www.baseball1.com** prepared by Sean Lahman. The first data file (batting.csv) contains player id, year id, games, at-bats, runs, hits, doubles, triples, home runs, rbis, stolen bases, caught stealings, walks, strikeouts, intentional walks, sacrifice hits, sacrifice flies, and ground into double plays for every player from 1871 to 2004. The Berry Value is defined as:

$$\begin{aligned} \mathit{Berry\ Value} = 0.34BB + 0.49(1B) + 0.72(2B) + 1.14(3B) + 1.51HR \\ + 0.26SB - 0.14CS - 0.10(OUT + K) \end{aligned}$$

For Outs, compute $Outs = AB - H + SH + SF$. A second file (master.csv) containing player age information was merged with batting.csv. For each player, the Berry Value for each season, the season indicator variable, and the player's age during that season create the dataset used for this analysis.

The year 1900 is often noted as the start of the modern game of baseball. A large majority of the rules that created the modern game were instituted during the last decade of the 19th century, and also the first decade of the 20th century. To focus on the modern game, only players in the years 1900 to present are modeled. As part of cleaning the data, all of the individuals who had missing ages from the data file are omitted. Also, all individuals who had less than three seasons with more than 100 plate appearances are removed since we want to avoid any outliers (players with only one fantastic season in the model). This eliminates all of the pitchers from the hitting data. Appendix A shows the model in WinBUGS. For this model, we ran 200,000 iterations.

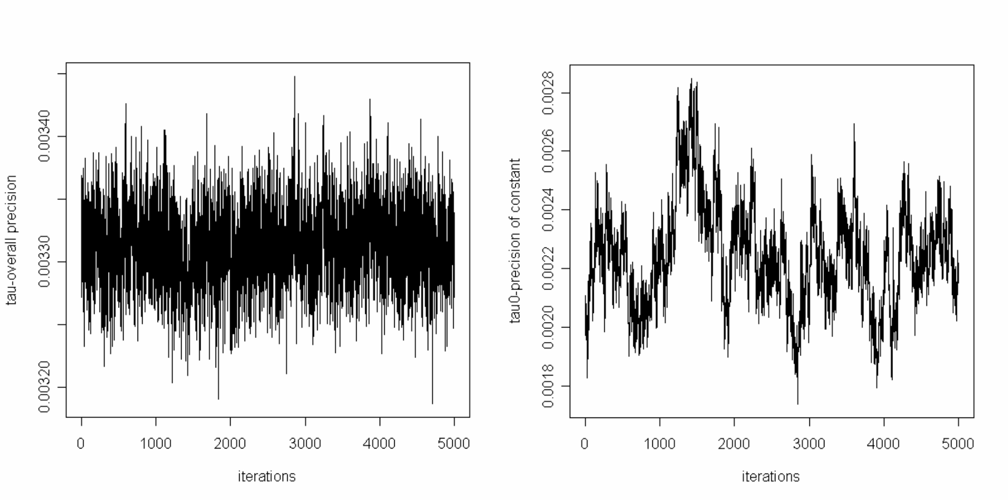
Section 4.3: Convergence

The most important criterion for testing the validity of a BHM model that uses Gibbs Sampling or Adaptive Rejection Sampling within Gibbs Sampling is the test of convergence. Because the Gibbs Sampler uses an iterative process to update the posterior distributions of each parameter, it is important that it converges.

The first thing to be done in checking convergence is to look at the trace plots of the draws from the posterior distribution. The trace plot charts the value of the parameter after each iteration. Figure 4.2 shows the trace plots for the last 5000 iterations of two posterior distribution parameters from this model.

A simple criterion for assessing convergence of draws to the joint posterior distribution is to examine trace plots of the draws. We look to see if there are trends or patterns in the trace plots regarding noticeable shifts in the values of the parameters. If there are no noticeable trends, then it is reasonable to assume that the chain has converged to the joint posterior distribution.

Figure 4.2: Trace Plot of the overall precision of the model and the precision of the model constant to illustrate model convergence and the level of confidence we should have in the model.



Section 4.4: Results

In this section, we discuss the two main effects: the season effect and the age function. The season effects show the relative difficulty of scoring runs in each season. We look into the trends of the season effect and analyze the impact of the changes in baseball over the last 100 years. From the age effect, we estimate the peak performance, and analyze the age curve of multiple players. Since we have linked players from different eras we can rank players according to their ability to score runs.

Section 4.4.1: Season Effects

Figure 4.3: Average Season Effect Coefficient by Year which illustrates the low season effects for the decades of 1910-1920 and 1985-1995 and high season effects for the 1920-1940 and the current state of the game.

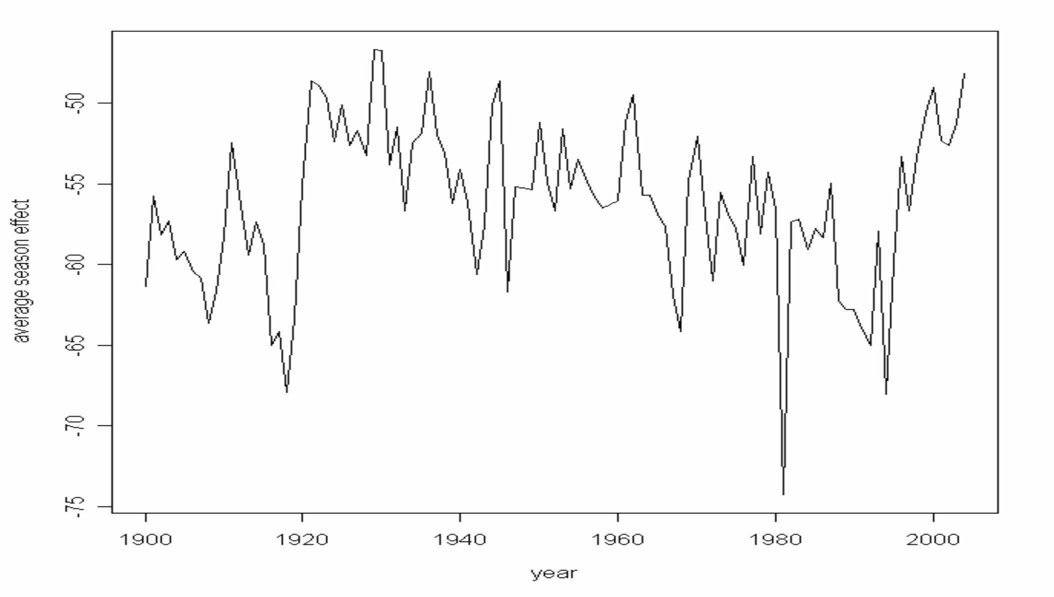


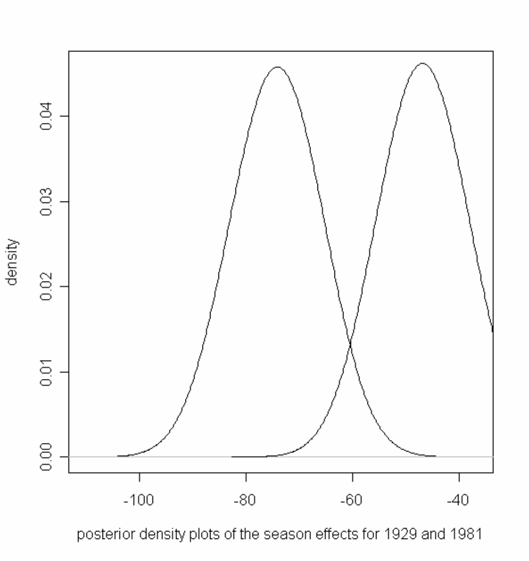
Figure 4.3 illustrates the yearly season effect for Berry Values in the Hierarchical Bayes Model, and a table of the posterior means of the δ_j can be found in Appendix C. The decades with the highest season effects were the 1920s, 1930s and the 2000s. The Berry, et al. (1999) model also had very high seasonal effects for home runs and batting averages during those decades. Lowest season effects were recorded during 1900-1910, the "dead-ball" era, and in the 1980s, when offensive production was low. Some of the drastic downward spikes seen in Figure 4.4 represent strike-

shortened seasons (1981 and 1994) and those affected by war and scandal (1918-1919).

In Section 4.1, we noted the fact that Major League Baseball has seen many rule changes over the years of its existence, which necessitates the inclusion of season effects in the model. The timing of these rule changes can be seen in the season effects. The year 1920 (the year that the RBI was introduced) is the start of a drastic increase in season effects. The year 1947 (the year that Robinson broke the color barrier) is also a part of a slight upturn in season effects; this effect is very slight because of the very slow introduction of African-American players into the Major Leagues. The lowering of the pitcher's mound in 1969 caused a large increase in the season effect for that year. The strike zone change in 2001 had a negative effect on hitters, and this is shown by the fall in the season effects from a peak in 2000 to a lower value in 2001.

The next issue we address is the significance of the season effect in the model. To examine this more closely, we chart the predictive posterior distributions for the seasons with the highest season effect (1929, -46.66), and the lowest season effect (1981, -74.28). Figure 4.4 displays the predictive posterior distributions.

Figure 4.4: Predictive Posterior Distributions for 1929 and 1981 the season with the high and lowest season effects to show that the season effect is indeed significant.



These distributions overlap, due to high variance. To measure the significance of this overlap, we analyze the area shared by the distributions. The two predictive posterior densities intersect at -60. The area in 1981 season density (the lowest season effect) above -60 is .017. The area in the 1929 season density (the highest season effect) below -60 is .022. Therefore, the distributions are statistically significantly different at the 5% level. We conclude that season has a very strong and significant impact in the model.

Section 4.4.2: Age Function

Figure 4.5 plots the mean age function using the hierarchical aging parameters. We make three observations: First, the peak of the distribution is 28 years; second, the peak Berry Value is around 90, which approaches the overall average Berry Value of 44 when the general season effect (which is around -60) is added; and, third, the decline period generally has a much steeper slope than the maturation period. Baseball players tend to decline faster on average than they mature.

Figure 4.5: Overall Age Curve for the Hierarchical Bayes Model to illustrate how the average player had a slow and short period of maturation and a long and steep period of decline.

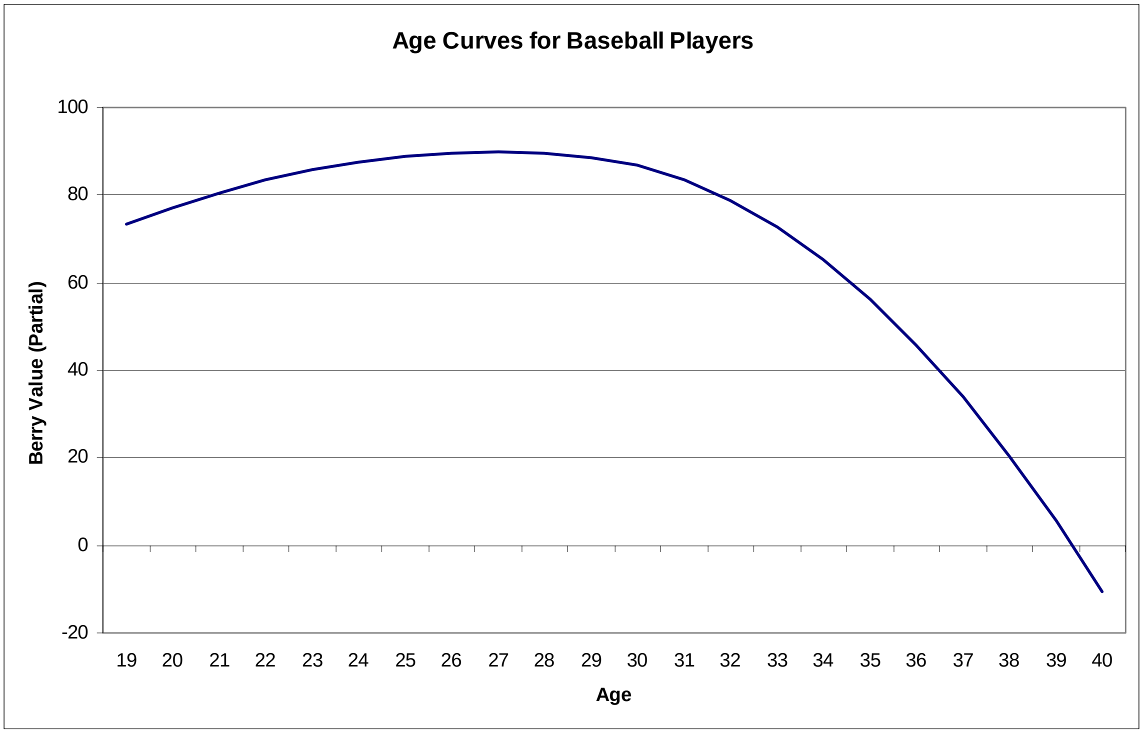
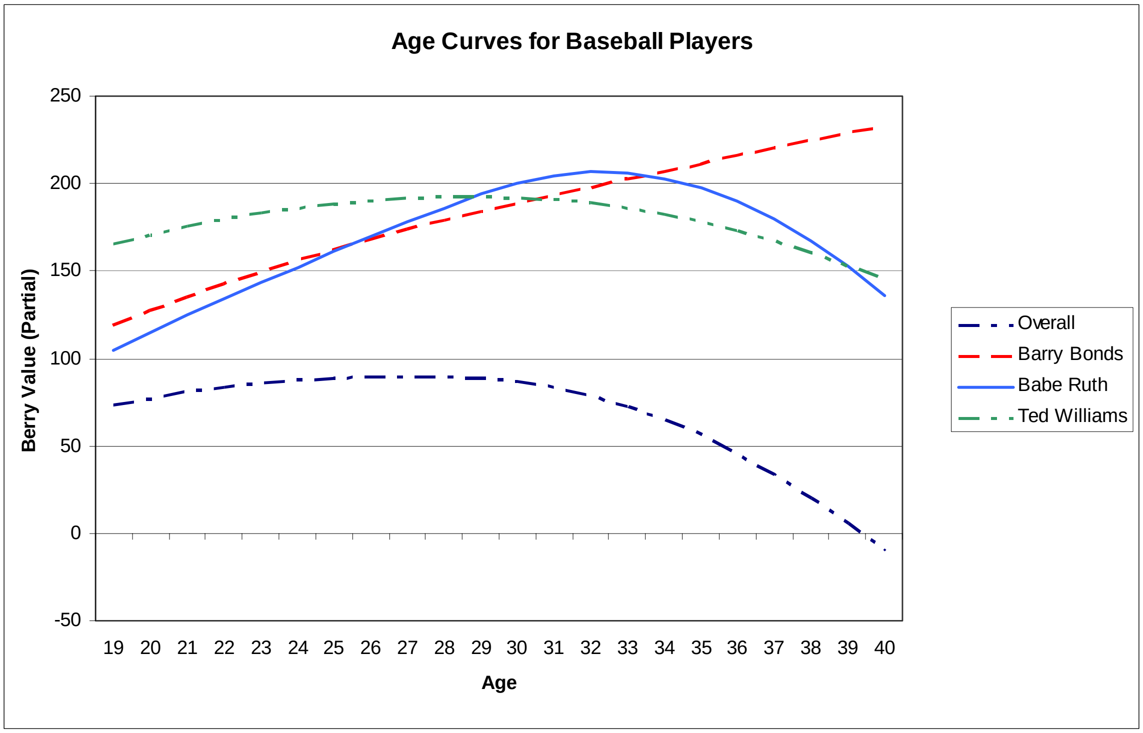


Figure 4.6 includes age curves for some of the best players in the model with the overall curve overlaid for a comparison. Barry Bonds, Babe Ruth, and Ted Williams were chosen because they were in the top 10 as far as the peak of their age curves. This can be seen by the distance of their curves away from the overall average age curve. These three players were also chosen because their age curves represent the different types of age curves that come out of the model.

Figure 4.6: Multiple Age Curves for Some Top Players in the model showing the three different general types of individual age curves which include the straight and steady curve, the parabolic curve, and the Barry Bonds curve.



The first type of age curve that is observed in this model is the straight and consistent age curve represented by Ted Williams. Williams had a consistent and high performing career. He did have a period of maturation and decline, but it is much less pronounced and fairly straight. The second type of age curve is the more parabolic age curve represented by Babe Ruth. Ruth had a higher peak, but much steeper rates of incline and decline than the mean curve. Age clearly had a strong impact on Ruth's performance. Looking at the overall curve, we conclude that a majority of the players in the model have a stronger parabolic age curve and would be in this second group.

The third group is unique. Most players peak at 27 or 28, near the overall model peak. Bonds' age curve, however, is always increasing. There is no peak in his age function, and he is the only player out of the top 100 that does not have a negative piecewise coefficient. Barry Bonds' best four seasons have come in the last 4 years included in the model (2000-2004). This could be due to his training, current equipment, alleged steroid use, or an unknown quantity in the current state of the game that allows a person to continue playing at a high level even at older ages.

The age curve in general suggests that overall offensive performance peaks at a young age. Very few individuals have a peak age of more than 32 (only 8 out of the top 50 offensive players). This agrees with Berry, et al. who concluded that the peak age for home run hitting was 29 and the peak age for batting was 27. It seems that in terms of offensive player performance, any benefit gained by experience is offset by the deteriorating effects of age.

Figures 4.7 and 4.8 are the scatter plots of the actual and predicted Berry Values for two players (Hank Aaron and Babe Ruth). The predicted Berry Values are computed by calculating each player's age curve and adding in the season effects for the seasons in which he played. The connected dots are the predictive values and the unconnected dots are the actual values.

Figure 4.7: Actual vs. Predicted Berry Values for Hank Aaron from which we can see how well the model does at predicting Berry Value for a player with a relatively stable career.

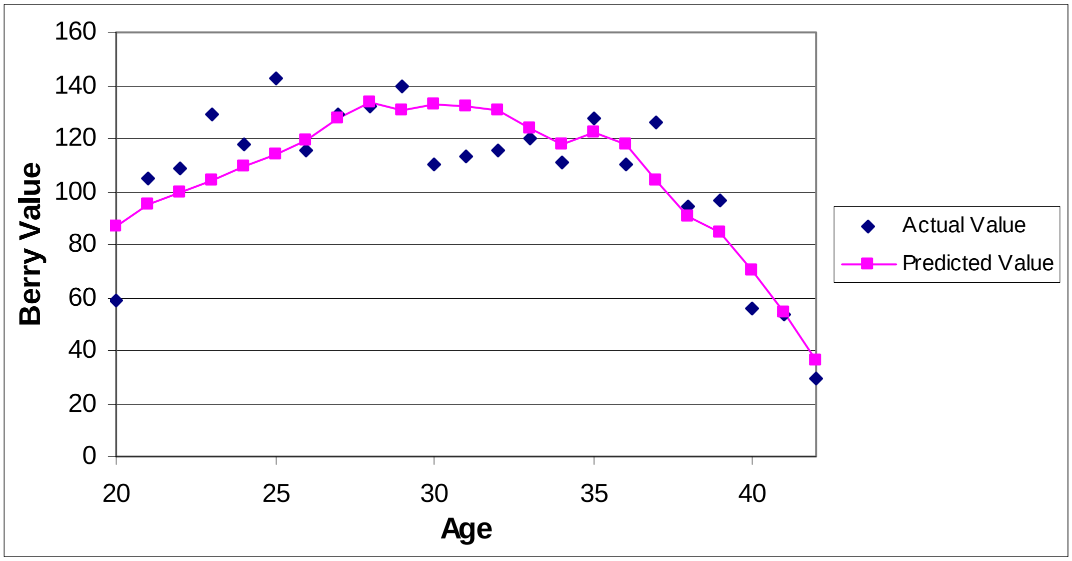
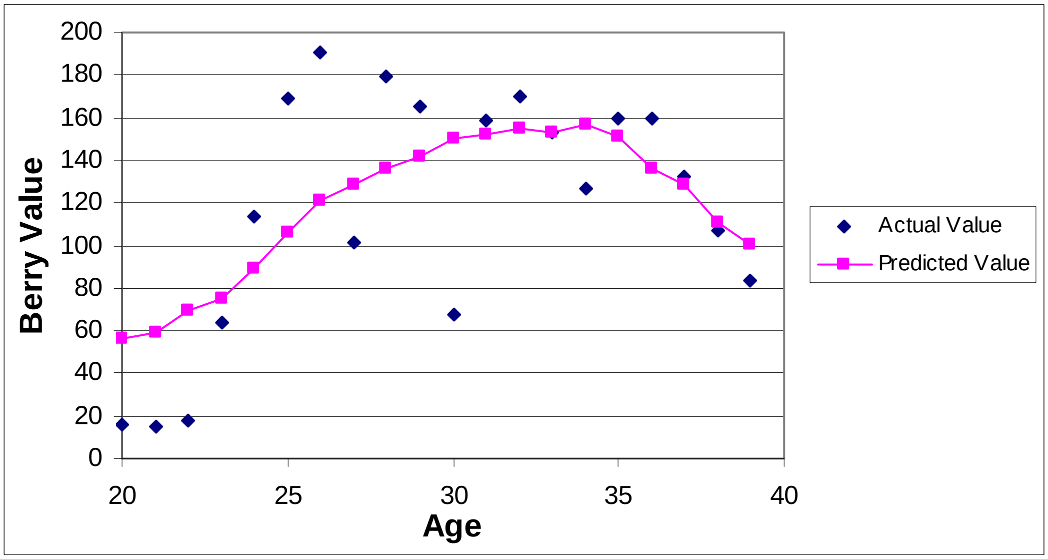


Figure 4.8: Actual vs. Predicted Berry Values for Babe Ruth from which we can see how well the model does at predicting Berry Value for a player with many extreme values.



From the above figures, we can see that the model accurately predicts Berry Value at the individual level. The BHM will not, however, catch outliers (drastic changes in Berry Value due to effects outside the model) like we see in Babe Ruth's data.

The predicted curves hit right in the middle of the actual values and seem to fit the actual data quite accurately. The predicted curves do seem to predict the later years of a player's career better than the earlier years. This makes sense because there is an additional parameter in the model to help predict the second half of a player's career. Considering the borrowing of information between players the model performs better when a player is consistent in his performance.

Section 4.4.3: Rating Player Values According to Berry Value

Because we have built a model to help us predict Berry Value by linking players from different eras, we have a predictive curve for each individual. In ranking the best individuals based on their Berry Value, we need a fixed number or value. In analyzing a curve that has a different slope and peak for each individual, we can rank the players in multiple ways. The option we use is the overall peak value of the aging curve. Because we allow each individual to have his own peak, we rank the player based on the distribution of his peak value. To calculate the peak value for each player, we use the first derivative of each player's age function. As stated before, the following is our age function: $\beta_{0i} + \beta_{1i}(a_{ij}) + \beta_{2i}(a_{ij})^2 + \beta_{3i}z_{ij}(a_{ij} -29)^2$ with the z_{ij} variable taking a value of 1 after the age of 29 years and a 0 before. We therefore, have two first derivative functions. The first derivative function, for ages under 29, is: $\beta_{1i}+ 2\beta_{2i}(a_{ij})$. The second derivative function, for ages 29 and above, is: $\beta_{1i} – 58\beta_{3i} + (2\beta_{2i} + 2\beta_{3i}) (a_{ij})$. The age at which the peak value occurs can be found by setting these functions to zero. The peak age for the first derivative function is: $\frac{-\beta_{1i}}{2\beta_{2i}}$. The peak age for the second derivative function is: $\frac{58\beta_{3i} - \beta_{1i}}{2\beta_{3i} + 2\beta_{2i}}$. If the peak age for the first derivative function is larger than 29 years, then the peak age for the second derivative is used to calculate peak value. After the age at which the curve peaks is found, the peak value can then be deduced by using the peak age in the formula. To calculate the distribution of the peak value, we use the saved age coefficients for the last 5000 iterations using the overall peak age as the age in the function. Therefore, each individual will have 5000 estimates for peak value. A mean and a standard deviation are calculated for this for each individual's posterior distribution. To ensure the selection of the highest peak value with the lowest variability, both the mean and the standard deviation are important measures. The coefficient of variation (mean/standard deviation) is used as the summary statistic of the peak value posterior distribution for each individual. Table 4.1 shows the top 25 players ranked by the value of their coefficient of variation for peak value. According to the model, the top five players of all time in respect to innate player value are: Babe Ruth, Lou Gehrig, Stan Musial, Hank Aaron, and Ted Williams. This list is topped by players who played many seasons at a high level of offensive productivity, and are recognized as the best offensive players of all time. The model consistently identifies current superstars, and those players inducted into the Baseball Hall of Fame. The list includes the all-time home run leader (Hank Aaron), the all time hits leader (Pete Rose), the all-time doubles leader (Tris Speaker), and the all time stolen base leader (Rickey Henderson). The top ranked players in our list include the top players in all time career slugging percentage with Babe Ruth, Lou Gehrig, and Ted Williams. These rankings are strongly correlated with high Berry

Values throughout a players career. Lou Gehrig has the highest career Berry Value average. Babe Ruth and Ted Williams were also in the top ten.

Table 4.1: Innate Player Rankings based on the peak value of the age curve and the variability of that peak value which includes many top echelon Hall of Famers.

Rank	Player	Peak Value CV
1	Babe Ruth	30.10
2	Lou Gehrig	29.28
3	Stan Musial	28.93
4	Hank Aaron	28.08
5	Ted Williams	27.58
6	Jimmie Foxx	27.00
7	Ty Cobb	26.68
8	Willie Mays	26.61
9	Tris Speaker	26.23
10	Rogers Hornsby	25.92
11	Ken Griffey Jr.	25.52
12	Frank Robinson	25.43
13	Paul Waner	25.31
14	Jeff Bagwell	25.29
15	Wade Boggs	25.20
16	Rickey Henderson	25.01
17	Pete Rose	24.89
18	Rafael Palmeiro	24.65
19	Mickey Mantle	24.62
20	Billy Williams	24.59
21	Honus Wagner	24.53
22	Eddie Mathews	24.49
23	Al Simmons	24.43
24	Frank Thomas	24.42
25	Harmon Killebrew	24.29

Very few current players appear on this list. Because they have yet to finish their careers, these contemporary players have fewer seasons in the model and thus more variability. As variability is part of the ranking equation, the high variability of today's players kept them off the list. As current All-Stars like Alex Rodriguez play more seasons at a strong offensive level, they will move up in the rankings. Our focus in this ranking is high peak value and high confidence in that peak value. An alternative ranking is shown in Appendix D.

There are a couple of surprise names in the top 25. While Jeff Bagwell is a perennial All-Star, his name is not among the first that commonly arise when thinking of the top 15 players of all time. His peak Berry Value is high because he is a player that gets on base. He is consistently one of the top 10 players in walks, hit by pitches, and on-base percentage. He also scores many runs and produces large numbers of home runs. Bagwell was also helped by having many good seasons in an era of low season effects.

There are a couple of names missing which according to contemporary opinion, should be in the top 25. Barry Bonds is one of the best current players in the game and is frequently compared to Babe Ruth and Hank Aaron. Bonds has had 14 seasons with a Berry Value over 100 and the season with the highest Berry Value ever (213.11). Barry Bonds is not in the top 25 due to the variability associated with estimation. The dilemma with Bonds is that the final four seasons were the best of his career and his age curve has no declining portion. Because of this anomaly, he has higher variability than other players in the model that have played as long as he has.

Another player that seems to be missing is Reggie Jackson, a Hall of Fame player who finished his career in the top ten all time in home runs and in the top ten in slugging twelve times. There are a couple of reasons that Jackson is not as high in the model. First, he played in an era with the highest season effects (the 1970s). Second, he is top all-time in the category of strikeouts.

Section 4.5: Conclusion

Using Bayesian Hierarchical methodology with a Gibbs Sampler, we have developed a model permitting the linking of players from different eras. Using this model, we parse out the intrinsic offensive value of each Major League Baseball player by accounting for different sources of variation. The effect of the relative difficulty of each season and the effects of age on performance are modeled. We use individual aging curves to account for different aging effects for each individual. The top 50 professionals are all Hall of Fame players, current All-Stars, or had multiple All-Star game appearances. Looking at the seasonal effects, we can conclude that the game of baseball is currently in a state of high run production. The current state of the game is as hitting-dominated as any period in the history of the game including the 1920s and 1930s. This can be due to the lack of pitching depth caused by expansion, and the high amount of conditioning hitters go through today. Looking at the ranked innate player values, only 3 of the top 25 are current players, with the rest representing past eras equally. It is interesting to see a representation of all eras and the fact that even the current state (taking into account the high variability of current players) does not seem too far out of line from other times in baseball history.

Some extensions of this work include accounting for other areas of potential impact, like coaches and ballparks. Allowing another piecewise point to show the level or stable period of a player’s career could also strengthen the age curves. The model created is successful in meeting our objective, which was to construct a system from which we could simultaneously compare players of different eras and ages and assess them on equal grounds.

APPENDIX A

MICKEY MANTLE PROBLEM PROGRAM IN WinBUGS:

```
model
{
for ( p in 1 : 18 ) {
berry[p] ~ dnorm(mu[p],tau)
mu[p] ~ dnorm(theta,tau0)
}
 theta ~ dnorm(80,0.01)
tau ~ dgamma(2,100)
tau0 ~ dgamma(2,100)
}
 list(theta = 80, tau=.01, tau0 = .01, mu = c(80,80,80,80,80,80,80,80,80,80,80,80,80,80,80,80,80,80))
 list(berry = c(46.75, 97.61, 82.41, 105.52, 124.36, 151.15, 153.09, 129.78, 102.07, 111.46, 144.51, 107.27, 43.84, 109.86,
61.37, 63.51, 73.15, 69))
```

FINAL PROJECT PROGRAM IN WinBUGS:

```
model
        {
          for(p in 1 : 29570) {
            Berry[p] ~ dnorm(mu[p], tau)
            mu[p] <- beta0[ind[p]] + beta1[ind[p]] * X1[p]
                  + beta2[ind[p]] * X2[p] + beta3[ind[p]] * X3[p]
                  + gamma[season[p]]
          }
          # Priors for betas:
          for (k in 1 : 5392) {
        beta0[k] ~ dnorm(theta0,tau0)
        beta1[k] ~ dnorm(theta1,tau1)
        beta2[k] ~ dnorm(theta2,tau2)
        beta3[k] ~ dnorm(theta3,tau3)
        }
          # Priors for gamma:
          for (j in 1 : 105) {
            gamma[j] ~ dnorm(thetagamma,taugamma)
          }
          # Hyper-priors:
        theta0 ~ dnorm(0,.01)
        theta1 ~ dnorm(0,.02)
        theta2 ~ dnorm(0,.1)
        theta3 ~ dnorm(0,.25)
        thetagamma ~ dnorm(0,.02)
        tau ~ dgamma(2,50)
        tau0 ~ dgamma(2,200)
```

```
                    tau1 ~ dgamma(2,128)
                    tau2 ~ dgamma(2,18)
                    tau3 ~ dgamma(2,8)
                    taugamma ~ dgamma(2,72)
                    }
```

```
list(gamma = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), theta0 = 0, theta1 = 0, theta2 = 0, theta3 = 0, thetagamma = 0,
tau=.04, tau0 = .01, tau1 = .03, tau2 = .1, tau3 = .25, taugamma = .03)
```

SYNTAX FOR PLOTS:

## READ IN DATA##

```
n <- 580116
m <- 116
m1 <- 105
o <- 5001
alldatag <- rep(0,n*2)
alldatag.m <- matrix(alldatag,n,2)
alldatag.m <- read.table("G:\\Pierre\\HBReg\\Baseball\\taugammaind12-20-05.txt",header=FALSE)
inits <- rep(0,m*3)
inits.m <- matrix(inits,m,3)
inits.m <- read.table("G:\\Pierre\\HBReg\\Baseball\\taugamma12-20-05.txt",header=FALSE)
gammam <- rep(0,m1)
for (i in 1:105) {
gammam[i] <- sum(alldatag.m[inits.m[i,2]:inits.m[i,3],2])/5001
}

m2 <- 11
tau <- rep(0,o)
tau <- alldatag.m[inits.m[106,2]:inits.m[106,3],2]
tau0 <- rep(0,o)
tau0 <- alldatag.m[inits.m[107,2]:inits.m[107,3],2]
tau1 <- rep(0,o)
tau1 <- alldatag.m[inits.m[108,2]:inits.m[108,3],2]
tau2 <- rep(0,o)
tau2 <- alldatag.m[inits.m[109,2]:inits.m[109,3],2]
tau3 <- rep(0,o)
tau3 <- alldatag.m[inits.m[110,2]:inits.m[110,3],2]
taugamma <- rep(0,o)
taugamma <- alldatag.m[inits.m[111,2]:inits.m[111,3],2]
theta0 <- rep(0,o)
theta0 <- alldatag.m[inits.m[112,2]:inits.m[112,3],2]
theta1 <- rep(0,o)
theta1 <- alldatag.m[inits.m[113,2]:inits.m[113,3],2]
theta2 <- rep(0,o)
theta2 <- alldatag.m[inits.m[114,2]:inits.m[114,3],2]
theta3 <- rep(0,o)
theta3 <- alldatag.m[inits.m[115,2]:inits.m[115,3],2]
thetagamma <- rep(0,o)
thetagamma <- alldatag.m[inits.m[116,2]:inits.m[116,3],2]

## CALCULATE MEANS OF PRIOR AGE COEFFICIENT PARAMETERS TO CREATE OVERALL AGE CURVE ##
theta0m <- sum(theta0)/5001
theta1m <- sum(theta1)/5001
theta2m <- sum(theta2)/5001
theta3m <- sum(theta3)/5001

## PRIOR PARAMETER PLOTS FOR CONVERGENCE ##
```

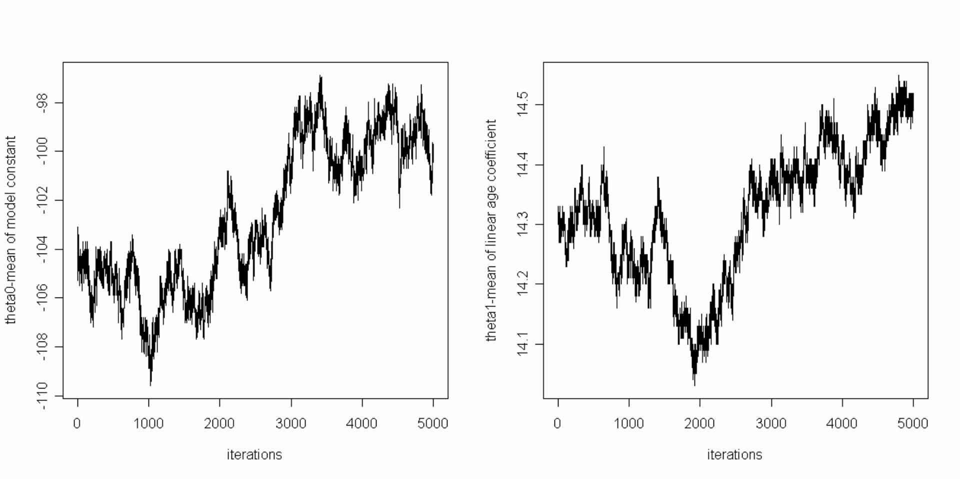
```
plot(x[1:5001],tau,"l",xlab="iterations",ylab="tau-overall precision")
plot(x[1:5001],tau0,"l",xlab="iterations",ylab="tau0-precision of constant")
plot(x[1:5001],tau1,"l",xlab="iterations",ylab="tau1-precision of linear age coefficient")
plot(x[1:5001],tau2,"l",xlab="iterations",ylab="tau2-precision of quadratic age coefficient")
plot(x[1:5001],tau3,"l",xlab="iterations",ylab="tau3-precision of piecewise age coefficient")
plot(x[1:5001],taugamma,"l",xlab="iterations",ylab="taugamma-precision of season effect")
plot(x[1:5001],theta0,"l",xlab="iterations",ylab="theta0-mean of model constant")
plot(x[1:5001],theta1,"l",xlab="iterations",ylab="theta1-mean of linear age coefficient")
plot(x[1:5001],theta2,"l",xlab="iterations",ylab="theta2-mean of quadratic age coefficient")
plot(x[1:5001],theta3,"l",xlab="iterations",ylab="theta3-mean of piecewise age coefficient")
plot(x[1:5001],thetagamma,"l",xlab="iterations",ylab="thetagamma-mean of season effect")

## SEASON EFFECT PLOT ##
plot(x[1900:2004],gammam,"l",xlab="year",ylab="average season effect")

## DENSITY PLOTS ##
plot(density(tau1,bw=.02),xlab="tau1-precision of linear age coefficient",ylab="density",main="")
plot(density(theta1,bw=.05),xlab="theta1-mean of linear age coefficient",ylab="density",main="")

## GROUPING THE SEASON EFFECTS FOR THE LAST 5000 ITERATIONS ##
gamma <- rep(0,o*m1)
gamma.m <- matrix(gamma,o,m1)
for (i in 1:105) {
gamma.m[,i] <- alldatag.m[inits.m[i,2]:inits.m[i,3],2]
}

## CREATING PREDICTIVE POSTERIOR DENSITIES FOR SEASONS WITH HIGHEST AND LOWEST SEASON
EFFECT ##

 ppost1 <- rnorm(length(gamma.m[,82]),gamma.m[,82],sqrt(1/taugamma))
 ppost2 <- rnorm(length(gamma.m[,30]),gamma.m[,30],sqrt(1/taugamma))

 plot(density(ppost1,bw=5),xlab="posterior density plots of the season effects for 1929 and
1981",ylab="density",main="")
 lines(density(ppost2,bw=5))

> mean(ppost1>-60)
[1] 0.01739652
> mean(ppost2>-60)
[1] 0.9776045


#### PLAYER RANKING SYNTAX ###


## LOADING IN THE AGE COEFFICIENTS FOR THE LAST 5000 ITERATIONS FOR THE TOP 150 INDIVIDUALS
###

n <- 3000000
m <- 600
o <- 5000
alldatab <- rep(0,n*2)
alldatab.m <- matrix(alldatab,n,2)
alldatab.m <- read.table("G:\\Pierre\\HBReg\\Baseball\\beta12-22-05.txt",header=FALSE)
inits <- rep(0,m*3)
inits.m <- matrix(inits,m,3)
inits.m <- read.table("G:\\Pierre\\HBReg\\Baseball\\betaind12-21-05.txt",header=FALSE)

## CREATING AN MEAN AGE COEFFICIENT FOR EACH INDIVIDUAL ##
betam <- rep(0,150*4)
betam.m <- matrix(betam,150,4)
for (a in 1:4) {
for (b in 1:150) {
betam.m[b,a] <- sum(alldatab.m[inits.m[(a-1)*150+b,2]:inits.m[(a-1)*150+b,3],2])/5000
}
}

beta <- rep(0,o*150*4)
```

```
beta.m <- matrix(beta,o*150,4)
for (a in 1:4) {
for (b in 1:150) {
beta.m[((b-1)*5000+1):(b*5000),a] <- alldatab.m[inits.m[(a-1)*150+b,2]:inits.m[(a-1)*150+b,3],2]
}
}

## COMPUTING THE PEAK AGE DEPENDING ON THE CORRECT DERIVATIVE ##

peakage1 <- rep(0,150)
peakage2 <- rep(0,150)
peakage <- rep(0,150)
z <- rep(0,150)
peakvalue <- rep(0,o*150)
rpeakvalue <- rep(0,o*150)
peakage1 <- read.table("G:\\Pierre\\HBReg\\Baseball\\peakage1.txt",header=FALSE)
peakage2 <- read.table("G:\\Pierre\\HBReg\\Baseball\\peakage2.txt",header=FALSE)

peakvalue.m <- matrix(peakvalue,o,150)
rpeakvalue.m <- matrix(rpeakvalue,o,150)

for (i in 1:150) {

peakage[i] <- peakage1[i,1]
if(peakage1[i,1] > 29) {
z[i] <- 1
peakage[i] <- peakage2[i,1]
}

## COMPUTING THE PEAK VALUE FOR EACH ITERATION FOR EACH INDIVIDUAL ##

peakvalue.m[,i] <- beta.m[((i-1)*5000+1):(i*5000),1] + beta.m[((i-1)*5000+1):(i*5000),2]*peakage[i] + beta.m[((i-
1)*5000+1):(i*5000),3]*peakage[i]*peakage[i] + beta.m[((i-1)*5000+1):(i*5000),4]*z[i]*(peakage[i]-29)*(peakage[i]-
29)
}

## RANKING THE PEAK VALUES FOR EACH ITERATION ##
for (q in 1:5000) {
rpeakvalue.m[q,] <- rank(peakvalue.m[q,])
}

## DETERMINING THE MEAN, STANDARD DEVIATION, AND AVERAGE RANKING FOR EACH INDIVIDUAL’S
PEAK VALUE ##

peakvaluem <- rep(0,150)
peakvaluesd <- rep(0,150)
rpeakvaluem <- rep(0,150)
for (j in 1:150) {
peakvaluem[j] <- sum(peakvalue.m[,j])/5000
rpeakvaluem[j] <- sum(rpeakvalue.m[,j])/5000
peakvaluesd[j] <- sd(peakvalue.m[,j],na.rm=TRUE)
}
```

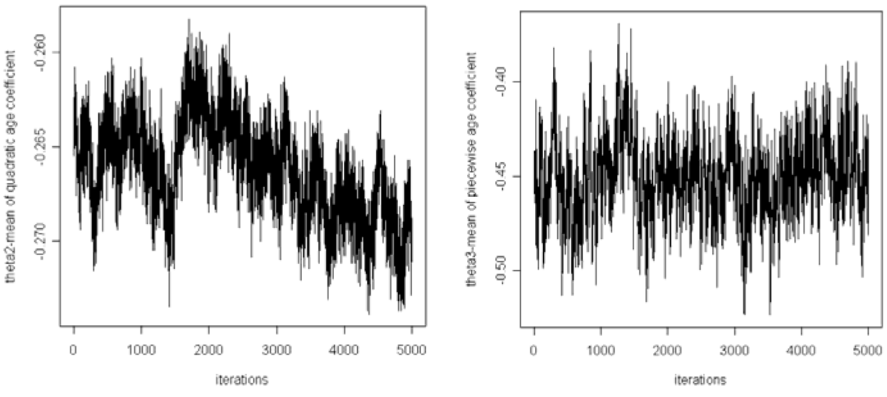
APPENDIX B

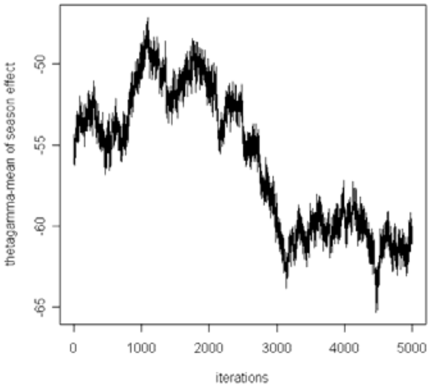
TRACE PLOTS:

$\theta_{\beta 0}$ **and** $\theta_{\beta 1}$**:**

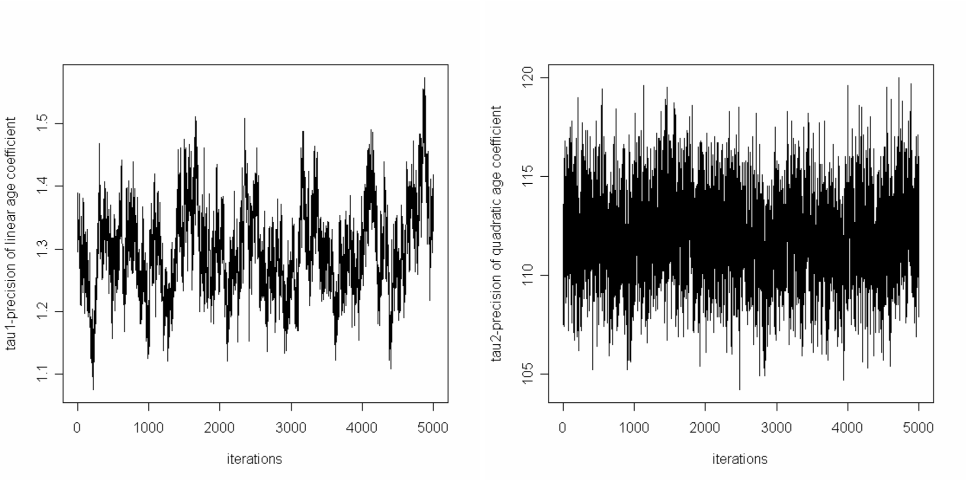


$\theta_{\beta 2}$, $\theta_{\beta 3}$ **and** θ_{δ}

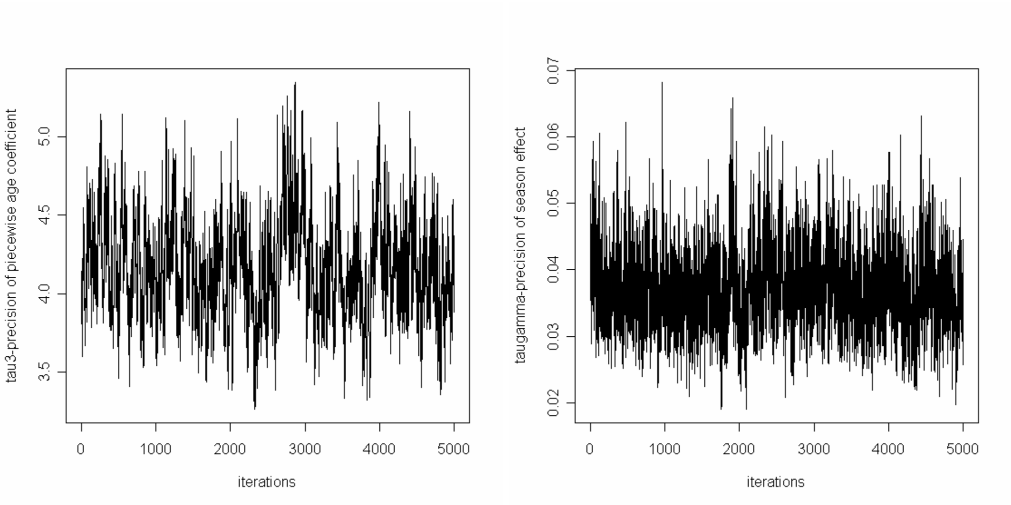




$1/\sigma^2_{\theta\beta 1}$ **and** $1/\sigma^2_{\theta\beta 2}$



$1/\sigma^2_{\theta\beta 3}$**, and** $1/\sigma^2_{\theta\delta}$



APPENDIX C

Table of Seasonal Effects

Season	Season Effect
1900	-61.3562
1901	-55.7748
1902	-58.16586
1903	-57.30684
1904	-59.72716
1905	-59.18174
1906	-60.47011
1907	-60.84239
1908	-63.65127
1909	-61.6669
1910	-58.2812
1911	-52.476
1912	-55.70569
1913	-59.42173
1914	-57.33899
1915	-58.77599
1916	-65.00832
1917	-64.14889
1918	-67.9242
1919	-63.3826
1920	-55.2053
1921	-48.61959
1922	-48.93676
1923	-49.70814
1924	-52.41274
1925	-50.11561
1926	-52.6363
1927	-51.70264
1928	-53.26561
1929	-46.66143
1930	-46.7568
1931	-53.83236
1932	-51.50171
1933	-56.67253
1934	-52.46582
1935	-51.90463
1936	-48.05154
1937	-51.92456
1938	-53.05392
1939	-56.21745
1940	-54.08959
1941	-56.16418
1942	-60.60437
1943	-57.53428
1944	-50.08217
1945	-48.63667
1946	-61.68944
1947	-55.21074
1948	-55.23304
1949	-55.37566
1950	-51.19298
1951	-54.87921
1952	-56.68676
1953	-51.6013
1954	-55.29115
1955	-53.48589
1956	-54.93151
1957	-55.81533
1958	-56.50956
1959	-56.26181
1960	-56.0608
1961	-51.16509
1962	-49.50571
1963	-55.72948
1964	-55.72727
1965	-56.88061
1966	-57.65597
1967	-61.88602
1968	-64.17238
1969	-54.82197
1970	-52.05176
1971	-56.92757
1972	-61.00692
1973	-55.50404
1974	-56.92131
1975	-57.76942
1976	-60.03128
1977	-53.30578
1978	-58.12292
1979	-54.2994
1980	-56.62886
1981	-74.28144
1982	-57.3576
1983	-57.20021
1984	-59.07162
1985	-57.74974
1986	-58.33673
1987	-54.97353
1988	-62.1898
1989	-62.78775
1990	-62.7894
1991	-63.92542
1992	-65.0262
1993	-57.92553
1994	-68.05338
1995	-59.94663
1996	-53.29998
1997	-56.69065
1998	-53.3372
1999	-50.5783
2000	-49.01495
2001	-52.3368
2002	-52.59873
2003	-51.22361
2004	-48.17858

APPENDIX D

Top 50 Innate Player Values Part 1

Rank	Name	Peak Age	Peak Value AVG	Peak Value SD	Peak Value CV
1	Albert Pujols	35.40	261.06	40.71	6.41
2	Barry Bonds	40.00	233.41	11.84	19.72
3	Todd Helton	35.18	208.32	24.33	8.56
4	Lou Gehrig	31.07	207.47	7.09	29.28
5	Babe Ruth	32.27	206.63	6.87	30.10
6	Vladimir Guerrero	32.42	198.99	17.75	11.21
7	Alex Rodriguez	32.55	194.26	16.81	11.56
8	Ted Williams	28.87	192.44	6.98	27.58
9	Stan Musial	30.94	192.30	6.65	28.93
10	Hank Aaron	30.99	189.00	6.73	28.08
11	Jimmie Foxx	29.85	185.82	6.88	27.00
12	Willie Mays	30.91	185.54	6.97	26.61
13	Lance Berkman	33.72	183.48	23.39	7.85
14	Jeff Bagwell	30.64	182.20	7.20	25.29
15	Bobby Abreu	33.13	181.93	16.58	10.97
16	Hank Greenberg	29.83	181.88	8.75	20.78
17	Frank Thomas	29.55	180.76	7.40	24.42
18	Joe DiMaggio	29.63	179.13	8.01	22.37
19	Ty Cobb	30.97	178.52	6.69	26.68
20	Rogers Hornsby	29.53	176.93	6.83	25.92
21	Manny Ramirez	31.51	176.82	9.52	18.58
22	Carlos Beltran	32.47	176.15	19.93	8.84
23	Johnny Mize	30.12	175.99	7.60	23.17
24	Tris Speaker	30.84	174.52	6.65	26.23
25	Billy Williams	30.58	173.11	7.04	24.59

An alternative ranking focuses on the peak value of the age curve and disregards the variability of the estimates. We provided the peak age value average and standard deviation over the last 5000 iterations. The coefficient of variation number is also included for comparison. It does seem very intuitive to look at the peak of the age curve to determine the rank value of the players. Many modern All-Stars like Albert Pujols and Alex Rodriguez make their way into this ranking. The main concern with current players is the extrapolation of the peak value beyond observed data.

Therefore, we would have much less confidence in these estimates because current players may or may not reach those values.

Top 50 Innate Player Values Part 2

Rank	Name	Peak Age	Peak Value AVG	Peak Value SD	Peak Value CV
26	Chipper Jones	30.02	172.98	8.16	21.20
27	Albert Belle	31.44	172.63	7.70	22.41
28	Rafael Palmeiro	33.19	172.51	7.00	24.65
29	Earl Averill	30.49	171.95	7.56	22.75
30	Mike Schmidt	31.56	171.80	7.32	23.47
31	Frank Robinson	29.87	171.66	6.75	25.43
32	Wade Boggs	30.65	171.34	6.80	25.20
33	Honus Wagner	31.17	170.90	6.97	24.53
34	Mel Ott	29.45	169.37	7.01	24.16
35	Mickey Mantle	29.47	169.12	6.87	24.62
36	Joe Jackson	29.74	169.04	9.21	18.35
37	Sammy Sosa	31.85	168.63	7.08	23.81
38	Jim Thome	33.25	168.04	9.52	17.66
39	Eddie Murray	30.39	167.56	6.97	24.05
40	Charlie Gehringer	31.50	167.54	7.22	23.20
41	Gary Sheffield	33.26	167.22	7.50	22.29
42	Rickey Henderson	30.73	166.87	6.67	25.01
43	Pete Rose	32.34	166.64	6.69	24.89
44	Ichiro Suzuki	31.70	166.50	11.93	13.96
45	Joe Morgan	31.26	166.43	7.18	23.20
46	Jesse Burkett	30.59	166.09	11.29	14.71
47	Ralph Kiner	29.25	165.15	7.27	22.72
48	Brian Giles	32.58	164.57	9.71	16.95
49	Carlos Delgado	31.42	164.54	9.25	17.79
50	Harmon Killebrew	30.71	164.03	6.75	24.29

Bibliography

Albert, J. (1994), "Exploring Baseball Hitting Data: What About Those Breakdown Statistics?". *Journal of the American Statistical Association*, 89, 1066-1074.

Bennet, J.M. and Flueck, J.A. (1983), "An Evaluation of MLB Offensive Models." *The American Statistican*, 37, 76-82.

Bennet, J.M. and Flueck, J.A. (1994), "Player Game Percentage." *Proceedings of the Social Statistics Section, American Statisticial Association*, 378-380.

Berry, S.M. (2000), "Modelling Offensive Ability in Baseball." *Chance*, 13, 52-57.

Berry, S. M., Reese, C.S., and Larkey,P.D. (1999),"Bridging Different Eras in Sports." *Journal of the American Statistical Association*, 94, 661-676.

Bukiet, B., Harold, E.R., and Palacios, J.L. (1997), "A Markov Chain Approach to Baseball.", *Operations Research*, 45, 14-23.

Covers, T.M., and Keilers, C.W. (1977), "An Offensive Earned-Run Average for Baseball.", *Operations Research*, 25, 729-740.

Draper, D., Gaver, D., Goel, P., Greenhouse, J., Hedges, L., Morris, C., Tucker, J., and Waternaux, C. (1992), Combining Information: Statistical Issues and Opportunities for Research. Washington, D.C.: National Academy Press.

Efron, B., and Morris, C. (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimators – Part 1: The Bayes Case", *Journal of the American Statistical Association*, 67, 130-139.

Efron, B., and Morris, C. (1975), "Data Analysis Using Stein's Estimator and its Generalizations", *Journal of the American Statistical Association*, 70, 311-319.

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1995), Markov Chain Monte Carlo In Practice. London: Chapman & Hall.

Hoffman, T. (1989), "The Search for the Ultimate Baseball Statistic." *Chance*, 2 (3), 37-39.

Lee, P.M. (1997), Bayesian Statistics: An Introduction. New York: John Wiley & Sons Inc.

Lewis, Michael (2003), Moneyball: The Art of Winning an Unfair Game. New York: W.W. Norton & Company, Inc.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes estimates for the linear model." *Journal of the Royal Statistical Society*, Series B, 34, 1-41.

Lindsey, G.R. (1963), "An Investigation of Strategies in Baseball." *Operations Research*, 11, 477-501.

Pankin, M.D. (1978), "Evaluating Offensive Performance in Baseball." *Operations Research*, 26 (4), 610-619.

Schutz, Robert W. (1995), "The Stability of Individual Performance in Baseball: An Examination of Four 5-Year Periods, 1928-32, 1948-52, 1968-72 and 1988-92." *Proceedings of the Section of Statistics in Sports, American Statistical Association*, 39-44.

Sobel, M.J. (1993), "Bayes & Empirical Bayes Procedures for Comparing Parameters." *Journal of the American Statistical Association*, 88, 687-693.

Thorn, J., Palmer, P., Gershman, M., Silverman, M., Lahman, S., and Spira, G. (eds.) (2001), Total Baseball. Kingston, NY: Total Sports.