Smoothing Career Trajectories of Baseball Hitters

Jim Albert
Department of Mathematics and Statistics
Bowling Green State University

August 22, 2002

Abstract

This paper considers the estimation of the career hitting trajectories for a large number of historical baseball players. A linear-weight batting statistic is used to measure a player's ability to create runs for his team, and a quadratic model is used to summarize a player's trajectory. A Bayesian exchangeable model is used to simultaneously estimate the trajectories for all players with at least 5000 career at-bats who are born in a particular decade. The estimated trajectories are used to analysis the aging patterns of players from different decades. In addition, they are also used to estimate the peak ability and peak age for players born in the 1930's and to make comparisons between great players from the same era.

1. Introduction

One general topic of discussion among baseball fans is the comparison of players. Fans will compare two players from the same era with respect to a number of aspects, including their talent to hit, their fielding ability, and their speeds in running the bases. However, there is one confounding issue that complicates any comparison. Generally, most of the best baseball players start playing professional baseball in their early 20's and finish in their late 30's, and it is well known that a player's ability does not remain constant over the 15-20 years of his career. In fact, a player's ability is thought to generally start at a relatively low level, increase until a particular peak age, and then deteriorate gradually until retirement, as shown in the graph of Figure 1. We will call this ability pattern the *career trajectory* of a player.

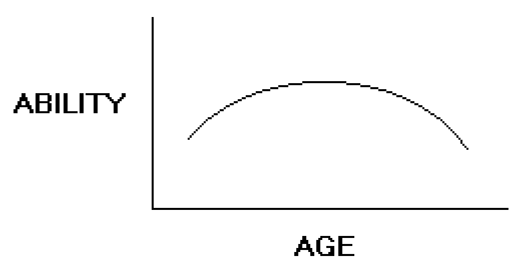

Figure 1
Basic shape of the batting ability of a player as a function of age.

Because of this general pattern of aging of baseball players, the abilities of two players in a particular season should be judged in the context of their career trajectories. It is a bit unfair to compare the hitting accomplishments of a 30-year-old player with a 40-year-old player in a particular season, since the first player is close to his peak performance and the second player is close to retirement. Instead, it is better to compare the entire career trajectories of the two players. In this way, one is comparing the hitting accomplishments of the two players controlling for the aging process.

In this paper, we estimate the career hitting trajectories for a large number of historical baseball players. We restrict attention to those hitters with at least 5000 career plate appearances and group the players by the decade of their birth year. For each player's season of his career, we observe the number of plate appearances and an estimate of the player's hitting performance. In Section 3, we justify the use of a linear weight batting statistic as a "good" measure of hitting performance. A simple way of modeling the career trajectory of a single player is by a quadratic fit. For many players, this quadratic model gives a reasonable estimate of the career trajectory. However, it will be seen in Section 5 that the quadratic fit can give odd estimates of the trajectory for some players. This analysis motivates the use of a Bayesian exchangeable model that reflects the belief in similarity in the career trajectories for the players born during the same decade. The trajectory estimates from this exchangeable model will be seen to correct some of the anomalous features of the individual trajectory estimates.

The estimated career trajectories provide a useful way of comparing hitters in a particular decade. From a player's estimated trajectory, we can estimate the player's age where his ability is maximized, the hitting ability of the player at this peak, and the size of the increase/decrease of the batter's ability about the peak. In Section 6, we use these three measures (peak age, ability at peak, and rate of growth and decline) to describe the trajectories of all of the hitters in a decade. We use these measures to judge the hitting accomplishments of all players during a particular decade. In Section 7, we compare these model-based estimates with naïve estimates of peak ability and peak age based on a player's career statistics. We use these estimated trajectories in Section 8 to compare the hitting accomplishments of some famous players.

2. The data

From Sean Lahman's baseball database (obtainable from www.baseball1.com), one can obtain the season batting statistics for all players in the history of Major League Baseball. We focus on the players who were born on or after 1910, and we divide the players into six groups by the decade in which they were born (1910's, 1920's, 1930's, 1940's, 1950's, and 1960's).

For each season for each player, we record (1) the season year, (2) the age of the player on July 1 of that season, and (3) the basic hitting statistics (ab, h, 2b, 3b, hr, bb, sf, sh, hbp).

We wish to focus our analysis on players who played many seasons with many plate appearances. Also we wish to exclude pitchers and part-time players with a small number of plate appearances. So we restrict attention in our analysis to the players who had at least 5000 career plate appearances. Through the 2001 baseball season, there were

473 players born on or after 1910 who had at least 5000 plate appearances. Table 1 shows the number of players born in each of the six decades and lists some famous hitters from each decade.

Table 1
Number of players with at least 5000 plate appearances born in each of six decades and some famous players in each decade.

Birthyear	Number of players with 5000 PA	Some famous hitters in the decade.
1910-1919	50	Hank Greenberg, Joe DiMaggio, Ted Williams
1920-1929	50	Ralph Kiner, Duke Snider, Stan Musial
1930-1939	61	Mickey Mantle, Willie Mays, Hank Aaron
1940-1949	109	Mike Schmidt, Willie Stargell, Reggie Jackson
1950-1959	97	George Brett, Eddie Murray, Jim Rice
1960-1969	106	Barry Bonds, Sammy Sosa, Mark McGwire

3. Measure of batting performance

Given a player's hitting statistics for a season, we wish to use a good estimate of the player's hitting ability. The standard batting measures that are commonly reported in the media are the batting average (AVG), the slugging percentage (SLG), and the on-base percentage (OBP). It is well known by sabermetricians that these three measures are relatively weak measures of batting performance. The batting average ignores the plate appearances where the player gets on-base by walks, and places equal weights on all types of hits. The slugging percentage, like the batting average, ignores walks, and weights hits by the number of bases attained. (We will see shortly that the number of bases achieved is not the best way of weighting the different types of hits.) The on-base percentage does account for walks, but (like the batting average) places equal value on all types of hits.

We would like to use a measure of batting performance that

- accounts for the value of hits and walks
- weights the different types of hits (single, double, triple, and home run) by numbers that reflect the values of these hits

Albert and Bennett (2001) give an extensive review of different batting measures. Since the goal of hitting is to produce runs, it makes sense to give weights to different batting events that correspond to the values of these events in producing runs. One can assess the value of these events by fitting a regression model to team offensive statistics. Thorn and Palmer (1984) (also see Thorn et al (2001)) found that a good measure of batting performance is the linear weights formula

$$LW = .46\times(\text{1B}) + .80\times(\text{2B}) + 1.02\times(\text{3B}) + 1.40\times(\text{HR}) + .33\times(\text{BB+HBP}),$$

where 1B, 2B, 3B, HR are respectively the number of singles, doubles, triples, and home runs of a player, and BB, HBP are the number of walks and hit-by-pitch.

The linear weights statistic LW measures the total run production of a player. To measure the ability of a player to create runs during a particular plate appearance, we divide the linear weights statistic by the number of plate appearances, obtaining the *average linear weight*

$$ALW = \frac{LW}{PA}$$

where we define the number of plate appearances to be

$$PA = AB + BB + HBP,$$

where AB is the number of at-bats. Note that we are ignoring the benefit of sacrifice flies in this definition. Although sacrifice flies do help in scoring runs, Albert and Bennett (2001) show that it is difficult to assess their benefit using the linear weights regression model and so they exclude this type of batting play in the model. Since sacrifice flies are relatively rare batting events in baseball, their exclusion will have little effect in our player comparisons.

4. Quadratic regression model

We are interested in modeling a player's batting performance in terms of his age, where we use the average linear weight (ALW) as our batting measure. We expect a player's ability to grow during his early years in the Major Leagues, reach a peak, and then decrease in his final years as a professional. That is, we expect a player's ability to have the basic shape shown in Figure 1.

We can obtain this shape by use of the quadratic model

$$\beta_0 + \beta_1 \text{ age} + \beta_2 \text{ age}^2 .$$

To summarize a particular quadratic fit, it is helpful to reparameterize $(\beta_0, \beta_1, \beta_2)$ by the *peak value*, $P = \beta_0 - \frac{\beta_1^2}{4\beta_2}$ the *peak age* $AGE^* = -\frac{\beta_1}{2\beta_2}$, and the *curvature* β_2. The peak value is the maximum hitting ability of the player, the peak age is the age where the player achieved this maximum ability, and the curvature is informative about the rate at which the player's ability changes around the peak value.

Let's illustrate the use of a quadratic fit using batting statistics for Sal Bando displayed in Table 2. Figure 2 constructs a scatterplot of the (age, ALW) data and overlays the quadratic fit

$$-0.27 + 0.033 \text{ age} - 0.00058 \text{ age}^2 .$$

From this fit, we compute the peak value $P = 0.197$ and the peak age $AGE^* = 28.4$ – both these values are shown in Figure 2. We can conclude that Bando's peak ability is approximately .2 and he achieved it about age 28. The coefficient value β_2=-.00058 reflects the shape of the quadratic fit about the modal value.

Table 2: Batting statistics for Sal Bando.

AGE	PA	ALW
22	24	0.1780
23	130	0.1336
24	605	0.1641
25	609	0.2165
26	502	0.2050
27	538	0.2043
28	535	0.1756
29	592	0.2159
30	498	0.1946
31	562	0.1720
32	550	0.1901
33	580	0.1813
34	540	0.1986
35	476	0.1651
36	254	0.1487
37	65	0.1577

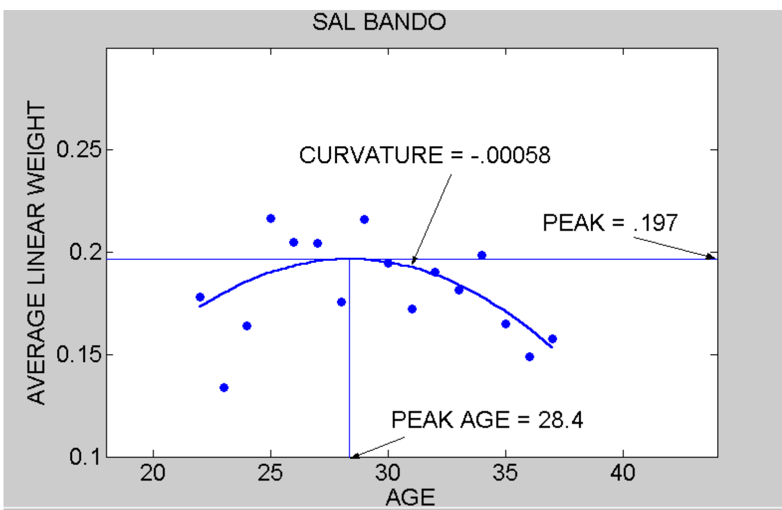

Figure 2: Scatterplot of (age, ALW) data for Sal Bando with quadratic smoothing curve placed on top.

5. Modeling

5.1 Separate Regression Estimates

For a particular player, let (y_j, n_j, x_j) denote respectively the average linear weight, the number of plate appearances, and the age of the player in the jth season. Since the number of plate appearances n_j varies across seasons, the variability of the response y_j will not be constant across seasons and this should be accounted for in our modeling. We assume that y_j is distributed normal with mean given by the regression model $\mu_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2$ and variance $v_j = \sigma^2 / n_j$. If we fit this model, the maximum likelihood estimates are essentially weighted least-squares estimates with weights n_j.

Generally this model appears to give reasonable estimates at the career trajectory of the players' batting abilities. However, if one looks at these estimates for many players, some estimates appear unsatisfactory. Many players, such as Sal Bando (see Figure 2), exhibit a large season-to-season variability in their ALW values, making it difficult to detect the underlying quadratic structure. Also, unusual ALW values for small or large ages can distort the regression fit. This is illustrated in Figure 3, which plots the data and quadratic fits for Norm Cash and Frank Malzone. For Cash, note that the fit (solid line) indicates that he had his greatest ability as a rookie and his ability leveled out for later years. This behavior is inconsistent with our general beliefs about the aging pattern. For Malzone, his relatively poor batting performance at age 26 has a significant effect on the quadratic fit. It seems that the fit has more curvature than we would expect for a player.

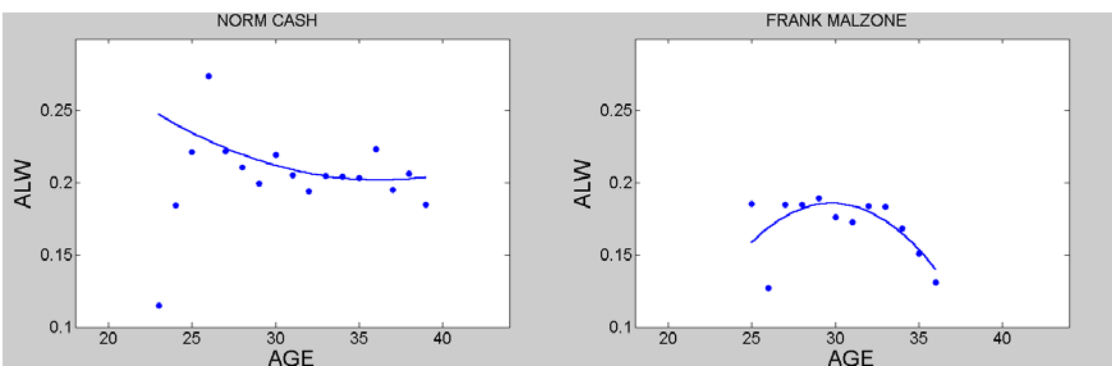

Figure 3: Scatterplots of (age, ALW) data and separate regression estimates (solid lines) for Norm Cash and Frank Malzone.

5.2 Combining Regression Estimates

For each player born in a particular decade, we fit the normal regression model for the (age, ALW) data. We observed in Section 5.1 that some of the individual regression estimates were unsatisfactory since each fit is based on a relatively small sample and the fit can be easily distorted by a couple of extreme points. We are interested in combining the individual regression estimates in a way that reflects our belief about the common aging behavior of major league hitters.

Let $\hat{\beta}_i = (\hat{\beta}_{i0}, \hat{\beta}_{i1}, \hat{\beta}_{i2})$ denote the vector of regression estimates for the *i*th player born in a particular decade, and let V_i denote the corresponding variance-covariance matrix (from the maximum likelihood fit) of this regression estimate. We assume that $\hat{\beta}_i$ is distributed $N(\beta_i, V_i)$, $i = 1, \ldots, p$. We wish to simultaneously estimate the underlying regression parameters $\beta_1, \ldots, \beta_p$.

A Bayesian exchangeable model is a convenient way of combining the individual regression estimates. (A good discussion of the rationale and use of Bayesian exchangeable models is contained in Gelman et al (1995), chapters xx and xx.) We believe that the *p* players born in the particular decade have similar career trajectories, and we represent this belief by assuming that $\beta_1, \ldots, \beta_p$ are a random sample from a common multivariate normal distribution with mean vector β^0 and variance-covariance matrix Σ. The values of the parameters β^0 and Σ are unknown and we represent this lack of knowledge by placing a uniform distribution on (β^0, Σ).

Expressions for the posterior distribution and a description of the simulation algorithm for simulating from this distribution are contained in the Appendix. We learn about the regression vectors by taking a simulated sample from the posterior distribution and $\beta_1, \ldots, \beta_p$ are estimated by their respective posterior means.

To understand how this exchangeable model gives "improved" estimates of the career trajectories, Figure 4 compares two estimates of the trajectories for Norm Cash and Frank Malzone. The individual estimates are represented by thin lines and the estimates using the exchangeable model are shown by thick lines. Note that the effect of the exchangeable model is to move the individual estimates towards a common career trajectory estimate. The exchangeable estimate corrects the nonintuitive decreasing estimate for Cash, and corrects the strong curvature of the individual estimate for Malzone.

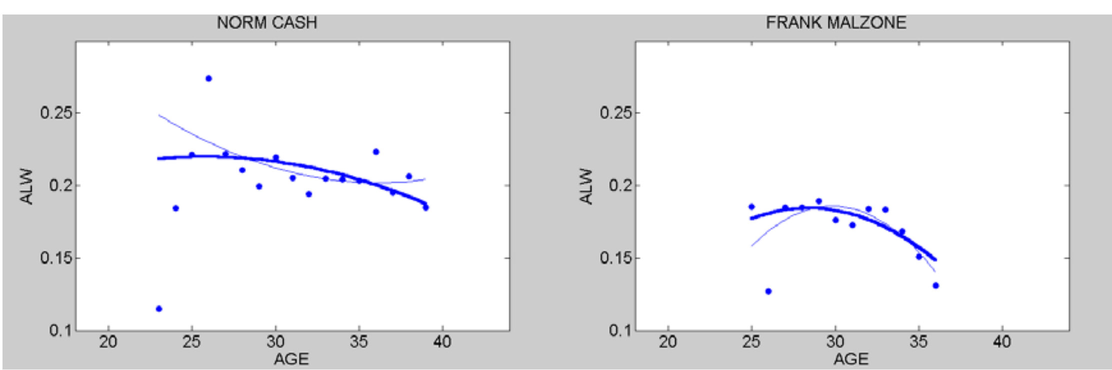

Figure 4: Scatterplots of (age, ALW) and separate regression (thin line) and exchangeable (thick line) estimates for Norm Cash and Frank Malzone.

6. Analysis of the estimated career trajectories

For each of the six groups of players categorized by decade, we used the Bayesian exchangeable model to simultaneously estimate the trajectories of the players. For each fitted trajectory (estimate at β_i), we can estimate a player's peak age, his peak hitting ability, and the curvature. Table 3 summarizes these estimates for all players in each decade. Although there have been large changes in the offensive performances of players over the hundred years of baseball, it is interesting to note the similarity of the career trajectories across decades. Although the peak age estimates vary greatly between players, the median player estimate is between 27.1-29.8 for all six decades. In addition, the median peak ability estimate is about .2 for all decades, and likewise there are similarities of the curvatures across decades.

Table 3: Summaries (lower quartile, median, upper quartile) of the estimated trajectories of all of the players with at least 5000 plate appearances born between 1910 and 1969.

Decade	PEAK AGE	PEAK	CURVATURE (x 1000)
1910s	(24.1, 28.0, 30.4)	(.192, .201, .212)	(-0.550, -0.283, -0.115)
1920s	(27.3, 28.6, 30.0)	(.189, .200, .210)	(-0.694, -0.484, -0.316)
1930s	(25.6, 27.1, 28.5)	(.178, .196, .218)	(-0.617, -0.389, -0.264)
1940s	(27.6, 28.9, 30.1)	(.179, .193, .205)	(-0.493, -0.350, -0.229)
1950s	(27.5, 28.7, 30.0)	(.180, .194, .204)	(-0.482, -0.365, -0.241)
1960s	(27.9, 29.8, 32.0)	(.188, .209, .221)	(-0.684, -0.383, -0.169)

Next, we focus on the estimated career trajectories of the players born in the 1930's. There are three dimensions of a player's trajectory, the age where he peaks, the peak ability, and the curvature (rate of increase and decrease) about the peak. Figure 5 plots the peak age estimates against the peak estimates for the 61 players with at least 5000 career plate appearances, and Figure 6 plots the curvature estimates against the peak estimates for the same players.

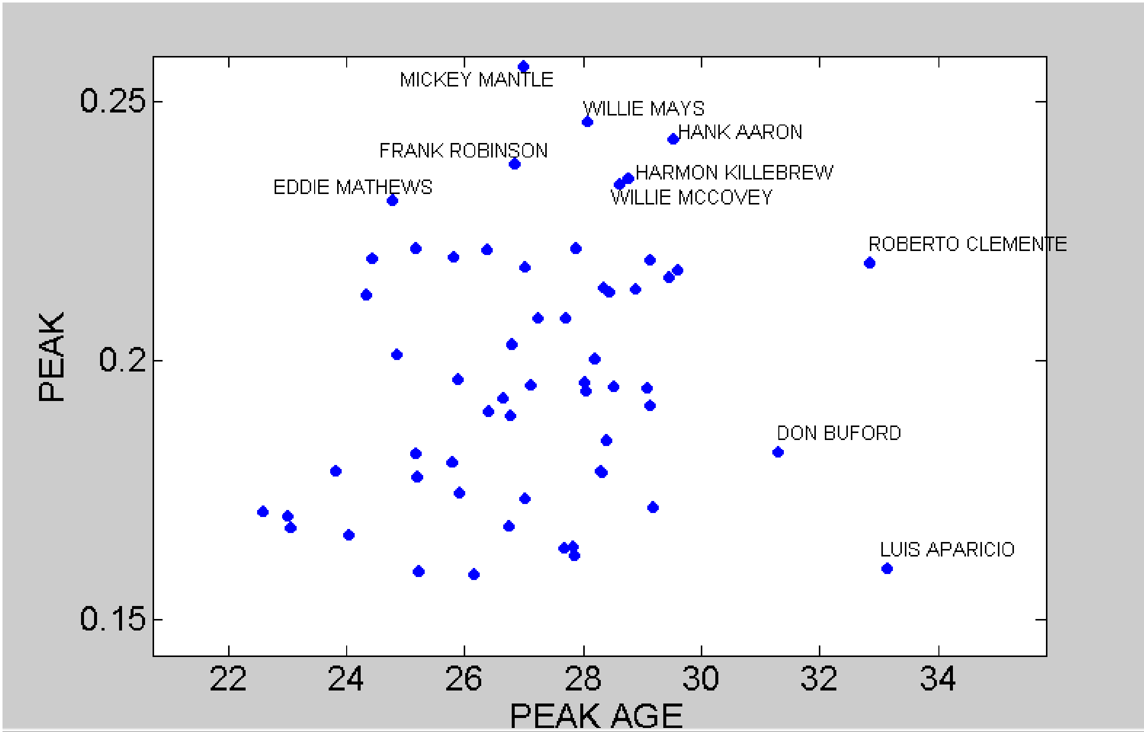

Figure 5
Scatterplot of peak age and peak estimates for all players born in the 1930's with at least 5000 plate appearances.

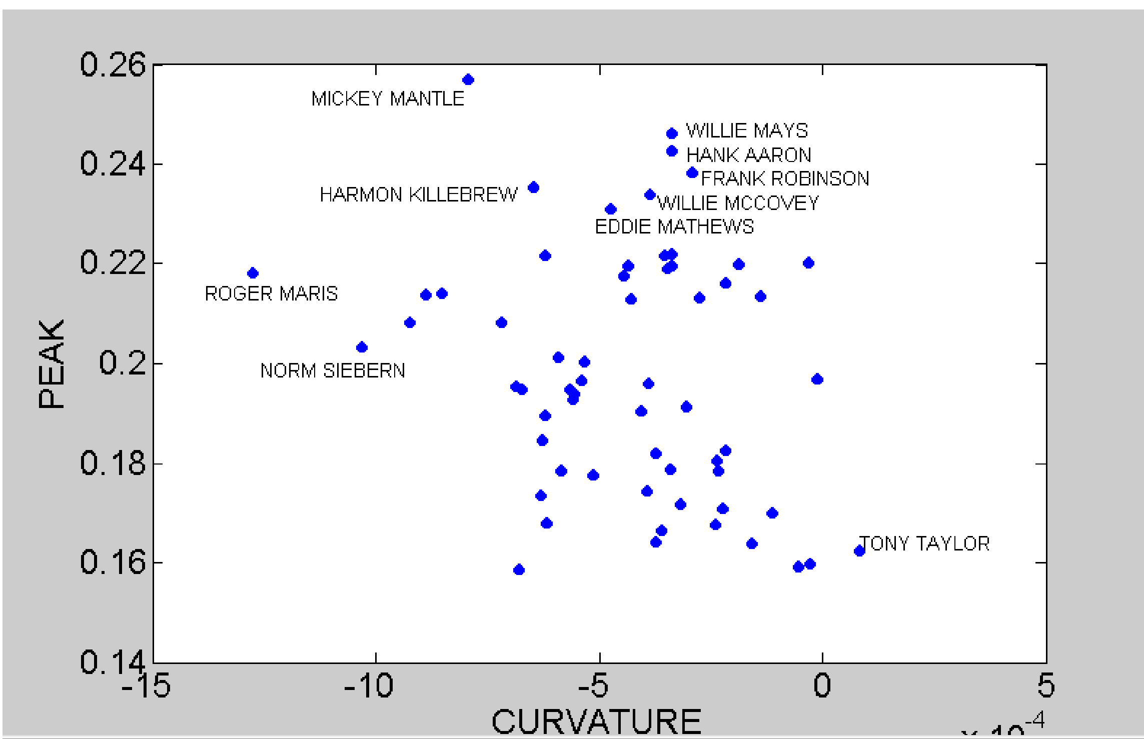

Figure 6
Scatterplot of curvature and peak estimates for all players born in the 1930's with at least 5000 plate appearances.

A number of points are labeled in the two plots corresponding to some of the famous hitters of this decade. Mickey Mantle stands out as the best hitter with regards to peak performance. From Figure 6, Mantle is an extreme point in this group of players, both with regards to his peak performance and his large curvature of his trajectory about the peak value. The next two best hitters, Willie Mays and Hank Aaron, had a peak age a bit later than Mantle, and both hitters had smaller curvature than Mantle about the peak value. That is, Mays and Aaron were better than Mantle in maintaining their high batting performance over many years. Some interesting extreme points are labeled. Roberto Clemente's estimated peak age is relatively high. This particular estimate may have been affected by the premature end of his career at age 38. Roger Maris, despite having 61 home runs in 1961, has an estimated peak ability of only .22, and he has a large curvature, which is reflective of his rapid rise and decline from his peak ability.

Figure 7 plots the estimated career trajectories for eight of the best hitters who were born in the 1930's. Visually, the career trajectories of Hank Aaron and Willie Mays look very similar. They had similar peak abilities, but Aaron's ability deteriorated less with increasing age. The size of the decline of some of the great hitters, such as Harmon Killebrew, Eddie Mathews, and Willie McCovey, is notable.

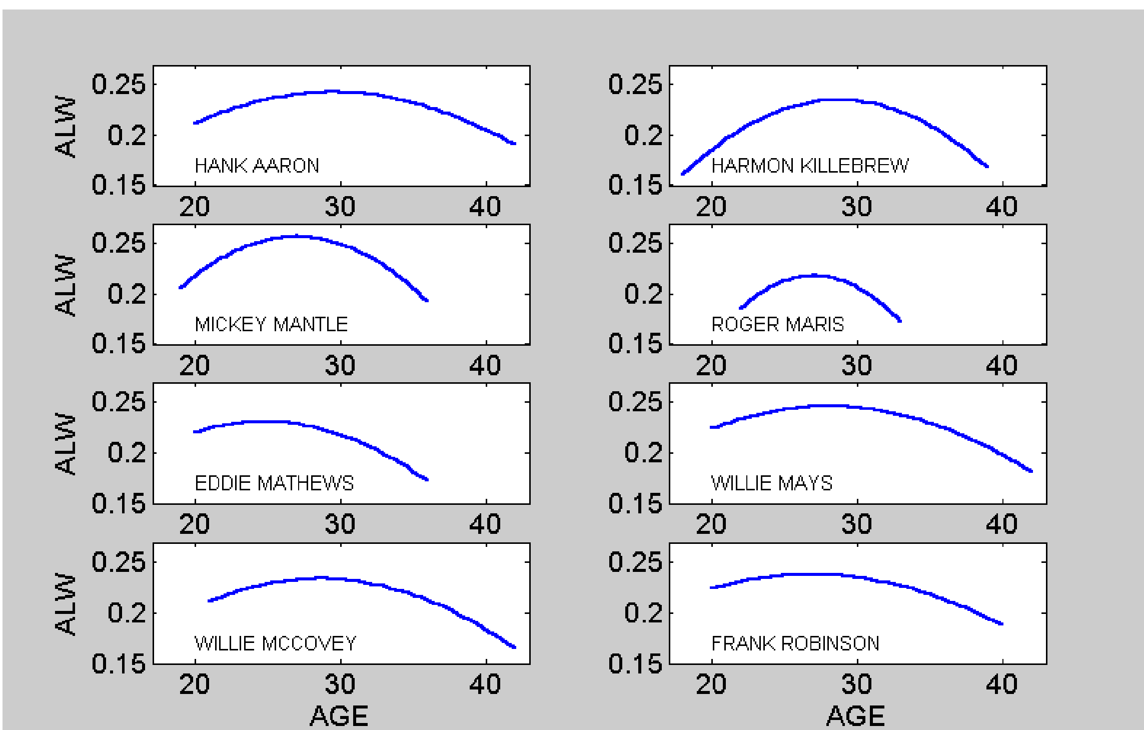

Figure 7
Estimated career trajectories for eight great hitters who were born in the 1930's.

7. Comparison of naïve and model-based peak value and peak age estimates

It is instructive to compare the peak value and peak age estimates using the exchangeable model with naïve estimates based only on the observed data. Given a player's career hitting statistics, the naïve estimate of his peak ability is the maximum average linear weight

$$\max_j y_j .$$

Likewise, the naïve estimate of a player's peak age is the age where his average linear weight is maximized.

A scatterplot of the naïve and model-based peak values is shown in Figure 8. The line through the origin with unit slope is drawn on the plot to help in comparison. Note that all of the points fall under the line, indicating that the model-based peak values are always smaller than the observed peak values. This is expected since the naïve estimates ignore the large season-to-season variability of the average linear weights. The line

$$ESTIMATED\ PEAK = OBSERVED\ PEAK - 0.022$$

is a reasonable fit to the points, indicating that the exchangeable peak value estimate is generally .02 smaller than the observed peak value.

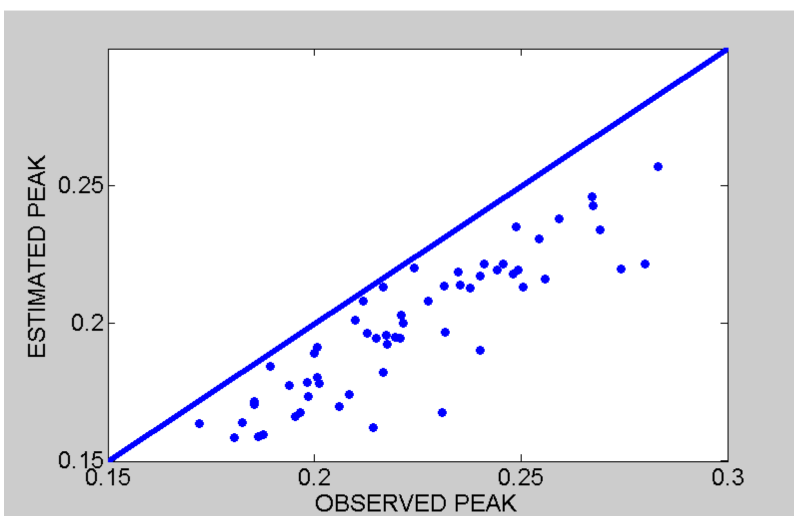

Figure 8
Scatterplot of observed and model-based estimates of peak values for hitters born in the 1930's.

Figure 9 displays a scatterplot of the naïve and model-based peak age estimates. Note that there is a wide variability in the observed peak ages for the players. This indicates that it is relatively difficult to estimate a player's peak age without using some smooth model. In contrast, the estimates of the peak ages using the exchangeable model are stable with most of the values between 25 and 30 years. There is a weak association in the scatterplot, indicating that the year in which a player has the best performance is not a good predictor of his model-based peak age.

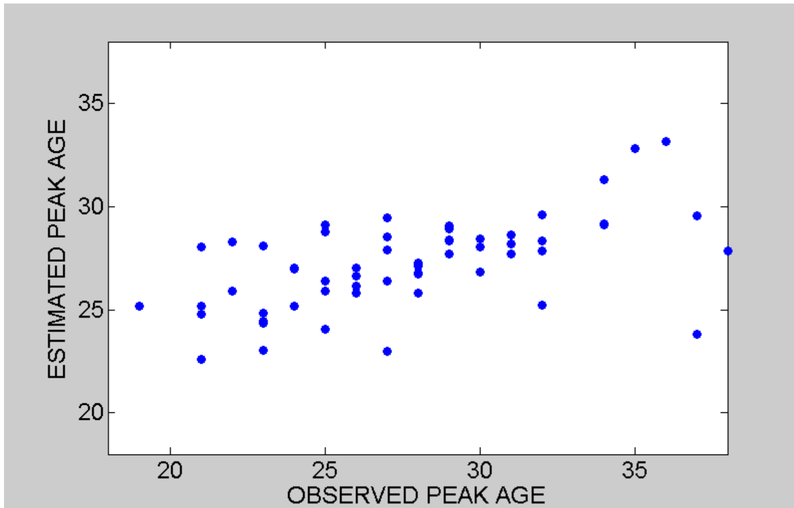

Figure 9
Scatterplot of observed and model-based estimates of peak ages for hitters born in the 1930's.

In Figure 7, we observe that some players like Roger Maris had short careers with large curvatures, and other players such as Hank Aaron had long careers with small curvatures. Is there a general relationship between a player's length of career (defined by the range of ages of his career) and the curvature in the model-based fit? To answer this question, Figure 10 shows a scatterplot of the career lengths and the curvature estimates for the players born in the 1930's. A loess smoother (Cleveland, 1979) is placed on top of the scatterplot to show the basic pattern in the plot. Note for career lengths between 10 and 19 years, there appears to be a positive association in the plot – in this range of career lengths, players with longer careers tend to have smaller curvature in their career trajectories

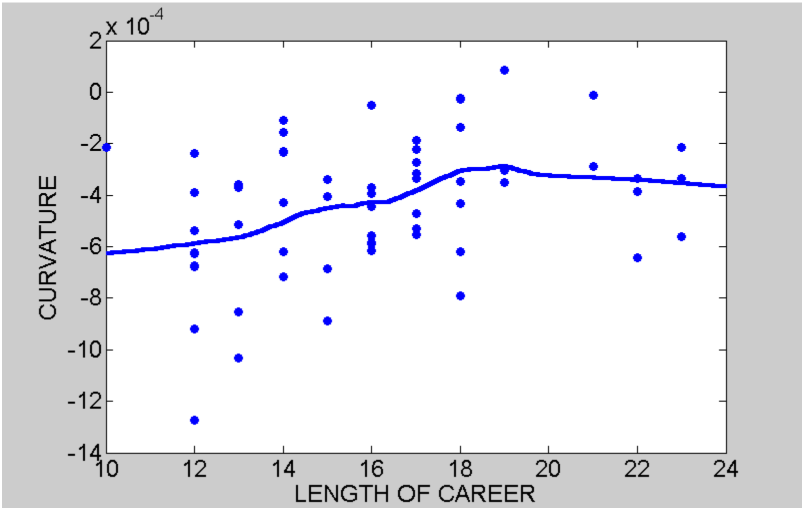

Figure 10
Scatterplot of length of career and curvature estimates for hitters born in the 1930's. A lowess smoother is drawn on top of the scatterplot.

8. Comparison of players

The estimated career trajectories are helpful in the comparison of players from a given era. Several chapters in Berra (2002) involve these type of player comparisons. Among the players born in the 1910s, the dominant two hitters were Ted Williams and Joe DiMaggio. Table 4 gives the age, average linear weight, and number of plate appearances for Williams and DiMaggio for the seasons of their careers. Figure 11 plots the values of ALW for the two players and superimposes the fitted trajectories. With respect to hitting, it is clear from the figure that Williams was the superior hitter. What is remarkable is the flatness of Williams' trajectory, and this is even more remarkable given the extra knowledge that there were two significant breaks in his career due to military service in World War II and the Korean Conflict.

Table 4: Average linear weight, number of plate appearances, and age for Ted Williams and Joe DiMaggio for the seasons of their careers.

	TED WILLIAMS		JOE DIMAGGIO	
Age	ALW	PA	ALW	PA
20	0.2542	674		
21	0.2508	660	0.2325	665
22	0.3007	604	0.2684	690
23	0.2726	671	0.2388	660
24			0.2739	518
25			0.2573	572
26			0.2643	621
27	0.2766	672	0.214	680
28	0.2689	692		

29	0.2652	638		
30	0.2722	730		
31	0.2665	416	0.2164	564
32	0.2448	675	0.2235	601
33	0.3333	12	0.2449	669
34	0.3414	110	0.2533	329
35	0.2716	523	0.2414	606
36	0.2864	413	0.1935	482
37	0.2598	503		
38	0.2972	544		
39	0.2517	513		
40	0.1965	326		
41	0.2659	388		

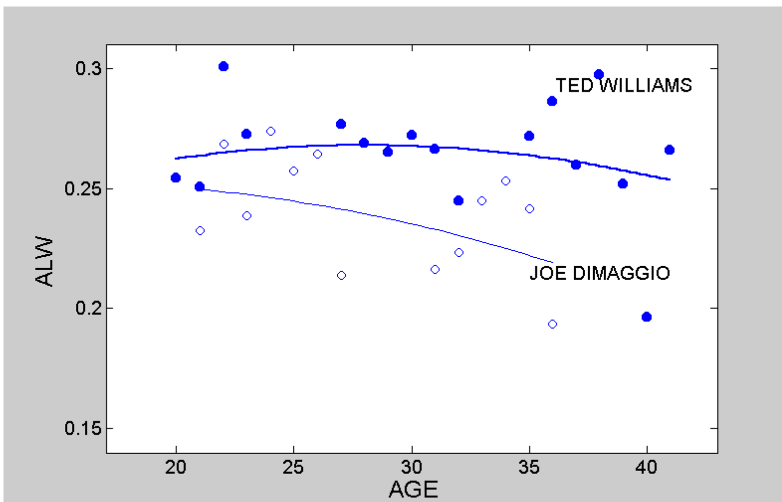

Figure 11
Scatterplot of ALW and fitted trajectories for Ted Williams and Joe DiMaggio.

Table 5 give the hitting statistics and Figure 12 plots the estimated career trajectories for Mickey Mantle and Willie Mays, two great hitters who were born in the 1930's. Here the comparison is not quite as clear as it was for Williams and DiMaggio. Mantle's estimated peak ability is a bit higher than Mays, but Mays sustained his pattern of great hitting for a long time.

Table 5: Average linear weight, number of plate appearances, and age for Mickey Mantle and Willie Mays for the seasons of their careers.

	MICKEY MANTLE		WILLIE MAYS	
Age	ALW	PA	ALW	PA
19	0.1948	384		
20	0.2268	624	0.2045	523
21	0.2189	540	0.1812	144

22	0.2272	645		
23	0.2534	633	0.2673	633
24	0.2832	647	0.2622	663
25	0.2784	620	0.2302	647
26	0.2509	650	0.2539	662
27	0.2223	636	0.2451	679
28	0.2354	639	0.2403	642
29	0.2752	640	0.2313	660
30	0.2602	500	0.242	655
31	0.2588	212	0.2491	703
32	0.2483	564	0.2394	664
33	0.2049	434	0.2457	661
34	0.2288	390	0.259	634
35	0.204	548	0.2304	624
36	0.1936	542	0.196	539
			0.212	567
			0.1962	455
			0.2206	560
			0.2214	532
			0.1968	305
			0.1625	237

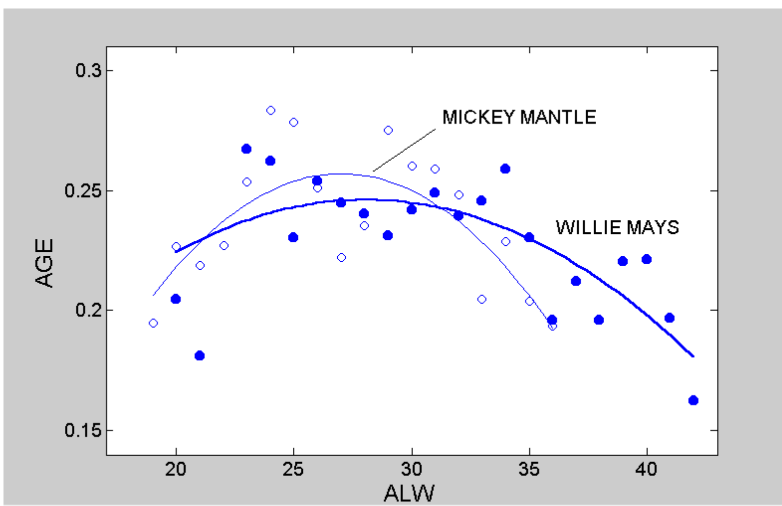

Figure 12
Scatterplot of ALW and fitted trajectories for Mickey Mantle and Willie Mays.

Last, we compare Pete Rose and Tim Raines, who were both great contact hitters in the modern era. Table 6 displays the hitting statistics for Rose and Raines, and Figure 13 shows the estimated trajectories. Although Rose is commonly thought by baseball fans to be the superior hitter, this figure seems to indicate that the two hitters had very similar trajectories. Most fans believe that Pete Rose would be easily elected to the Hall of

Fame if he were eligible. If so, then this analysis indicates that Tim Raines is also deserving of election to the Hall of Fame.

Table 6: Average linear weight, number of plate appearances, and age for Pete Rose and Tim Raines for the seasons of their careers.

	PETE ROSE		TIM RAINES	
Age	ALW	PA	ALW	PA
			0.0938	26
			0.2007	360
22	0.1716	683	0.1752	724
23	0.1548	554	0.1996	714
24	0.2007	747	0.2014	711
25	0.1988	692	0.2127	659
26	0.1973	644	0.2145	660
27	0.2091	686	0.2315	624
28	0.2269	720	0.1925	484
29	0.2082	724	0.1981	613
30	0.1926	703	0.1875	530
31	0.1929	725	0.1706	697
32	0.2006	751	0.1904	632
33	0.1884	763	0.2142	482
34	0.2026	762	0.1899	446
35	0.2065	757	0.1946	575
36	0.1964	726	0.2103	236
37	0.1918	720	0.2091	312
38	0.2039	725	0.1889	379
39	0.1714	727	0.1701	161
40	0.1864	480		
41	0.1652	707	0.211	107
42	0.1461	547		
43	0.167	417		
44	0.1737	495		
45	0.1438	271		

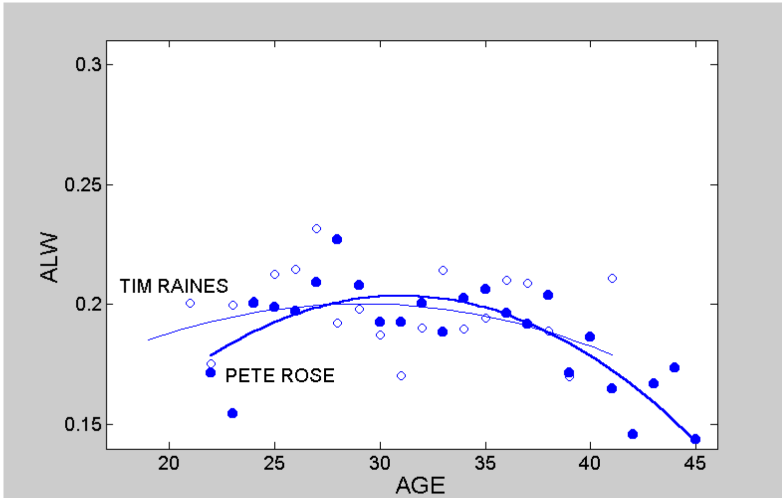

Figure 13
Scatterplot of ALW and fitted trajectories for Tim Raines and Pete Rose

9. Related work

Much of the sabermetrics literature is devoted to the evaluation of a player by use of his season statistics. A player's career batting average that is commonly quoted in the media is a relatively poor measure of average performance since it ignores a player's career trajectory and the average will underestimate a player's peak ability. James (2001), in his evaluation of the best players of all time, implicitly assumes that players have career trajectories by taking the mean of the win shares of a player's five best consecutive seasons as one of his measures of performance. James (1982) discusses the career projession of players and gives evidence that players generally peak at age 27. He compares his research with that of Pete Palmer, who found that ballplayers achieve constant level performance from ages 23 to 40. James explains that there is a bias in Palmer's findings, since only the better hitters and pitchers are playing at advanced ages. Schell (1999) adjusts his batting averages of historical players by a "longevity adjustment" that truncates a player's hitting data at 8000 at-bats. This adjustment was made to account for the decreasing performance in players' career trajectories at the end of their careers. Schall and Smith (2000) recently discuss the observed career trajectories for hitters and pitchers. One of their objectives of their study was to see if one could predict a player's career length on the basis of his performance in his rookie season.

From a modeling perspective, Morris (1983) estimated Ty Cobb's batting average trajectory. In this paper, he illustrated the use of empirical Bayes procedures to shrink Cobb's observed batting averages towards a quadratic fit curve. Albert (1992) used a random effects model to smooth the career trajectory of a batter's home run rates. Berry et al (1999) performed an extensive study in which they estimated the career trajectories for athletes in baseball, hockey, and golf. They used a nonparametric aging function in there modeling in contrast to the quadratic function used here. Using their model, they rated the top 25 hitters of all time using the criteria of batting average and home run rate. As noted by Albert (1999), Berry et al (1999) make several questionable assumptions –

they assume at each player peaks at the same age and that the maturing and declining period is the same across all players. One advantage of the parametric modeling of this paper is that one obtains smooth estimates of the career trajectories and the characteristics of the trajectory (the peak height and the peak age) are defined in terms of the regression parameters.

Appendix: Posterior calculations for the Bayesian exchangeable model.

The model

The exchangeable model introduced in Section 5 can be defined in three stages. At the sampling stage, the regression estimate for the *i*th player, $\hat{\beta}_i$, is assumed to have a multivariate normal distribution with mean β_i and known variance matrix V_i.

1. $\hat{\beta}_i$ distributed $N(\beta_i, V_i), i = 1, \ldots, p$

At the second stage, one assumes that the regression parameters for the *p* players, $\beta_1, \ldots,$ β_p, are a random sample from a multivariate normal density with mean β^0 and variance Σ.

2. $\beta_1, \ldots, \beta_p$ distributed $N(\beta^0, \Sigma)$

The locations of the hyperparameters (β^0 , Σ) at the second stage are unknown, and so they are assigned a uniform density at the third stage of the model.

3. (β^0, Σ) distributed from the density $g(\beta^0, \Sigma) = c$.

The posterior distribution

There are *p+2* parameters in the model, the *p* regression vectors $\beta_1, \ldots, \beta_p$, and the hyperparameters (β^0 , Σ) that describe the common distribution for the regression vectors. It is convenient to represent the posterior distribution of the complete set of parameters as the product

$$g(\beta_1, \ldots, \beta_p, \beta^0, \Sigma \mid data) = g(\beta_1, \ldots, \beta_p \mid \beta^0, \Sigma, data)\ g(\beta^0, \Sigma \mid data)$$

where the first term in the product is the joint posterior distribution of the regression vectors conditional on values of the hyperparameters, and the second term in the product is the posterior distribution of the hyperparameters. If we are given values of the hyperparameters (β^0 , Σ), then, using standard results for linear models, the posterior distributions of $\beta_1, \ldots, \beta_p$ are independent where β_i is distributed $N(\beta_i^*, V_i^*)$ where

$$\beta_i^* = \left(V_i^{-1} + \Sigma^{-1}\right)^{-1} \left(V_i^{-1}\hat{\beta}_i + \Sigma^{-1}\beta^0\right), \quad V_i^* = \left(V_i^{-1} + \Sigma^{-1}\right)^{-1}.$$

The posterior density of the hyperparameters is given by

$$g(\beta^0, \Sigma \mid data) = \prod_{i=1}^{p} \left|V_i + \Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(\hat{\beta} - \beta^0)'(V_i + \Sigma)^{-1}(\hat{\beta} - \beta^0)\right\}$$

Simulating from the posterior distribution

The joint posterior distribution of ($\beta_1,\ldots,\ \beta_p, \beta^0$, Σ) can be simulated by first simulating a value of (β^0 , Σ) from its posterior distribution, and then simulating values of $\beta_1,\ldots,\ \beta_p$ from the conditional posterior distribution. Since the posterior distribution of (β^0 , Σ) has a nonstandard form, the independence Metropolis algorithm (Robert and Casella, 1999) is used to simulate from this simulation using a suitable proposal density.

References

Albert, J. (1992), “A Bayesian analysis of a Poisson random effects model for homerun hitters,” *The American Statistician* 46, 246-253.

Albert, J. (1999), Comment to "Bridging Different Eras in Sports," *Journal of the American Statistical Association*, 94, 677-679.

Albert, J. and Bennett, J. (2001). *Curve Ball*, Copernicus Books.

Berra, A. (2002). *Clearing the Bases: The Greatest Baseball Debates of the Last Century*, Dunne books.

Berry, S. M., Reese, S. C., and Larkey, P. D. (1999), "Bridging Different Eras in Sports," *Journal of the American Statistical Association*, 94, 661-678.

Cleveland, W. S. (1979). “Robust Locally Weighted Regression and Smoothing Scatterplots,” *Journal of the American Statistical Association*, 74:829-836, 1979

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman and Hall.

James, B. (1982). *The Bill James Baseball Abstract*, Ballantine Books.

Morris, C. (1983). “Parametric Empirical Bayes Inference: Theory and Applications,” *Journal of the American Statistical Association*, 78, 47-55.

Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer.

Schall, T. and Smith, G. (2000), “Career Trajectories in Baseball,” *Chance*, 13, pages

Schell, M. J. (1999) *Baseball All-Time Best Hitters*, Princeton University Press.

Thorn, J., Palmer, P., and Gershman, M. (2001). *Total Baseball: The Official Encyclopedia of Major League Baseball*, Total Sports.

Thorn, J. and Palmer, P. (1984). *The Hidden Game of Baseball*, Doubleday.