By the Numbers

Volume 12, Number 3 The Newsletter of the SABR Statistical Analysis Committee August, 2002

Review

"Win Shares" Has Meat On Its Bones

Robert T. Allen

Bill James latest work, "Win Shares,"has much material to digest. Five years in the making, it is a welcome addition to the sabermetrician's bookshelf.

Preceded by a presentation at last year's SABR convention and then a brief treatment in *The New Bill James Historical Baseball Abstract,* the full exposition of James' long-awaited Win Shares metric finally arrived in April, 2002. Was it worth the wait?

There is considerable meat on the bones of this new measure, and much material to be digested in the book. The Win Shares statistic is an attempt to place all of a player's contributions, both offensively and defensively, within the context of his team's bottom-line accomplishment: its wins. The process starts with the arbitrary assignment of three Win Shares for each game won – if a team wins 85 games, it has 255 Win Shares to be divvied up among all the players. Through a series of formulas, some staggering in their complexity, this total is first divided into offense and defense, and then the defense's portion is further broken down into pitching and fielding. At that point, the apportionment to individual players in each category begins. When the process is completed, each player's assigned shares for batting, fielding and pitching are recombined into a single number which has historically ranged from 0 to 59 (the 59 belonging to Honus Wagner in 1908, the post-1900 record). In most seasons, 30 Win Shares would probably get a player into the MVP race, 25 could suggest an All-Star, and 15-20 might represent a regular on a winning team.

The first three sections of the book are devoted to describing the rationale and formulas. To illustrate the process, James shows how the formulas work through for three actual teams from three different eras: the 1890 New York Giants, the 1932 Athletics, and the 1998 Cardinals. This portion of the book makes for heavy going, but one comes through it with an appreciation for the thought that went into each step along the way. Many of the component parts will be familiar to readers of James' body of work. Runs Created is the foundation on the offensive side, while the portions concerned with fielding bear a strong resemblance to his work on Defensive Winning Percentages used in annual *Abstracts* of the early 1980s.

It is in applying Win Shares to pitching that James is perhaps on shakiest ground. Apart from the methodology, which rests on an unusual number of arbitrary assumptions, especially with regard to the value of saves, the problem can be seen in the results. By far the highest yearly totals for any players occur among 19th century pitchers – in fact, the 14 highest annual Win Shares performances are all by pitchers who performed in the 1880s. In contrast, today's starting pitchers, working longer schedules but in many fewer innings, don't have a prayer of matching those totals, or even competing with hitters. Pedro Martinez, for his amazing seasons in 1999 and 2000, is credited with 27 and 29 Win Shares, figures exceeded by 9 and 6 AL batters in those years. Oddities such as this hamper the application of the WS concept to career totals without the use of some type of "time line", as the author himself did in his updated *Historical Abstract*.

Although the fielding portion of Win Shares is the smallest by weight (typically about a third of the defense, which itself is generally half of the total), it is here that much of the detailed and original work is done. The formulas differ for each position, and they attempt to take account of more than just the customary counting categories (Putouts, Assists, Errors and DPs). Primary

In this issue

calculations are done at the team level, with adjustments for such factors as the composition of the pitching staff (LHP/RHP, ground ball/fly ball tendencies), the opportunity for double plays, the interaction between positions, and even park effects – possibly the first time the latter have been incorporated in fielding measures. To increase the complexity, for some positions there are different formulas to be used for different eras.

Once past these initial sections, James offers the reader an entertaining series of "Random Essays," wherein he applies the fruits of his Win Shares research to various general questions, such as MVP and Rookie of the Year selections, Gold Glove teams, trades, team age analysis, career projections, and so on. In some of these pieces, he tries to anticipate questions that may arise as the concept becomes familiar. How, for example, if the system is fair to all, can players with virtually identical sets of statistics wind up with different Win Shares? His answers to this question, and others, may not satisfy every challenge, but they don't have to for the overall concept to be a useful one.

The final two-thirds of the book serve as a Win Shares encyclopedia. First there is a listing of Win Shares for every significant player (1 or more WS) on every major league team from 1876 through 2001. This is followed by tables showing the year-by-year WS for prominent players in each decade, then career totals for every player not included in the decade summaries. Next are various leader boards, for batting, pitching and fielding; franchise leaders; and annual leaders in each category for each season. Finally, there are detailed breakdowns showing how each team fared with regard to Win Share components in the 2001 season.

All in all, this project represents a prodigious undertaking, one that the author claims was at least five years in the making, covering the thinking, the research and finally the writing. Only time will tell whether the Win Shares concept becomes a standard arrow in the quiver of sabermetric tools or is supplanted by even more complex measures yet to come. One problem is that it is not a procedure that can be easily carried forward, unless an organization such as STATS, Inc. elects to include it in an annual publication. The task of setting up the various spreadsheets necessary to carry out all the calculations on one's own is daunting, as anyone who attempts it by following the book will quickly discover. But swear by it or not, *Win Shares* is a welcome (and perhaps landmark) addition to the sabermetrics bookshelf.

Robert T. Allen, 6917 Wrentree Dr., Charlotte, NC, 28210, JustBob898@aol.com ♦

Baseball At Altitude – SABR 33

Rod Nelson

The 33rd Annual SABR National Convention will be held July 10-13, 2003 at the Mariott City Center Hotel in Denver, Colorado. This provides an excellent opportunity for baseball's best and brightest statistical analysts to address the phenomenon that is "Baseball at Altitude". After ten years of major league play, the subject is still something of an enigma and presents a most complex challenge to the field of sabermetrics. Whether attempting to simply correlate data from games played at Mile High Stadium and Coors Field or charged with developing an organizational philosophy whose objective is to build a consistent winner, adding the altitude factor to the equation presents a formidable task like none other. The Rocky Mountain Chapter invites all SABR members to present their observations of the extraordinary game which is played in such a unique environment and the problems dealing with the disparity of home and road conditions.

The SABR33 Organizing Committee is working on several concepts for panel discussions that will be of particular interest to members of the Statistical Analysis Committee. The "Baseball at Altitude" panel will feature a cross-section of baseball experts on physics, statistics, history, and player personnel. Others intriguing topics include "The Relief Pitcher and the Hall of Fame", "Baseball Simulation Games" and others.

Members of the Statistical Analysis Committee are invited to develop presentations or poster displays of their work with the possibility for post-convention publication of papers on the "Baseball at Altitude" theme. The Presentations Chair is Gail Rowe (growes36@attbi.com). Watch SABR-L and the SABR Bulletin for additional details.

Academic Research: Disabled List Trends and Racially-Motivated HBP

Charlie Pavitt

The author describes two more academic studies, one on player disability rates, the second on whether hit batsmen have a racially-motivated component .

This is one in a series of occasional reviews of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at *www.udel.edu/johnc/faculty/pavitt.html*, use it for your research, and let me know what I'm missing.

Stan Conte, Ralph K. Requa, and James G. Garrick, Disability Days in Major League Baseball, American Journal of Sports Medicine, Volume 29 Number 4, 2001, pages 431-436

John Matthew IV was kind enough to send me a copy of this study. The first author is the Giants' head trainer, while the other authors are associated with a sports medicine center in San Francisco. The authors used major league disabled list data to examine the total number of disabled players during each season from 1989 through 1999, along with the number of days these players were disabled. Despite advances in sports medicine and training techniques, there has been a statistically significant increase over time in the average number of days for which teams disabled players, and an insignificant increase in the average number of days that players spent on the disabled list. Pitchers seem to be responsible for much, but not all, of the increase. There has also been a significant increase in the total number of disabled players and total number days for which players were lost, although in these cases the data presented does not distinguish between an increase due to actual injuries and an artifactual increase due to the addition of four expansion teams over that time period. These findings leave open the question of how improved medical techniques could be correlated with more disablement. In answer, the authors speculate that advances in sports medicine and training may be allowing players to continue playing despite injuries that would have formerly ended their careers, but that these players would be more susceptible to injury and thus relatively frequent visitors to the disabled list. They claim that the absence of specific diagnoses in the information present in disabled lists make it impossible for the authors to evaluate this speculation. My first impression is to agree to the extent that we don't have any precise way of knowing how many careers have been ended by injury over the years. However, it seems to me that a study of patterns of disabled list stretches for players with relatively long careers would be relevant to this issue.

Thomas A. Timmerman, Violence and Race in Professional Baseball: Getting Better or Getting Worse? Aggressive Behavior, Volume 28, 2002, pages 109-116

I am tired of all the articles on sports and race asking the same questions about discrimination over and over again, seemingly functioning as little more than an easy way to pad some academic's vitae. This one is a refreshing exception, using batter hit-by-pitch data to explore whether Blacks and Hispanics are more susceptible to this form of "covert aggression" than Whites. The author finds that Black and Hispanic players were indeed more likely to be hit during the 1950s and 1960s, but that this discrepancy disappeared in the 1970s and 1980s, and Blacks were actually hit less than Whites and Hispanics during the 1990s. Further, during 1997, 1998, and 1999 seasons there was no relationship between the race of the pitcher and the race of the batter in hit-by-pitch events. This article serves as another piece of evidence from, thankfully, a new source of information implying that racism in baseball is slowly but surely dying.

Charlie Pavitt, 812 Carter Road, Rockville, MD, 20852, chazzq@udel.edu ♦

The Impact of Lineup Balance on Scoring, 1920-1989

Cyril Morong

Suppose two teams have exactly equal aggregate offensive stats. Now, suppose one team has nine players of roughly equal ability, but the other has some very strong hitters and some very weak hitters. Which team should score more runs? In this study, the author looks back at 70 years of baseball records to search for evidence on this question.

This paper analyzes whether or not a more balanced lineup, holding overall team batting percentages (on-base percentage or OBP, slugging percentage or SLG, etc.) constant, has, in the past, increased run scoring.

Team balance (BAL) is the sample standard deviation (SD) of various hitting statistics like OBP and SLG calculated for a team's most frequently used eight position players. Only teams that had eight players with at least 400 at-bats at each of the non-pitching positions according to the 8th edition of the Macmillan Baseball Encyclopedia were included in the study. This allows for a fairly constant lineup throughout the season. There were 50 such teams in this time period.[1]

The question was analyzed using ordinary least-squares regression analysis. In the first regression, the dependent variable was team runs per game (R/G). The independent variables were team OBP, team SLG and the league error rate (ER).[2] The ER is added since it varied over the period and impacts scoring. For example, it was much different in the 1921 AL (.035) than it was in the 1982 AL (.020). I assumed that the error rate committed against all teams in a league in a season is the same. This is not perfect, but it is an improvement over not including an error rate variable at all.

The results for the first regression, which used the 50 teams in the study:

(1) R/G = -7.34 + 24.97*OBP + 8.39*SLG + 11.01*ER

The r-squared was .918, meaning that 91.8% of the variation across teams in R/G is explained by the independent variables. The T-values for the 3 independent variables were 10.42, 5.97, and 1.44, respectively. When this regression was extended to all teams from 1920-98, instead of just the original 50 teams, the equation was:

(2) R/G = -5.87 + 17.63*OBP + 10.7*SLG + 13.51*ER

The r-squared was .922. In this case ER had a much higher T-value (11.11). The other big difference was the coefficient value of OBP. It was much lower than for the group of 50 teams that I am working with here.

When the balance variables were added to the first regression, the results were:

(3) R/G = -7.04 + 25.40*OBP + 6.96*SLG + 11.60*ER −3.54*BAL/OBP + 3.30*BAL/SLG

The r-squared was .923, not much higher than equation (1). So adding in variables to represent balance does not add much to our ability to explain R/G. Neither BAL variable was significant, with T-values of -0.89 and 1.58. But having a more balanced team in OBP increased scoring since the sign on the coefficient is negative. The higher the standard deviation, the less balanced the team is in OBP and the lower the scoring. BAL/SLG is the opposite. It helped to be less balanced in SLG.

The mean standard deviation of OBP for the eight players on each team (or BAL/OBP) is .036. The extreme high and low both differed from it by about .020. Multiplying this times the 3.54 coefficient from equation (1) we get -.07 R/G. For 162 games this is about 11 runs. But 38 teams were within .010 of the .036 mean standard deviation for OBP. So for those teams this is a difference of five runs or less per season. The difference between the most balanced and least balanced teams is about .040 or 22 runs a season. The standard deviation of BAL/OBP was .0137. Multiplying this by –3.53 gives -.048 or -7.84 runs per season. So increasing your balance by one standard deviation added 7.84 runs per season. This does not seem to be a large effect.

The mean standard deviation for SLG (or BAL/SLG) is .070. The extreme high differed from it by about .053. Multiplying this times the 3.30 coefficient we get .26 R/G. For 162 games this is about 28.33 runs. So the team that was the least balanced in SLG scored 28.33 more

[1] Seven teams that played under the DH-rule were used. They each had nine players with 400 or more at-bats. There were two non-DH teams that actually had nine players with 400 or more at-bats, the 1937 Pirates and the 1971 Tigers. They were not used.

[2] The error rate is simply 1 minus the fielding percentage for the entire league in the year in which a given team played.

runs than the average team. For the most balanced team, the difference was .025. This cost them about 13 runs a season. But 39 teams were within .020 of the .070 mean standard deviation for SLG. So for those teams this is a difference of 10.69 runs or less per season. But all this says is that it paid to be unbalanced in SLG. The standard deviation of BAL/SLG was .041. Multiplying this by 3.30 gives .135 or 21.88 runs per season. So *decreasing* your balance by one standard deviation added 21.88 runs per season. This seems like a large effect.

Regression (3) was run using isolated power, total bases divided by at-bats and extra bases divided by at-bats in place of SLG. The results were generally similar. Adding in the BAL variables did not increase the r-squared very much. The signs on the coefficients were the same. None of the BAL variables had T-values of 2 or more (or even close), so they were not significant. But the coefficient on BAL/OBP in the regression that used isolated power was much higher than in the other regressions at –5.32. Multiplying that by the standard deviation of BAL/OBP of .0137 gives -.073 or –11.81 runs a season. So increasing your balance by one standard deviation added 11.81 runs per season.

One regression used OPS instead of OBP and SLG. Again, BAL had little impact on the model. In fact, less balance added more R/G. The range of OPS from the highest to lowest hitter on each team was also used as a BAL variable. It also had little impact on the model. In fact, the coefficient on the BAL variable was positive, meaning that the bigger the range in OPS from the highest to lowest hitter the higher the R/G.

One problem with the 50 teams in the data set is that they generally scored more runs than the average team in their league -- on average, 6% more. Only 16 of the 50 teams were below average in R/G for their league. So I ran a regression using the 32 lowest scoring teams (so there was an equal number of teams above and below their league average R/G). In that regression, the independent variables were team OPS, ER, and BAL/OPS. As with the above analysis, the BAL variable had little impact on the r-squared. It was not significant. Its coefficient was negative, but only -.29. This would cause a difference of less that .5 runs per season between he most and least balanced teams in OPS.

Then I ran a regression on those same 32 teams that was the same as (3). The results were:

```
(3A) R/G = -7.33 + 28.79*OBP + 5.98*SLG + 6.16*ER – 7.21*BAL/OBP + 0.12*BAL/SLG
```

In this case of the 32 lowest scoring teams, the coefficient is much stronger on BAL/OBP. But 24 of those 32 teams are within .010 of the mean standard deviation for OBP (or BAL/OBP) of .034. That .010 means a difference of 11.68 runs per season. The standard deviation of BAL/OBP was .022. Multiplying this by –7.21 gives -.158 or -25.66 runs per season. So increasing your balance by one standard deviation will add 25.66 runs per season. This shows a big advantage for making your lineup more balanced in OBP. The coefficient on BAL/SLG is very slight, indicating that being balanced in SLG has little effect one way or another.

To conclude, it seems that having a more balanced lineup generally has had little positive impact on scoring. The only analysis that supports the importance of balance is the last one which had a limited number of teams and that was only for OBP.

Cyril Morong, 723 W. French Place, San Antonio, TX 78212, cyrilmorong@aol.com ♦

Get Your Own Copy

If you're not a member of the Statistical Analysis Committee, you're probably reading a friend's copy of this issue of BTN, or perhaps you paid for a copy through the SABR office.

If that's the case, you might want to consider joining the Committee, which will get you an automatic subscription to BTN. There are no extra charges (besides the regular SABR membership fee) or obligations – just an interest in the statistical analysis of baseball.

To join, or for more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. Or write to him at 4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.

The Essential Sabermetric Library

Randy Klipstein

What books, magazines, or websites are the most essential for statistical baseball research? The author asked this question of several prominent sabermetricians, and presents the results in this article.

Charlie Pavitt's review of *Curve Ball* (*By the Numbers*, February 2002) started me thinking about what are the most important statistical analysis baseball books ever published. Years ago, a SABR publication compiled the essential baseball library (*The SABR Review of Books*, *Volume II*). I decided to do the same for the field of baseball statistical analysis.

Contributors: Clem Comly, Duke Rankin, Pete Palmer, Tom Ruane, Mark Pankin, Rob Wood, Charlie Pavitt

I queried a bunch household names (at least in my house) in this field, for their lists; the sources that they turn to most frequently or that most influenced them. I broadened the definition of library to include not just books and periodicals, but also Internet sites, software, or any other electronic medium. I received seven responses, with as few as three, and as many as twenty items. There were thirty-eight distinct nominations of books, periodicals, and web sites; and a couple of people added a second list of 'nice to have' but not essential entries. Books with more than one edition or annual publications were counted as one.

Entries that were mentioned at least twice give us a library of sixteen, a manageable number. We'll call this the Essential Sabermetric Library. It includes eight books, four annual publications, one periodical, and three web sites.

The Hidden Game of Baseball, Total Baseball, and James' historical and annual abstracts were the most mentioned titles. Of *The Hidden Game*, Mark Pankin said it "is still the best book in the field." Of James, and the 1982 *Abstract* in particular, Duke Rankin writes, "I suppose a book about baseball is, by definition, not terribly important, but I would argue the 1982 Abstract is one of the 100 most important books of the 20th century because of the originally and insight of the work, and the profound influence it's had on at least one portion of our society."

On *Curve Ball*, Mark Pankin writes, "It has a nice overview of the nature of performance models, but is often somewhat elementary. Would be useful for a novice wanting to learn about the subject."

The Essential Sabermetric Library

Books:

Curve Ball, by Jim Albert and Jay Bennett
Percentage Baseball, by Earnshaw Cook
The (Old and New) Bill James Historical Baseball Abstract, by Bill James
The Diamond Appraised, by Craig Wright and Tom House
The Hidden Game of Baseball, by Pete Palmer and John Thorn
The Politics of Glory, by Bill James
Total Baseball
Win Shares, by Bill James

Annuals:

Baseball Prospectus
The Big Bad Baseball Annual
The Bill James Baseball Abstract
The Elias Baseball Analyst

Periodicals and web sites:

By the Numbers
www.baseball-reference.com
www.baseball1.com
www.retrosheet.org

Charlie Pavitt characterizes *Percentage Baseball* as a "flawed, but classic and thought provoking work."

Retrosheet's web site was the most often cited Internet destination, particularly for its game logs for every game in the 20th century and play-by-play data for all games since 1974. Sean Lahman's site, www.baseball1.com was noted for its historical stats that can be easily downloaded into a spreadsheet or data base application. Rob Wood said www.baseball-reference.com "is a great site for historical data; I treat it as similar to an electronic baseball encyclopedia."

The entries that were mentioned once are listed in the box "The Secondary Sabermetric Library."

Pete Palmer included *Baseball Weekly*, *The New York Clipper*, *The Sporting Life*, and *The Sporting News* as a source of box scores.

For additional sources of information, Tom Ruane mentions the guides and registers; various publications from STATS, Inc; Baseball America's Draft Book; and Marshall Wright's minor league statistic books.

Duke Rankin adds some offbeat recommendations:

The Physics of Baseball, by Robert Adair. "I spent an afternoon hitting fungoes to examine his views on the flight of baseballs. The analysis really is counter-intuitive – I think it is possible for baseballs to increase speed after the impact with a bat, and softer bats are better than harder bats. Read it; it will blow your mind."

It's What Your Learn After You Know It All That Counts, by Earl Weaver. "As a Yankee fan, I hated Earl Weaver, but he sure could win ball games, and he talks about his philosophy in his book. It might be the best insight into a great manager's mind available on the market."

Rob Neyer's e-column at *sports.espn.go.com/mlb/neyer/index*. "Rob has his shortcomings, but he is probably the highest profile sabermetrician on the web, and he does try to implement sabermetric analysis to current questions."

Finally, even after both these lists, if you're still searching for more reading, the bibliography in *Curve Ball* offers a number of additional sabermetric sources.

In looking over these lists, I am struck by how much of this material was published in the recent past, within the past year or two. I will spare you any statistical analysis, though. It is this vibrancy that makes this field so fascinating and leads me to believe that many more essential works are to be published.

The Secondary Sabermetric Library

Books:

Baseball by the Numbers, by Willie Runquist
Green Cathedrals, by Philip Lowry
Optimal Strategies in Sports, edited by S. P. Ladany and R. E. Machol
STATS All-Time Major League Handbook
STATS All-Time Major League Sourcebook
The Bill James Guide to Baseball Managers, by Bill James
The Home Run Encyclopedia
The Macmillan Baseball Encyclopedia

Annuals:

Baseball Sabermetric
Great American Baseball Stat Book
Mike Gimbel's annual player and team rating books

Periodicals:

Baseball Research Journal
Baseball Weekly
Chance
The New York Clipper
The Sporting Life
The Sporting News

Web Sites:

www.astrosdaily.net
www.baseball-links.com
www.baseballprimer.com
www.baseballprospectus.com
www.stathead.com

Randy Klipstein, 65 Landing Drive, Dobbs Ferry, NY 10522, r.klipstein@verizon.net ♦

E-mail Changes

If you normally receive "By the Numbers" by e-mail, but you found this issue in your physical mailbox instead, it's probably because your e-mail address changed. If you'd like to switch back to an e-mail BTN, please drop me (Phil) a line with your new e-mail address, and I'll switch you back to the electronic version. I'm at birnbaum@sympatico.ca.

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis - that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any - I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics - confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Jim Box	im.box@duke.edu	Statistics
Keith Carlson	kcarlson2@mindspring.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
Keith Karcher	kckarcher@compuserve.com	General
Chris Leach	chrisleach@yahoo.com	General
John Matthew IV	john.matthew@rogers.com	Apostrophes
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Smoothing Career Trajectories of Hitters

Jim Albert

Prior studies, as well as theory, suggest that young players improve as they play major-league baseball, until a certain point where age catches up and they begin their decline. But for many players, their careers do not appear to follow this archetypical rise and fall. In this study, the author attempts to reconcile the empirical data with the theory, by showing how a rise-and-fall path can be inferred from a player's actual batting statistics.

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.

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.

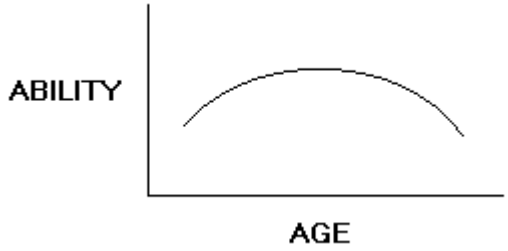

Figure 1

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).

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	Greenberg, J. DiMaggio, Ted Williams
1920-1929	50	Kiner, Snider, Musial
1930-1939	61	Mantle, Mays, Aaron
1940-1949	109	Schmidt, Stargell, Reggie Jackson
1950-1959	97	Brett, Eddie Murray, Rice
1960-1969	106	Barry Bonds, Sosa, McGwire

We will 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.

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. We will start with Thorn and Palmer's *Linear Weights* statistic, without the outs term. The negative contribution of outs has been omitted from the linear weights formula. The objective here was to use a reasonable measurement of the value of a player's on-base events.

$$LW^* = .46(1B) + .80(2B) + 1.02(3B) + 1.4(HR) + .33(BB+HBP)$$

Then, to convert the performance measure from a total to a rate, we will divide the LW* statistic by the number of plate appearances, obtaining the *average linear weight:*

$$ALW = \frac{LW^*}{AB + BB + HBP}$$

4. Quadratic regression model

We are interested in modeling a player's batting performance in terms of his age, where we use 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

$$\beta_0 + \beta_1 age + \beta_2 age^2$$

To summarize a particular quadratic fit, it is helpful to reparameterize $(\beta_0, \beta_1, \beta_2)$ by

$P = \beta_0 - \frac{\beta_1^2}{4\beta_2}$, the *peak value;*

$AGE^* = -\frac{\beta_1}{2\beta_2}$, the *peak age, and*

β_2 , the *curvature.*

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 (listed in the appendix). Figure 2 constructs a scatterplot of the (age, ALW) data and overlays the quadratic fit

$$-0.27 + 0.033 age - 0.00058 age^2$$

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.

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 *j*th 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.

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.

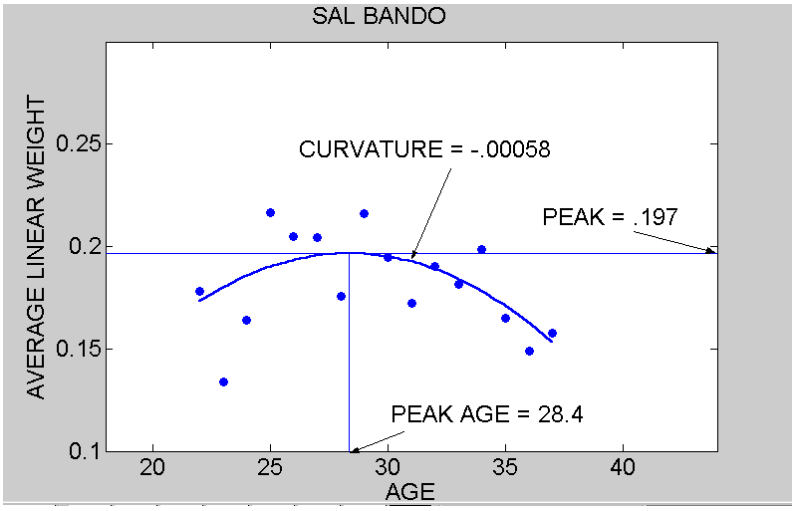

Figure 2 – Scatterplot of (age, ALW) data for Sal Bando with quadratic smoothing curve placed on top.

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).) 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.

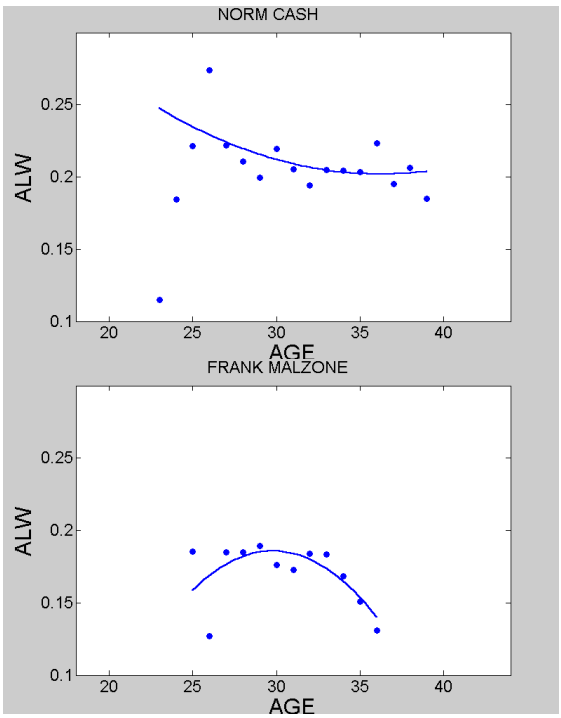

Figure 3: Scatterplots of (age, ALW) data and separate regression estimates (solid lines) for Norm Cash and Frank Malzone.

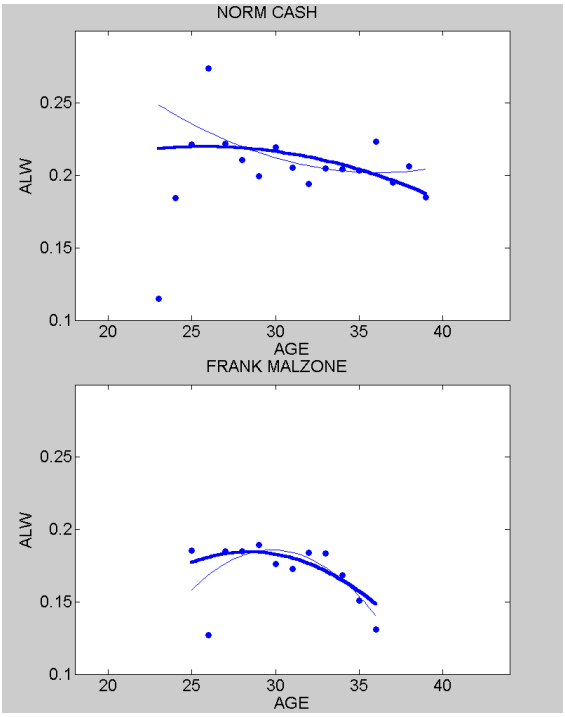

Figure 4 – Scatterplots of (age, ALW) and separate regression (thin line) and exchangeable (thick line) estimates for Norm Cash and Frank Malzone

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 2 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.

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.

Table 2 – 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 (x1000)
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)

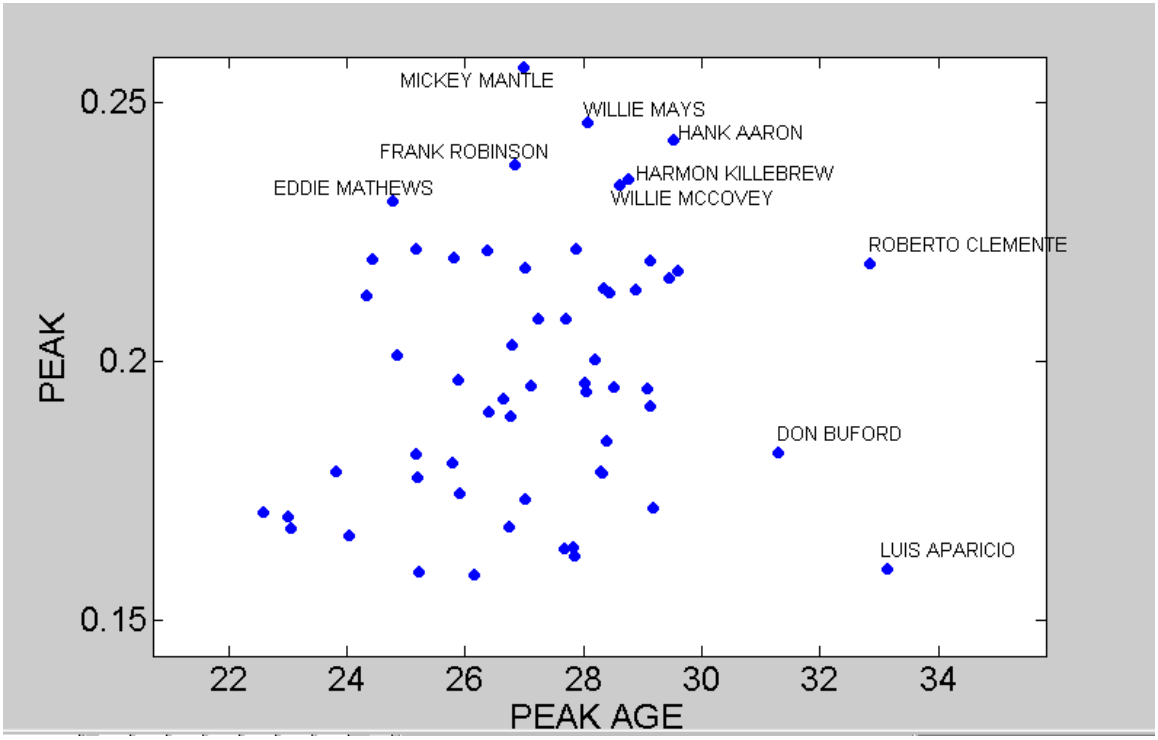

Figure 5 – Scatterplot of peak age and peak estimates for all players born in the 1930's with at least 5000 plate appearances

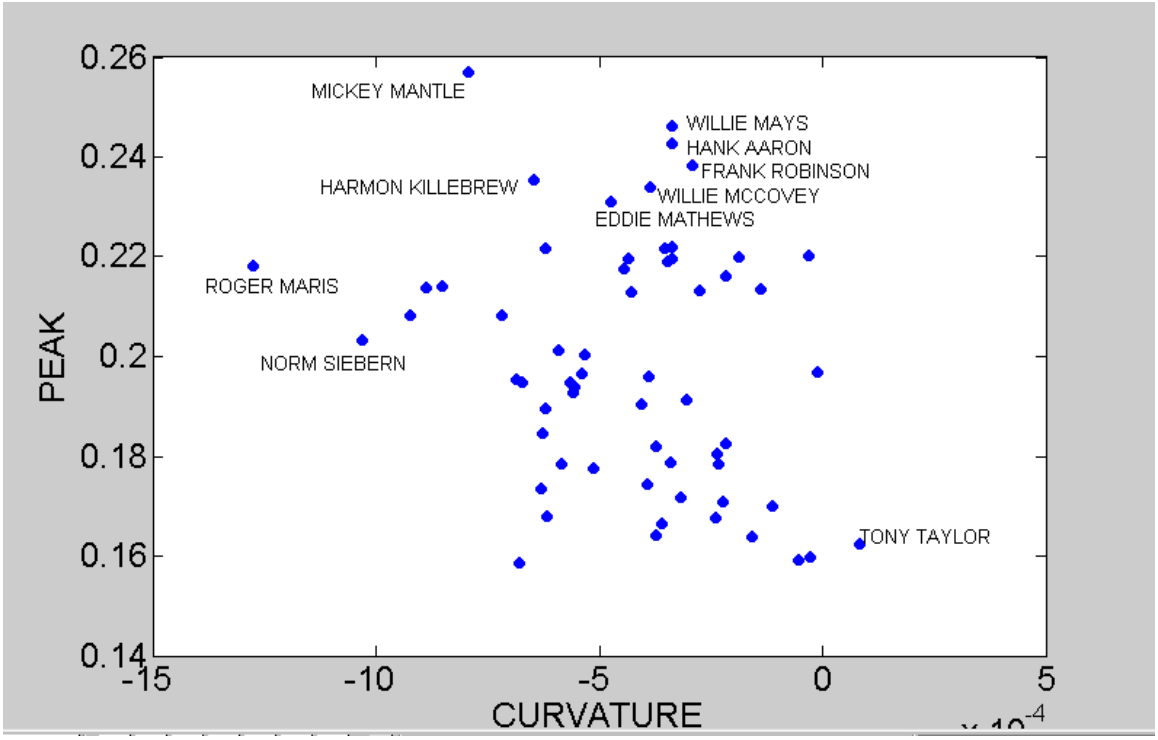

Figure 6 – Scatterplot of curvature and peak estimates for all players born in the 1930's with at least 5000 plate appearances

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.

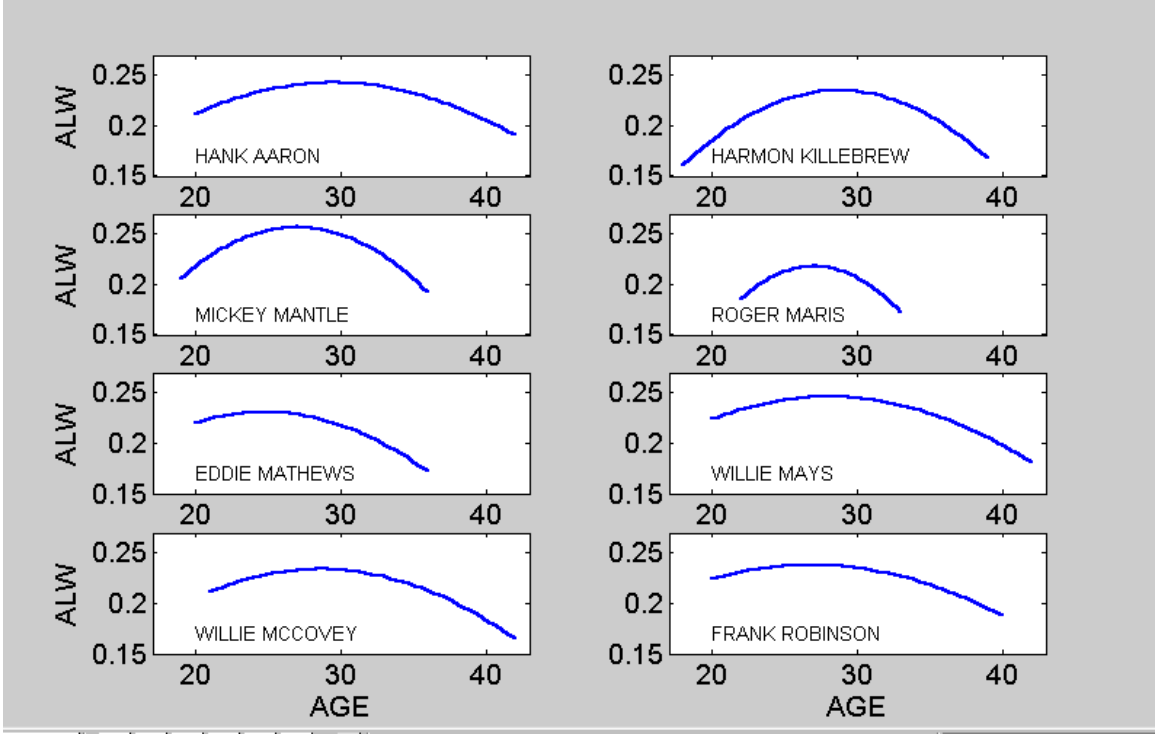

Figure 7 – Estimated career trajectories for eight great hitters who were born in the 1930's.

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.

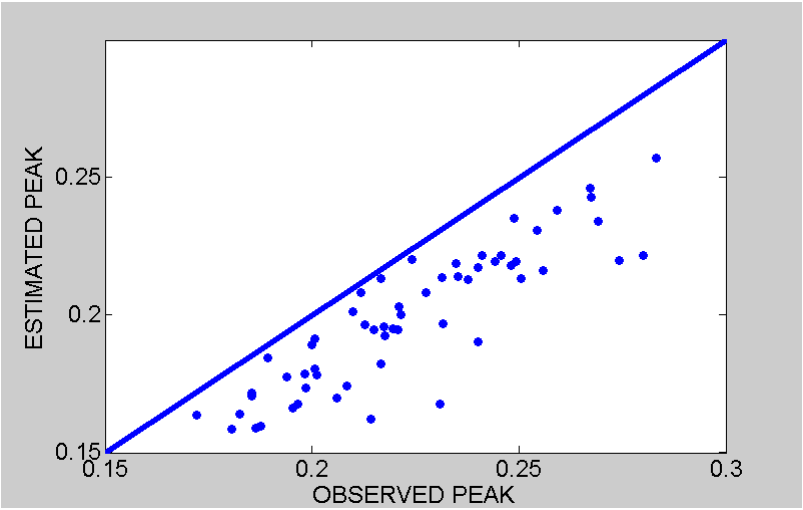

Figure 8 – Scatterplot of observed and model-based estimates of peak values for hitters born in the 1930's.

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.

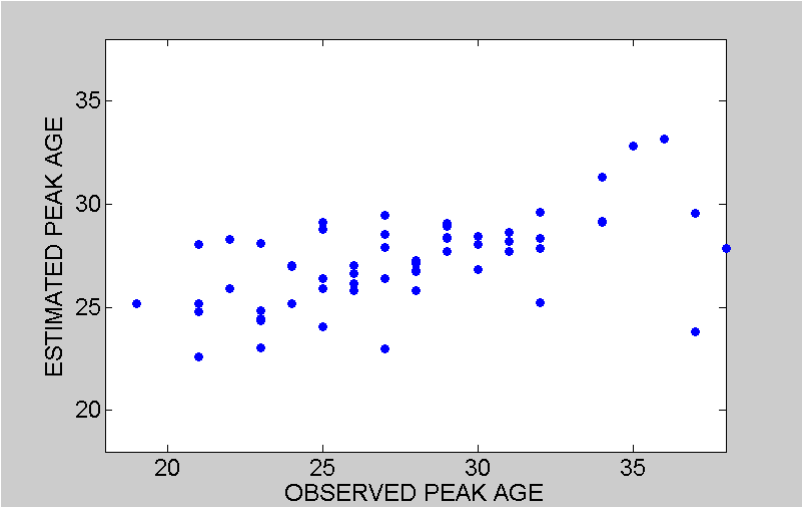

Figure 9 – Scatterplot of observed and model-based estimates of peak ages for hitters born in the 1930's

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

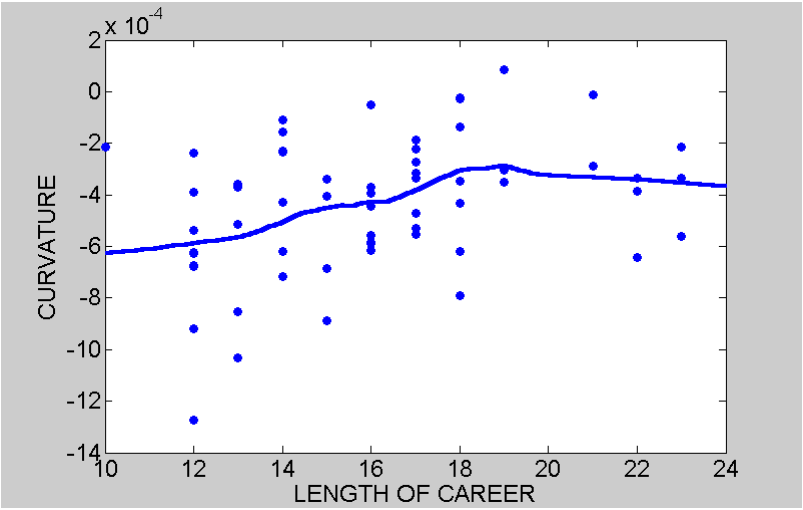

Figure 10 – Scatterplot of length of career and curvature estimates for hitters born in the 1930's. A lowess smoother is drawn on top of the scatterplot

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.

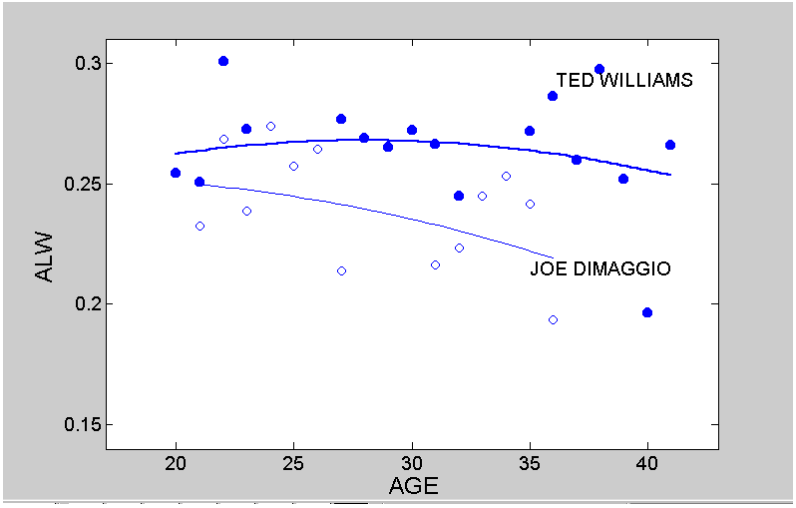

Figure 11 – Scatterplot of ALW and fitted trajectories for Ted Williams and Joe DiMaggio

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.

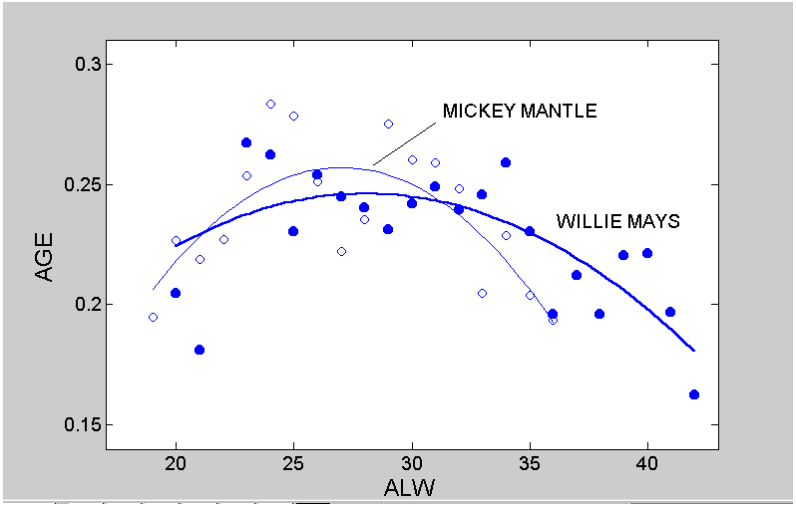

Figure 12 – Scatterplot of ALW and fitted trajectories for Mickey Mantle and Willie Mays

Last, we compare Pete Rose and Tim Raines, who were both great contact hitters in the modern era. 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.

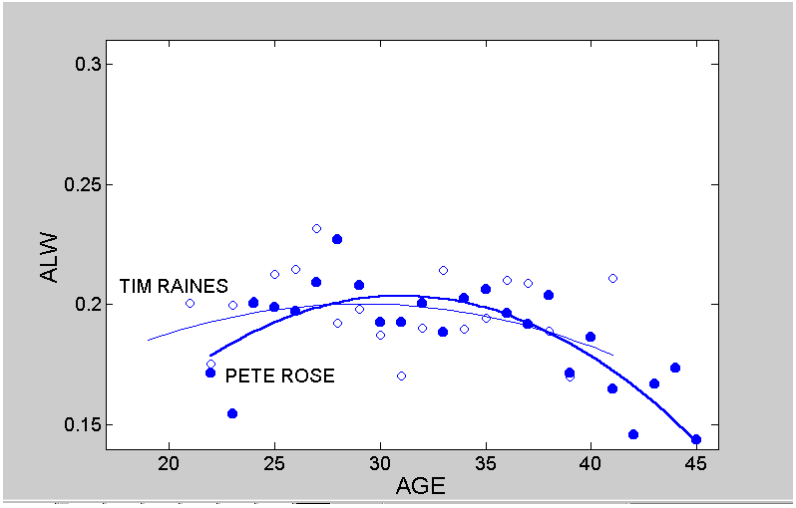

Figure 13 – Scatterplot of ALW and fitted trajectories for Tim Raines and Pete Rose

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 progression 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.

10. Notes

Details on the Bayesian model used, as well as the raw data used as the basis for the graphs in this paper, may be obtained by writing to the author at the address below.

Jim Albert, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, 43403,
albert@bgnet.bgsu.edu ♦

Erratum

In an article last issue, "Week-to-Week Consistency in Individual Offensive Performance," some printings of BTN mis-identified one of the authors of "Curve Ball". The correct authors are Jim Albert and Jay Bennett.

Apologies for the error.

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work (but no death threats, please) are all welcome.

Articles should be submitted in electronic form, either by e-mail or on PC-readable floppy disk. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

I usually edit for spelling and grammar. (But if you want to make my life a bit easier: please, use two spaces after the period in a sentence. Everything else is pretty easy to fix.)

If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does, and please include your byline at the end with your address (see the end of any article this issue).

Deadlines: January 24, April 24, July 24, and October 24, for issues of February, May, August, and November, respectively.

I will acknowledge all articles within three days of receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to:
Phil Birnbaum
18 Deerfield Dr. #608, Nepean, Ontario, Canada, K2G 4L1
birnbaum@sympatico.ca

Receive BTN by E-mail

You can help save SABR some money, and me some time, by receiving your copy of *By the Numbers* by e-mail. BTN is sent in Microsoft Word 97 format; if you don't have Word 97, a free viewer is available at the Microsoft web site (http://support.microsoft.com/support/kb/articles/Q165/9/08.ASP).

To get on the electronic subscription list, send me (Phil Birnbaum) an e-mail at birnbaum@sympatico.ca. If you're not sure if you can read Word 97 format, just let me know and I'll send you this issue so you can try

If you don't have e-mail, don't worry – you will always be entitled to receive BTN by mail, as usual.