1-29-2023

Filling in the Gaps: A Multiple Imputation Approach to Estimating Aging Curves in Baseball

Quang Nguyen
Carnegie Mellon University, nmquang@cmu.edu

Gregory J. Matthews
Loyola University Chicago, gmatthews1@luc.edu

Filling in the Gaps: A Multiple Imputation Approach to Estimating Aging Curves in Baseball

Quang Nguyen
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
nmquang@cmu.edu

Gregory J. Matthews
Department of Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660
gmatthews1@luc.edu

Abstract

In sports, an aging curve depicts the relationship between average performance and age in athletes' careers. This paper investigates the aging curves for offensive players in the Major League Baseball. We study this problem in a missing data context and account for different types of dropouts of baseball players during their careers. In particular, the performance metric associated with the missing seasons is imputed using a multiple imputation model for multilevel data, and the aging curves are constructed based on the imputed datasets. We first perform a simulation study to evaluate the effects of different dropout mechanisms on the estimation of aging curves. Our method is then illustrated with analyses of MLB player data from past seasons. Results suggest an overestimation of the aging curves constructed without imputing the unobserved seasons, whereas a better estimate is achieved with our approach.

Keywords: aging curve, baseball, multiple imputation, survival bias

1 Introduction

The rise and fall of an athlete is a popular topic of discussion in the sports media today. Questions regarding whether a player has reached their peak, is past their prime, or is good enough to remain in their respective professional league are often seen in different media outlets such as news articles, television debate shows, and podcasts. The average performance of players by age throughout their careers is visually represented by an *aging curve*. This graph typically consists of a horizontal axis representing a time variable (usually age or season) and a vertical axis showing a performance metric at each time point in a player's career.

One significant challenge associated with the study of aging curves in sports is *survival bias*, as pointed out by Lichtman (2009), Turtoro (2019), Judge (2020a), and Schuckers, Lopez, and Macdonald (2023). In particular, the aging effects are not often determined from a full population of athletes (i.e. all players who have ever played) in a given league. That is, only players that are good enough to remain are observed; whereas those who might be involved, but do not actually participate or are not talented enough to compete, are being completely disregarded. This very likely results in an overestimation of the aging curves.

As such, player survivorship and dropout can be viewed as a missing data problem. There are several distinct cases of player absence from professional sport at different points in their careers. Early on, teams may elect to assign their young prospects to their minor/development league affiliates for several years of nurture. Many of those players would end up receiving a call-up to join the senior squad, when the team believes they are ready. During the middle of a player's career, a nonappearance could occur due to various reasons. Injury is unavoidable in sports, and this could cost a player at least one year of their playing time. Personal reasons such as contract situation and more recently, concerns regarding a global pandemic, could also lead to athletes sitting out a season. Later on, a player, by either their choice or their team's choice, might head for retirement because they cannot perform at a level like they used to, leading to unobserved seasons that could have been played.

The primary aim of this paper is to apply missing data techniques to the estimation of aging curves. In doing so, we focus on baseball and pose the following research question: *What would the aging curve look like if all players competed in every season within a fixed range of age?* In other words, *what would have happened if a player who was forced to retire from their league at a certain age had played a full career?* The manuscript continues with a review of existing literature on aging curves in baseball and other sports in Section 2. Next, we describe our data and methods used to perform our analyses in Section 3. After that, our approach is implemented through simulation and analyses of real baseball data in Sections 4 and 5. Finally, in Section 6, we conclude with a discussion of the results, limitations, and directions for future work.

2 Literature Review

To date, we find a considerable amount of previous work related to aging curves and career trajectory of athletes. This body of work consists of several key themes, a wide array of statistical methods, and applications in many sports besides baseball such as basketball, hockey and track and field, to name a few.

A typical notion in the baseball aging curves literature is the assumption of a quadratic form for modeling the relationship between performance and age. Morris (1983) looks at Ty Cobb's batting average trajectory using parametric empirical Bayes and uses shrinkage methods to obtain a parabolic curve for Cobb's career performance. Albert (1992) proposes a quadratic random effects log-linear model for smoothing a batter's home run rates throughout their career. A nonparametric method is implemented to estimate the age effect on performance in baseball, hockey, and golf by Berry, Reese, and Larkey (1999). However, Albert (1999) weighs in on this nonparametric approach and questions the assumptions that the peak age and periods of growth and decline are the same for all players. Albert (1999) ultimately prefers a second-degree polynomial function for estimating age effect in baseball, which is a parametric model. Continuing his series of work on aging trajectories, Albert (2002) proposes a Bayesian exchangeable model for baseball hitting performance. This approach combines quadratic regression estimates and assumes similar careers for players born in the same decade. Fair (2008) and Bradbury (2009) both use a fixed-effects regression to examine age effects in the MLB, also assuming a quadratic aging curve form. A major drawback of Bradbury (2009)'s study is that the analysis only considers players with longer baseball careers.

In addition to baseball, studies on aging curves have also been conducted for other sports. Early on, Moore (1975) looks at the association between age and running speed in track and field and produces aging curves for different running distances using an exponential model. Fair (1994) and Fair (2007) study the age effects in track and field, swimming, chess, and running, in addition to their latter work in baseball, as mentioned earlier. In triathlon, Villaroel, Mora, and Parra (2011) assume a quadratic relationship between performance and age, as many have previously considered. As for basketball, Page, Barney, and McGuire (2013) use a Gaussian process regression in a hierarchical Bayesian framework to model age effect in the NBA. Additionally, Lailvaux, Wilson, and Kasumovic (2014) use NBA and WNBA data to investigate and test for potential sex differences in the aging of comparable performance indicators. Vaci et al. (2019) apply Bayesian cognitive latent variable modeling to explore aging and career performance in the NBA, accounting for player position and activity level. In tennis, Kovalchik (2014) studies age and performance trends in men's tennis using change point analysis.

Another convention in the aging curve modeling literature is the assumption of discrete observations. Specifically, most researchers use regression modeling and consider a data

measurement for each season played throughout a player's career. In contrast to previous approaches, Wakim and Jin (2014) take a different route and consider functional data analysis as the primary tool for modeling MLB and NBA aging curves. This is a continuous framework which treats the entire career performance of an athlete as a smooth function. In a similar functional data setting, Leroy et al. (2018) investigate the performance progression curves in swimming.

A subset of the literature on aging and performance in sports specifically studies the question: At what age do athletes peak? Schulz and Curnow (1988) look at the age of peak performance for track and field, swimming, baseball, tennis, and golf. A follow-up study to this work was done by Schulz et al. (1994), where the authors focus on baseball and find that the average peak age for baseball players is between 27 and 30, considering several performance measures. Later findings on baseball peak age also show consistency with the results in Schulz et al. (1994). Fair (2008) determines the peak-performance age in baseball to be 28, whereas Bradbury (2009) determines that baseball hitters and pitchers reach the top of their careers at about 29 years old. In soccer, Dendir (2016) determines that the peak age for footballers in the top leagues falls within the range of 25 to 27.

The idea of player survivorship is only mentioned in a small number of articles. To our knowledge, not many researchers have incorporated missing data methods into the estimation of aging curves to account for missing but observable athletes. Schulz et al. (1994) and Schell (2005) note the selection bias problem with estimating performance averages by age in baseball, as better players tend to have longer career longevity. Schall and Smith (2000) predict survival probabilities of baseball players using a logit model, and examine the link between first-year performance and career length. Lichtman (2009) studies different aging curves for different eras and groups of players after correcting for survival bias, and shows that survival bias results in an overestimation of the age effects. Alternatively, Judge (2020a) concludes that survival bias leads to an underestimation, not overestimation, of the aging curves. In analyzing NHL player aging, Brander, Egan, and Yeung (2014) apply their quadratic and cubic fixed-effects regression models to predict performance for unobserved players in the data.

Recently, researchers have noticed the benefits of accounting for missing data in modeling performance in sports. Stival et al. (2022) use a latent class matrix-variate state-space framework to analyze runners' careers, and find that missing data patterns greatly contribute to the prediction of performance. Perhaps the most closely related approach to our work is that by Schuckers, Lopez, and Macdonald (2023), which considers different regression and imputation frameworks for estimating the aging curves in the National Hockey League (NHL). First, they investigate different regression approaches including spline, quadratic, quantile, and a delta plus method, which is an extension to the delta method previously studied by

Lichtman (2009), Turtoro (2019), and Judge (2020b). This paper also proposes an imputation approach for aging curve estimation, and ultimately concludes that the estimation becomes stronger when accounting for unobserved data, which addresses a major shortcoming in the estimation of aging curves. Säfvenberg (2022) modifies Schuckers, Lopez, and Macdonald (2023)'s imputation algorithm to study aging trajectory in Swedish football.

However, it appears that the aging curves in the aforementioned papers are constructed without taking into account the variability as a result of imputing missing data. This could be improved by applying multiple imputation rather than considering only one imputed dataset. As pointed out by Gelman and Hill (2006) (Chapter 25), conducting only a single imputation essentially assumes that the filled-in values correctly estimate the true values of the missing observations. Yet, there is uncertainty associated with the missingness, and multiple imputation can incorporate the missing data uncertainty and provide estimates for the different sources of variability.

3 Methods

3.1 Data Collection

In the forthcoming analyses, we rely on one primary source of publicly available baseball data: the Lahman baseball database (Lahman 2021). Created and maintained by Sean Lahman, this database contains pitching, hitting, and fielding information for Major League Baseball players and teams dating back to 1871. The data are available in many different formats, and the `Lahman` package in `R` (Friendly et al. 2022; R Core Team 2022) is utilized for our investigation.

Due to our specific purpose of examining the aging curves for baseball offensive players, we consider the following datasets from the `Lahman` library: `Batting`, which provides season-by-season batting statistics for baseball players; and `People`, which contains the date of birth of each player, allowing us to calculate their age for each season played. In each table, an athlete is identified with their own `playerID`, hence we use this attribute as a joining key to merge the two tables together. A player's age for a season is determined as their age on June 30, and we apply the formula suggested by Marchi, Albert, and Baumer (2018) for age adjustment based on one's birth month.

Throughout this paper, we consider on-base plus slugging (OPS), which combines a hitter's ability to reach base and power-hitting, as the baseball offensive performance measure. We normalize the OPS for all players and then apply an arcsine transformation to ensure a reasonable range for the OPS values when conducting simulation and imputation. We also assume a fixed length for a player's career, ranging from age 21 to 39. In terms of sample restriction, we observe all player-seasons with at least 100 plate appearances, which means a

season is determined as missing if one's plate appearances is below that threshold.

3.2 Multiple Imputation

Multiple imputation (Rubin 1987) is a popular statistical procedure for addressing the presence of incomplete data. The goal of this approach is to replace the missing data with plausible values to create multiple completed datasets. These completed datasets can each then be analyzed and results are combined across the imputed versions. Multiple imputation consists of three steps. First, based on an appropriate imputation model, m completed copies of the dataset are created by filling in the missing values. Next, m analyses are performed on each of the m completed datasets. Finally, the results from each of the m datasets are pooled together to create a combined estimate, and standard errors are estimated to account for the between- and within-imputation variability. This last step can be accomplished using asymptotic theory with Rubin's rules (Little and Rubin 1987), which are as follows.

Let Q be a parameter of interest and $\widehat{Q}_i$ where $i = 1, 2, \ldots, m$ are estimates of Q obtained from m imputed datasets, with sampling variance U estimated by $\widehat{U}_i$. Then the point estimate for Q is the average of the m estimates

$$\overline{Q} = \frac{1}{m} \sum_{i=1}^{m} \widehat{Q}_i \,.$$

The variance for $\overline{Q}$ is defined as

$$T = \overline{U} + \left(1 + \frac{1}{m}\right) B \,,$$

where

$$\overline{U} = \frac{1}{m} \sum_{i=1}^{m} \widehat{U}_i$$

and

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\widehat{Q}_i - \overline{Q})^2$$

are the estimated within and between variances, respectively.

Inferences for Q are based on the approximation

$$\frac{Q - \overline{Q}}{\sqrt{T}} \sim t_\nu \,,$$

where t_ν is the Student's t-distribution with $\nu = (m-1)\left(1 + \frac{1}{r}\right)^2$ degrees of freedom, with $r = \left(1 + \frac{1}{m}\right) \frac{B}{\overline{U}}$ representing the relative increase in variance due to missing data.

Accordingly, a $100(1-\alpha)\%$ Wald confidence interval for Q is computed as

$$\overline{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T},$$

where $t_{\nu,1-\alpha/2}$ is the $1-\alpha/2$ quantile of t_ν.

It is important to understand the reasons behind the missingness when applying multiple imputation to handle incomplete data. Generally, there are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin 1976). MCAR occurs when a missing observation is statistically independent of both the observed and unobserved data. In the case of MAR, the missingness is associated only with the observed and not with the unobserved data. When data are MNAR, the probability of missingness is related to both observed and unobserved data.

Among the tools for performing multiple imputation, multivariate imputations by chained equation (MICE) (van Buuren and Groothuis-Oudshoorn 1999) is a flexible, robust, and widely used method. This algorithm imputes missing data via an iterative series of conditional models. In each iteration, each incomplete variable is filled in by a separate model of all the remaining variables. The iterations continue until apparent convergence is reached.

Here we implement the MICE framework in `R` via the popular `mice` package (van Buuren and Groothuis-Oudshoorn 2011). Moreover, we focus on multilevel multiple imputation, due to the hierarchical structure of our data. Specifically, we consider multiple imputation by a two-level normal linear mixed model with heterogeneous within-group variance (Kasim and Raudenbush 1998). In context, our data consist of baseball seasons (ages) which are nested within the class variable, player; and the season-by-season performance is considered to be correlated for each athlete. The described imputation model can be specified as the `2l.norm` method available in the `mice` library.

4 Simulation

In this simulation, we demonstrate our aging curve estimation approach with multiple imputation, and evaluate how different types of player dropouts affect the curve. There are three steps to our simulation. First, we fit a model for the performance-age relationship and utilize its output to generate fictional careers for baseball players. Next, we generate missing data by dropping players from the full dataset based on different criteria, and examine how the missingness affects the original aging curve obtained from fully observed data. Finally, we apply multiple imputation to obtain completed datasets and assess how close the imputed aging curves are to the true curve based on fully observed data.

4.1 Generating Player Careers

We fit a mixed-effects model using the player data described in Section 3.1. Our goal is to obtain the variance components of the fitted model to simulate baseball player careers. The model of interest is of the following form:

$$\begin{gathered} Y_{pq} = (\beta_0 + b_{0p}) + \beta_1 X_q + \beta_2 X_q^2 + \beta_3 X_q^3 + \epsilon_{pq} \\ b_{0p} \sim N(0, \tau^2) \\ \epsilon_{pq} \sim N(0, \sigma^2). \end{gathered}$$

In detail, this model relates the performance metric Y_{pq} (in our case, transformed OPS) for player p at age (season) q to a baseline level via the fixed effect β_0. The only covariate X in the model is age, which is assumed to have a cubic relationship with the response variable, transformed OPS. Another component is the observational-level error ϵ_{pq} with variance σ^2 for player p at age q. We also introduce the random effects b_{0p}, which represents the deviation from the grand mean β_0 for player p, allowing a fixed amount of shift to the performance prediction for each player. In addition, to incorporate the variability in production across the season q, a random effect parameter τ^2 is included. Our modeling approach is implemented using the `lme4` package in `R` (Bates et al. 2015). We utilize the estimated variance components from the fitted model to simulate 1000 careers for baseball players from the ages of 21 to 39.

4.2 Generating Missing Data

After obtaining reasonable simulated careers for baseball players, we create different types of dropouts and examine how they lead to deviations from the fully observed aging curve. We consider the following cases of player dropout from the league:

(1) Dropout players with 4-year OPS average below a certain threshold, say 0.55.

(2) Dropout players with OPS average from age 21 (start of career) to 25 of less than 0.55.

(3) 25% of the players randomly retire at age 30.

For the first two scenarios, the missingness mechanism is MAR, since players get removed due to low previously observed performance. Dropout case (3) falls under MCAR, since athletes are selected completely at random to retire without considering past or future offensive production.

Figure 1 displays the average OPS aging curve for all baseball players obtained from the original data with no missingness, along with the aging curves constructed based on data with only the surviving players associated with the dropout mechanisms mentioned above. These are smoothed curves obtained from loess model fits, and we use mean absolute error

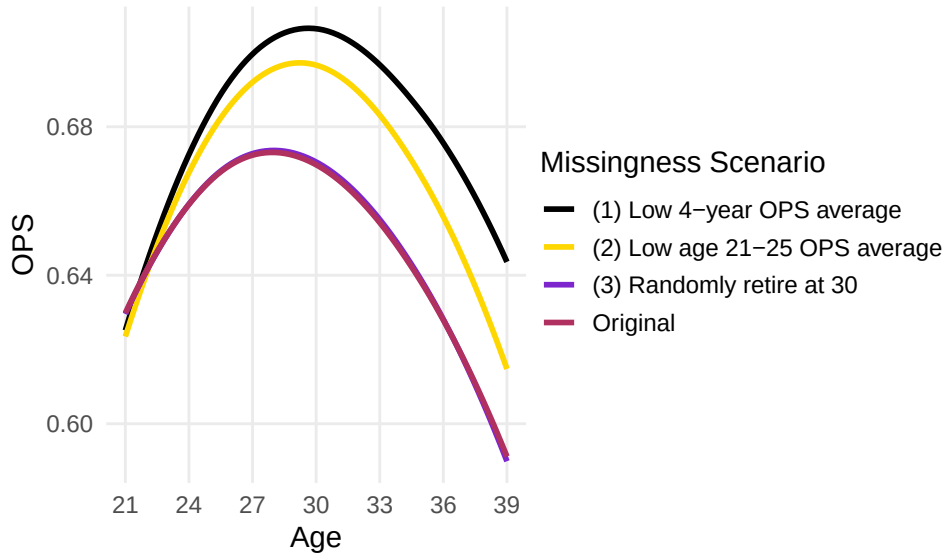

Figure 1: Comparison of the average aging curves constructed with the full simulated data (maroon) and different cases of dropouts (without imputation). The dropout mechanisms presented are (purple) randomly removing 25% of the players at age 30; (black) dropping players with any 4-year OPS average below 0.55; and (gold) dropping players with OPS average between age 20 and 25 of less than 0.55

(MAE) to evaluate the discrepancy between the dropout and true aging curves. It is clear that randomly removing players have minimal effect on the aging curve, as the curve obtained from (3) and the original curve essentially overlap (MAE $= 7.42 \times 10^{-4}$). On the other hand, a positive shift from the fully observed curve occurs for the remaining two cases of dropout based on OPS average (MAE $= 0.031$ for (1) and MAE $= 0.019$ for (2)). This means the aging curves with only the surviving players are overestimated in the presence of missing data due to past performance. More specifically, the estimated player performance drops off faster as they age when considering missing data than when it is estimated with only complete case analysis (i.e. only considering observed seasons). In addition to overestimating the aging curve, ignoring player dropout also pushes the estimated performance peak to a later point (29 years old) in a player's career.

4.3 Imputation

Next, we implement the multiple imputation with a hierarchical structure procedure described in Section 3 to the cases of dropout that shifts the aging effect on performance. We perform $m = 5$ imputations with each made up of 30 iterations of the MICE algorithm, and apply Rubin's rules for combining the imputation estimates. The following results are illustrated for dropout mechanism (2), where players with a low OPS average at the start of their careers (ages 21–25) are forced out of the league.

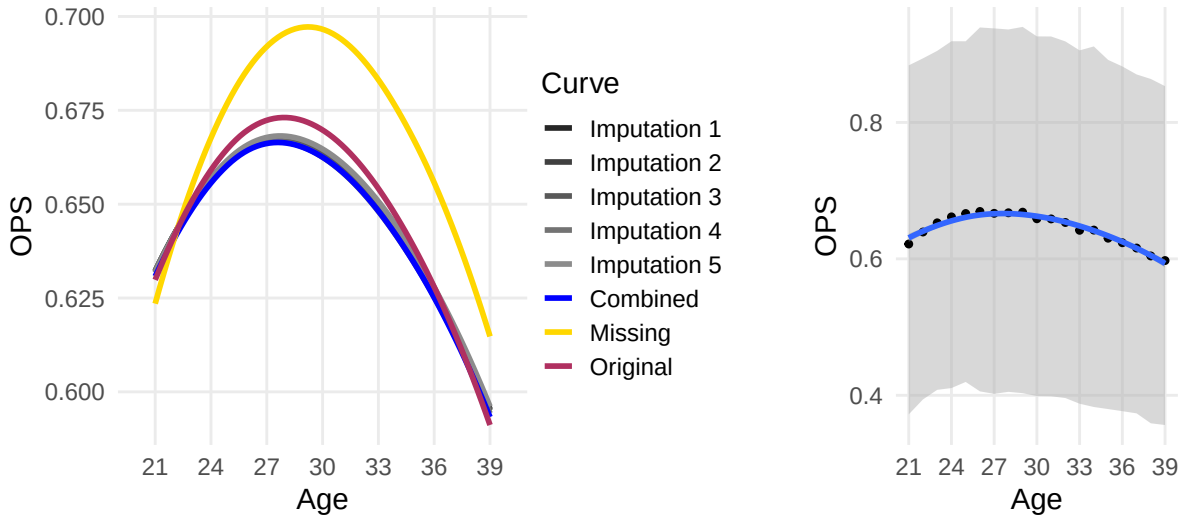

Figure 2: At left, comparison of the average OPS aging curves constructed with the fully observed data, only surviving players, and imputation. At right, combined imputed curve with 95% confidence intervals obtained from Rubin's rules. Results shown here are for dropout case of players having OPS average from age 21 to 25 below 0.55.

Figure 2 (left) shows smoothed fitting loess aging curves for all 5 imputations and a combined version of them, in addition to the curves constructed with fully observed and only surviving players data. The 95% confidence interval for the mean OPS at each age point in the combined curve obtained from Rubin's rules is further illustrated in Figure 2 (right). It appears that the combined imputed curve follows the same shape as the true, known curve. Moreover, imputation seems to capture the rate of change for the beginning and end career periods quite well, whereas the middle of career looks to be slightly underestimated. The resulting MAE of 0.0039 confirms that there is little deviation of the combined curve from the true one.

Additionally, we perform diagnostics to assess the plausibility of the imputations, and also examine whether the MICE algorithm converges. We first check for distributional discrepancy by comparing the distributions of the fully observed and imputed data. Figure 3 (left) presents the density curves of the OPS values for each imputed dataset and the fully simulated data. It is obvious that the imputation distributions are well-matched with the observed data. To confirm convergence of the MICE algorithm, we inspect trace plots for the mean and standard deviation of the imputed OPS values. As shown in Figure 3 (right), convergence is apparent, since no definite trend is revealed and the imputation chains are intermingled with one another.

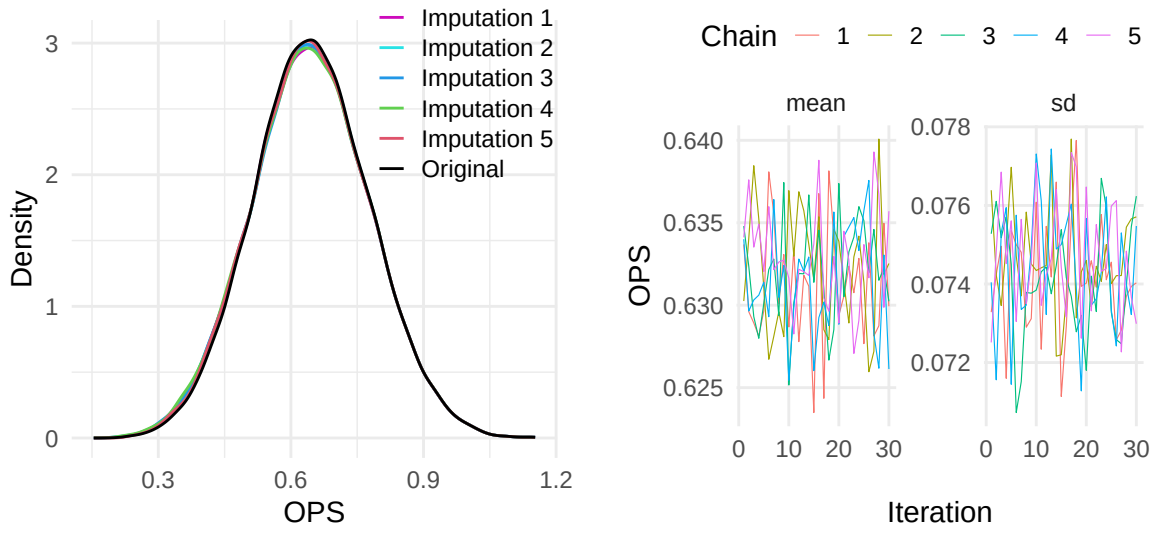

Figure 3: At left, kernel density estimates for the fully observed and imputed OPS values. At right, trace plots for the mean and standard deviation of the imputed OPS values against the iteration number for the imputed data. Results shown here are for dropout case of players having OPS average from age 21 to 25 below 0.55.

5 Application: MLB Data

Lastly, we apply the previously mentioned multilevel multiple imputation model to estimate the average OPS aging curve for MLB players. For this investigation, besides the data pre-processing tasks mentioned in Section 3.1, our sample is limited to all players who made their major league debut no sooner than 1985, resulting in a total of 2323 players. To perform imputation, we pass in the same parameters to our simulation study ($m = 5$ with 30 iterations for each imputation).

Figure 4 shows the OPS aging curves for MLB players estimated with and without imputation. The plot illustrates a similar result as our simulations as the combined curve based on imputations is lower than the curve obtained when ignoring the missing data. The peak age after imputation occurs much earlier (26 years old) compared to the peak age when player dropout is not accounted for (30 years old). It is clear that the aging effect is overestimated without considering the unobserved player-seasons. In other words, the actual performance declines with age more rapidly than estimates based on only the observed data.

6 Discussion

The concept of survival bias is frequently seen in professional sports, and our paper approaches the topic of aging curves and player dropout in baseball as a missing data problem. We utilize multiple imputation with a multilevel structure to improve estimates for the baseball

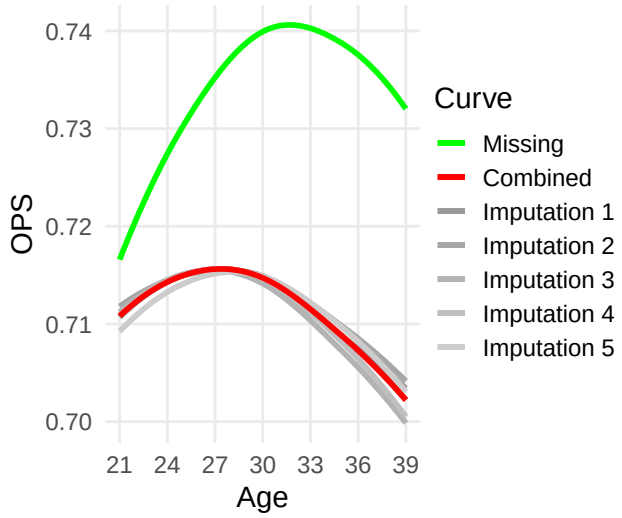

Figure 4: Comparison of the average OPS aging curves constructed with only observed players and imputation for MLB data.

aging curves. Through simulation, we highlight that ignoring the missing seasons leads to an overestimation of the age effect on baseball offensive performance. With imputation, we achieve an aging curve showing that players actually decline faster as they get older than previously estimated.

There are many limitations in our study which leave room for improvement in future work. In our current imputation model, age is the only predictor for estimating performance. It is possible to include more covariates in the imputation algorithm and determine whether a better aging curve estimate is achieved. In particular, we can factor in other baseball offensive statistics (e.g. home run rate, strikeout rate, WOBA, walk rate,...) in building an imputation model for OPS. We can also examine other performance metrics to see how age affects different statistics.

Furthermore, the aging curve estimation problem can be investigated in a completely different statistical setting. As noted in Section 2, rather than considering discrete observations, another way of studying aging curves is through a continuous approach, assuming a smooth curve for career performance. As pointed out by Wakim and Jin (2014), methods such as functional data analysis (FDA) and principal components analysis through conditional expectation (PACE) possess many modeling advantages, in regard to flexibility and robustness. There exists a number of proposed multiple imputation algorithms for functional data (He, Yucel, and Raghunathan 2011; Ciarleglio, Petkova, and Harel 2021; Rao and Reimherr 2021), which all can be applied in future studies on aging curves in sports.

Acknowledgements

We thank the organizers of the 2022 Carnegie Mellon Sports Analytics Conference (CMSAC) for the opportunity to present this work and receive feedback. We thank the anonymous reviewers of the Reproducible Research Competition at CMSAC 2022 for the insightful comments and suggestions. We thank Kathryne Piazza for her help in the early stages of this project.

References

Albert, Jim. 1992. "A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters." *The American Statistician* 46 (4): 246. https://doi.org/10.2307/2685306.

———. 1999. "Bridging Different Eras in Sports: Comment." *Journal of the American Statistical Association* 94 (447): 677. https://doi.org/10.2307/2669974.

———. 2002. "Smoothing Career Trajectories of Baseball Hitters." Unpublished manuscript, Bowling Green State University. https://bayesball.github.io/papers/career_trajectory.pdf.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Berry, Scott M., C. Shane Reese, and Patrick D. Larkey. 1999. "Bridging Different Eras in Sports." *Journal of the American Statistical Association* 94 (447): 661–76. https://doi.org/10.1080/01621459.1999.10474163.

Bradbury, John Charles. 2009. "Peak Athletic Performance and Ageing: Evidence from Baseball." *Journal of Sports Sciences* 27 (6): 599–610. https://doi.org/10.1080/02640410802691348.

Brander, James A., Edward J. Egan, and Louisa Yeung. 2014. "Estimating the Effects of Age on NHL Player Performance." *Journal of Quantitative Analysis in Sports* 10 (2). https://doi.org/10.1515/jqas-2013-0085.

Ciarleglio, Adam, Eva Petkova, and Ofer Harel. 2021. "Elucidating Age and Sex-Dependent Association Between Frontal EEG Asymmetry and Depression: An Application of Multiple Imputation in Functional Regression." *Journal of the American Statistical Association* 117 (537): 12–26. https://doi.org/10.1080/01621459.2021.1942011.

Dendir, Seife. 2016. "When Do Soccer Players Peak? A Note." *Journal of Sports Analytics* 2 (2): 89–105. https://doi.org/10.3233/jsa-160021.

Fair, Ray C. 1994. "How Fast Do Old Men Slow Down?" *The Review of Economics and Statistics* 76 (1): 103–18. https://ideas.repec.org/a/tpr/restat/v76y1994i1p103-18.html.

———. 2007. "Estimated Age Effects in Athletic Events and Chess." *Experimental Aging Research* 33 (1): 37–57. https://doi.org/10.1080/03610730601006305.

———. 2008. "Estimated Age Effects in Baseball." *Journal of Quantitative Analysis in Sports* 4 (1). https://doi.org/10.2202/1559-0410.1074.

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2022. *Lahman: Sean 'Lahman' Baseball Database*. https://CRAN.R-project.org/package=Lahman.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/cbo9780511790942.

He, Yulei, Recai Yucel, and Trivellore E. Raghunathan. 2011. "A Functional Multiple Imputation Approach to Incomplete Longitudinal Data." *Statistics in Medicine* 30 (10): 1137–56. https://doi.org/10.1002/sim.4201.

Judge, Jonathan. 2020a. "An Approach to Survivor Bias in Baseball." BaseballProspectus.com. https://www.baseballprospectus.com/news/article/59491/an-approach-to-survivor-bias-in-baseball.

———. 2020b. "The Delta Method, Revisited: Rethinking Aging Curves." BaseballProspectus.com. https://www.baseballprospectus.com/news/article/59972/the-delta-method-revisited.

Kasim, Rafa M., and Stephen W. Raudenbush. 1998. "Application of Gibbs Sampling to Nested Variance Components Models with Heterogeneous Within-Group Variance." *Journal of Educational and Behavioral Statistics* 23 (2): 93–116. https://doi.org/10.3102/10769986023002093.

Kovalchik, Stephanie Ann. 2014. "The Older They Rise the Younger They Fall: Age and Performance Trends in Men's Professional Tennis from 1991 to 2012." *Journal of Quantitative Analysis in Sports* 10 (2). https://doi.org/10.1515/jqas-2013-0091.

Lahman, Sean. 2021. "Lahman's Baseball Database." SeanLahman.com. https://www.seanlahman.com/baseball-archive/statistics/.

Lailvaux, Simon P., Robbie Wilson, and Michael M. Kasumovic. 2014. "Trait Compensation and Sex-Specific Aging of Performance in Male and Female Professional Basketball Players." *Evolution* 68 (5): 1523–32. https://doi.org/10.1111/evo.12375.

Leroy, Arthur, Andy Marc, Olivier Dupas, Jean Lionel Rey, and Servane Gey. 2018. "Functional Data Analysis in Sport Science: Example of Swimmers' Progression Curves Clustering." *Applied Sciences* 8 (10): 1766. https://doi.org/10.3390/app8101766.

Lichtman, Mitchel. 2009. "How Do Baseball Players Age? (Part 2)." The Hardball Times. https://tht.fangraphs.com/how-do-baseball-players-age-part-2.

Little, Roderick J. A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley.

Marchi, Max, Jim Albert, and Benjamin S. Baumer. 2018. *Analyzing Baseball Data with R*. 2nd ed. Boca Raton, FL: Chapman; Hall/CRC Press. https://doi.org/10.1201/9781351107099.

Moore, Dan H. 1975. "A Study of Age Group Track and Field Records to Relate Age and

Running Speed." *Nature* 253 (5489): 264–65. https://doi.org/10.1038/253264a0.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55. https://doi.org/10.1080/01621459.1983.10477920.

Page, Garritt L., Bradley J. Barney, and Aaron T. McGuire. 2013. "Effect of Position, Usage Rate, and Per Game Minutes Played on NBA Player Production Curves." *Journal of Quantitative Analysis in Sports* 9 (4): 337–45. https://doi.org/10.1515/jqas-2012-0023.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rao, Aniruddha Rajendra, and Matthew Reimherr. 2021. "Modern Multiple Imputation with Functional Data." *Stat* 10 (1). https://doi.org/10.1002/sta4.331.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92. https://doi.org/10.1093/biomet/63.3.581.

———, ed. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696.

Säfvenberg, Rasmus. 2022. "Age of Peak Performance Among Swedish Football Players." Master's thesis, Linköping University, The Division of Statistics; Machine Learning. https://www.diva-portal.org/smash/get/diva2:1674221/FULLTEXT01.pdf.

Schall, Teddy, and Gary Smith. 2000. "Career Trajectories in Baseball." *CHANCE* 13 (4): 35–38. https://doi.org/10.1080/09332480.2000.10542233.

Schell, Michael J. 2005. "Calling It a Career: Examining Player Aging." In *Baseball's All-Time Best Sluggers: Adjusted Batting Performance from Strikeouts to Home Runs*, 45–57. Princeton University Press. https://www.jstor.org/stable/j.ctt19705ks.

Schuckers, Michael, Michael Lopez, and Brian Macdonald. 2023. "Estimation of Player Aging Curves Using Regression and Imputation." *Annals of Operations Research*, January. https://doi.org/10.1007/s10479-022-05127-y.

Schulz, Richard, and Christine Curnow. 1988. "Peak Performance and Age Among Superathletes: Track and Field, Swimming, Baseball, Tennis, and Golf." *Journal of Gerontology* 43 (5): 113–20. https://doi.org/10.1093/geronj/43.5.p113.

Schulz, Richard, Donald Musa, James Staszewski, and Robert S. Siegler. 1994. "The Relationship Between Age and Major League Baseball Performance: Implications for Development." *Psychology and Aging* 9 (2): 274–86. https://doi.org/10.1037/0882-7974.9.2.274.

Stival, Mattia, Mauro Bernardi, Manuela Cattelan, and Petros Dellaportas. 2022. "Missing Data Patterns in Runners' Careers: Do They Matter?" arXiv. https://doi.org/10.48550/arxiv.2206.12716.

Turtoro, CJ. 2019. "Flexible Aging in the NHL Using GAM." RPubs. https://rpubs.com/cjtdevil/nhl_aging.

Vaci, Nemanja, Dijana Cocić, Bartosz Gula, and Merim Bilalić. 2019. "Large Data and

Bayesian Modeling—aging Curves of NBA Players." *Behavior Research Methods* 51 (4): 1544–64. https://doi.org/10.3758/s13428-018-1183-8.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 1999. *Flexible Multivariate Imputation by MICE*. Vol. PG/VGZ/99.054. Leiden: TNO Prevention; Health.

———. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Villaroel, Carlos, Rigoberto Mora, and Gilberto C. González Parra. 2011. "Elite Triathlete Performance Related to Age." *Journal of Human Sport and Exercise* 6 (2 (Suppl.)): 363–73. https://doi.org/10.4100/jhse.2011.62.16.

Wakim, Alexander, and Jimmy Jin. 2014. "Functional Data Analysis of Aging Curves in Sports." arXiv. https://doi.org/10.48550/arXiv.1403.7548.