Estimating Aging Curves: Using Multiple Imputation to Examine Career Trajectories of MLB Offensive Players

2022 CMSAC Reproducible Research Competition

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 metrics associated with the missing seasons are 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 · statistics in sports

1 Introduction

Athlete's career trajectory is a popular topic of discussion in the media nowadays. 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 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 et al. (2021). In particular, the aging effects are not often determined from a full population of athletes in a given league. That is, only players that are good enough to remain are observed; whereas those might be involved, but do not actually participate or not talented enough to compete, are being completely disregarded. This very likely results in an overestimation in the construction of aging curves.

As such, player survivorship and dropout can be viewed in the context of missing data. There are different cases of player absence from professional sport at different points in their careers. At the beginning, 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 one'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 might head for retirement because they cannot perform at a level like they used to.

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 one research question: What would the aging curve look like if players competed in every season from 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.

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 et al. (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 modeling 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.

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 et al. (2011) assume a quadratic relationship between performance and age, as many have previously considered. As for basketball, Page et al. (2013) use a Gaussian process regression in a hierarchical Bayesian framework to model age effect in the NBA. Additionally, Lailvaux et al. (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 levels. 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 & 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.

A subset of the literature on aging and performance in sports provides answers to the question: At what age do athletes peak? Schulz & 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 their career apex 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.

Last and most importantly, 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 & 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 survivorship bias leads to an underestimation, not overestimation, of the aging curves. In analyzing NHL player aging, Brander et al. (2014) apply their quadratic and cubic fixed-effects regression models to predict performance for unobserved players in the data.

Perhaps the most closely related approach to our work is that by Schuckers et al. (2021), 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. However, it appears that the aging curves 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 & 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, 1996 – 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., 2021; R Core Team, 2022) is utilized for our investigation.

Due to our specific purpose of examining the aging curves for baseball offensive players, the following datasets from the `Lahman` library are considered: `Batting`, which provides season-by-season batting statistics for baseball players; and `People`, to obtain the date of birth of each player and calculate their age for each season played. In each table, an athlete is identified with their own `playerID`, hence this attribute is used 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 the formula suggested by Marchi et al. (2018) for age adjustment based on one's birth month is applied.

Throughout this paper, we consider on-base plus slugging (OPS), which combines a hitter's ability to reach base on a swing and power-hitting, as the baseball offensive performance measure. We scale 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 datasets can each be analyzed and results are combined across the imputed versions. Multiple imputation consists of three steps. First, based on an appropriate imputation model, m copies of the dataset are created by filling 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 that account for the between and within imputation

variability. This last step can be accomplished using asymptotic theory with Rubin's combining rules (Little & 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. Based on the degree of bias attributable to the missingness, three types of missing data are defined: 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 with the observed but not with the unobserved data. When data are MNAR, there is a link between the missingness and the unobserved values in the dataset.

Among the tools for performing multiple imputation, multivariate imputations by chained equation (MICE) (van Buuren & 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.

In this paper, we implement the MICE framework in `R` via the popular `mice` package (van Buuren & 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 & 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.

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 consideration is of the form

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

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 sources of variance from the fitted model to simulate 1000 careers for baseball players of age 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 cause deviations from the fully observed aging curve. We consider the following cases of players' retirement 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 at random to retire without considering past or future offensive production.

Figure 1 displays the average OPS aging curves for all baseball players obtained from the original data with no missingness and 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 (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 with imputation than with only complete case analysis.

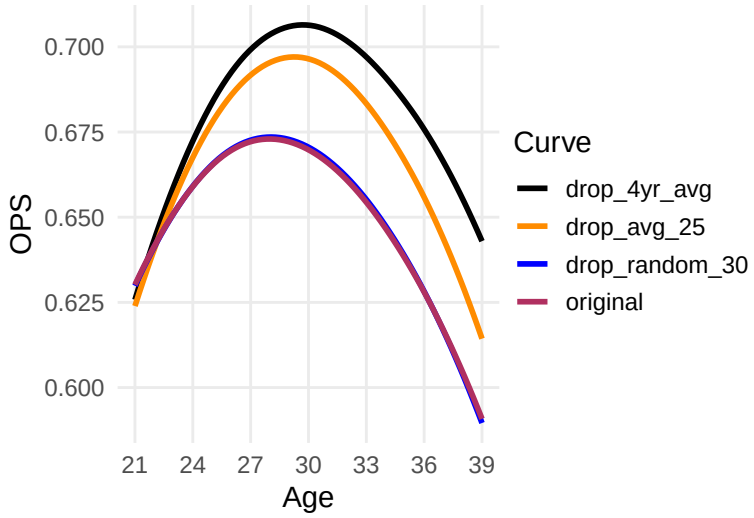

Figure 1: Comparison of the average OPS aging curve constructed with the fully observed data and different cases of dropouts (obtained only for the surviving players and without imputation).

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, 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 (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 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.

5 Application: MLB Data

Lastly, we apply the 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 parameters similar to our simulation study ($m = 5$ with 30 iterations for each imputation).

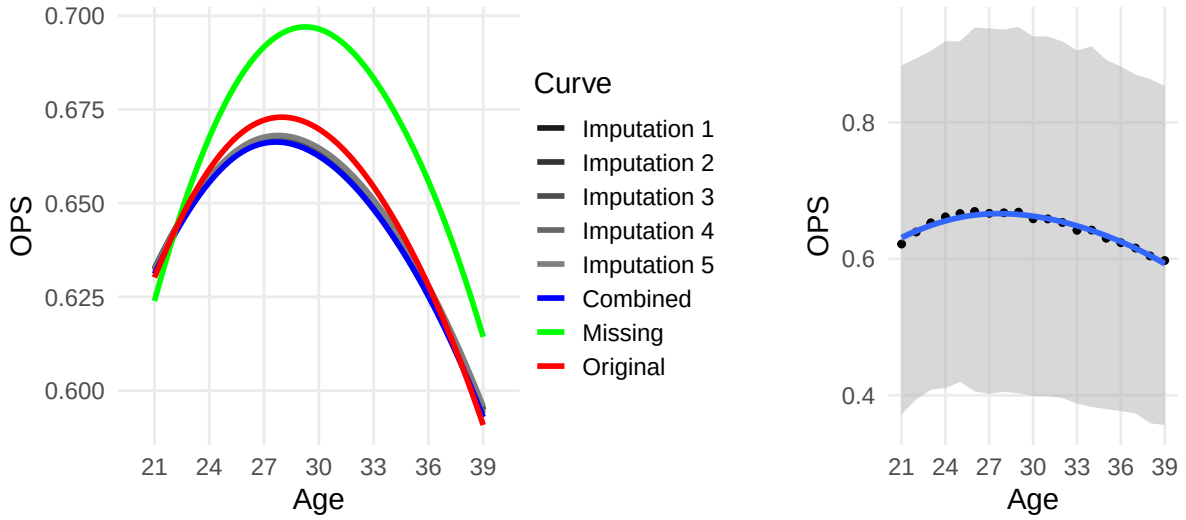

Figure 2: At left, comparison of the average OPS aging curve constructed with the fully observed data, only surviving players, and imputation. At right, combined imputed curve with 95% confidence intervals obtained from Rubin's rules. Results shown here are for dropout case of players having OPS average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

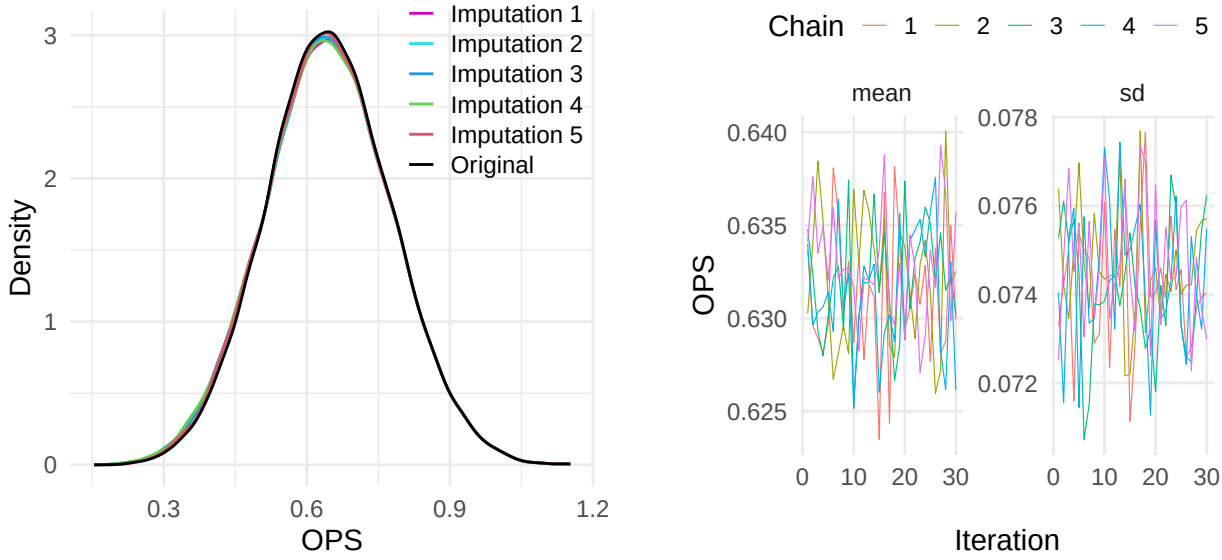

Figure 3: At left, kernel density estimates for the fully observed and imputed OPS values. At right, trace plots for the mean and standard deviation of the imputed OPS values against the iteration number for the imputed data. Results shown here are for dropout case of players having OPS average from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

Figure 4 shows the OPS aging curves for MLB players estimated with and without imputation. The plot illustrates a similar result as from simulation, as the combined imputed curve is lower than the curve obtained when ignoring the missing data. It is clear that the aging effect is overestimated without considering the unobserved player-seasons that are observable. In other words, the actual performance declines with age more rapidly than estimates based on only the observed data.

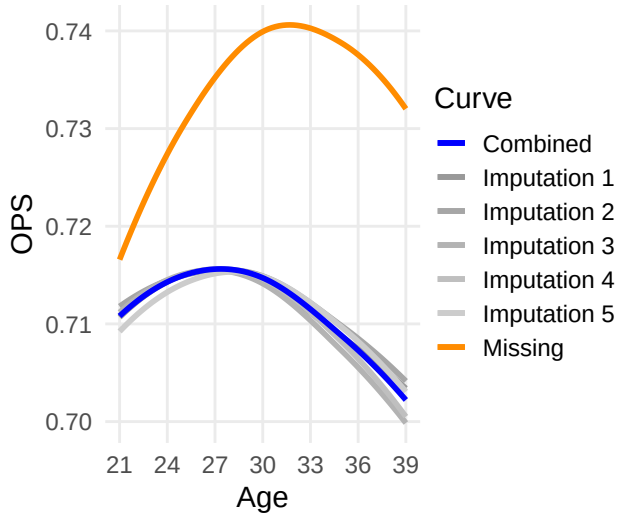

Figure 4: Comparison of the average OPS aging curve constructed with only observed players and imputation for MLB data.

6 Discussion

The concept of survivorship 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 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 a better aging curve which shows 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 algorithm and determine whether a better aging curve estimate is achieved. In particular, we could factor in other baseball offensive statistics (e.g. home run rate, strikeout rate, WOBA, walk rate,...) in building an imputation model for OPS.

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 & 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 (Ciarleglio et al., 2021; He et al., 2011; Rao & Reimherr, 2021), which all can be applied in future studies on aging curves in sports.

References

Albert, J. (2002). *Smoothing career trajectories of baseball hitters*. Unpublished manuscript, Bowling Green State University. https://bayesball.github.io/papers/career_trajectory.pdf

Albert, J. (1999). Bridging different eras in sports: comment. *Journal of the American Statistical Association*, *94*(447), 677. https://doi.org/10.2307/2669974

Albert, J. (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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (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, S. M., Reese, C. S., & Larkey, P. D. (1999). Bridging different eras in sports. *Journal of the American Statistical Association*, *94*(447), 661–676. https://doi.org/10.1080/01621459.1999.10474163

Bradbury, J. C. (2009). Peak athletic performance and ageing: Evidence from baseball. *Journal of Sports Sciences*, *27*(6), 599–610. https://doi.org/10.1080/02640410802691348

Brander, J. A., Egan, E. J., & Yeung, L. (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, A., Petkova, E., & Harel, O. (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, S. (2016). When do soccer players peak? A note. *Journal of Sports Analytics*, *2*(2), 89–105. https://doi.org/10.3233/jsa-160021

Fair, R. C. (2007). Estimated age effects in athletic events and chess. *Experimental Aging Research*, *33*(1), 37–57. https://doi.org/10.1080/03610730601006305

Fair, R. C. (2008). Estimated age effects in baseball. *Journal of Quantitative Analysis in Sports*, *4*(1). https://doi.org/10.2202/1559-0410.1074

Fair, R. C. (1994). How Fast Do Old Men Slow Down? *The Review of Economics and Statistics*, *76*(1), 103–118. https://ideas.repec.org/a/tpr/restat/v76y1994i1p103-18.html

Friendly, M., Dalzell, C., Monkman, M., & Murphy, D. (2021). *Lahman: Sean 'Lahman' Baseball Database*. https://CRAN.R-project.org/package=Lahman

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/cbo9780511790942

He, Y., Yucel, R., & Raghunathan, T. E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine*, *30*(10), 1137–1156. https://doi.org/10.1002/sim.4201

Judge, J. (2020a). *An approach to survivor bias in baseball*. BaseballProspectus.com. https://www.baseballprospectus.com/news/article/59491/an-approach-to-survivor-bias-in-baseball

Judge, J. (2020b). *The Delta Method, Revisited: Rethinking Aging Curves*. BaseballProspectus.com. https://www.baseballprospectus.com/news/article/59972/the-delta-method-revisited

Kasim, R. M., & Raudenbush, S. W. (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, S. A. (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, S. (1996 – 2021). *Lahman's baseball database*. SeanLahman.com. https://www.seanlahman.com/baseball-archive/statistics/

Lailvaux, S. P., Wilson, R., & Kasumovic, M. M. (2014). Trait compensation and sex-specific aging of performance in male and female professional basketball players. *Evolution*, *68*(5), 1523–1532. https://doi.org/10.1111/evo.12375

Lichtman, M. (2009). *How do baseball players age? (Part 2)*. The Hardball Times. https://tht.fangraphs.com/how-do-baseball-players-age-part-2

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley.

Marchi, M., Albert, J., & Baumer, B. S. (2018). *Analyzing baseball data with R* (2nd ed., p. 83). Boca Raton, FL: Chapman; Hall/CRC Press. https://doi.org/10.1201/9781351107099

Moore, D. H. (1975). A study of age group track and field records to relate age and running speed. *Nature*, *253*(5489), 264–265. https://doi.org/10.1038/253264a0

Morris, C. 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, G. L., Barney, B. J., & McGuire, A. T. (2013). Effect of position, usage rate, and per game minutes played on NBA player production curves. *Journal of Quantitative Analysis in Sports*, *0*(0), 1–9. https://doi.org/10.1515/jqas-2012-0023

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, A. R., & Reimherr, M. (2021). Modern multiple imputation with functional data. *Stat*, *10*(1). https://doi.org/10.1002/sta4.331

Rubin, D. B. (Ed.). (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schall, T., & Smith, G. (2000). Career trajectories in baseball. *CHANCE*, *13*(4), 35–38. https://doi.org/10.1080/09332480.2000.10542233

Schell, M. J. (2005). Calling it a career: Examining player aging. In *Baseball's all-time best sluggers: Adjusted batting performance from strikeouts to home runs* (pp. 45–57). Princeton University Press. https://www.jstor.org/stable/j.ctt19705ks

Schuckers, M., Lopez, M., & Macdonald, B. (2021). *What does not get observed can be used to make age curves stronger: Estimating player age curves using regression and imputation*. arXiv. https://doi.org/10.48550/arXiv.2110.14017

Schulz, R., & Curnow, C. (1988). Peak performance and age among superathletes: Track and field, swimming, baseball, tennis, and golf. *Journal of Gerontology*, *43*(5), 113–120. https://doi.org/10.1093/geronj/43.5.p113

Schulz, R., Musa, D., Staszewski, J., & Siegler, R. S. (1994). The relationship between age and major league baseball performance: Implications for development. *Psychology and Aging*, *9*(2), 274–286. https://doi.org/10.1037/0882-7974.9.2.274

Turtoro, C. (2019). *Flexible aging in the NHL using GAM*. RPubs. https://rpubs.com/cjtdevil/nhl_aging

Vaci, N., Cocić, D., Gula, B., & Bilalić, M. (2019). Large data and bayesian modeling—aging curves of NBA players. *Behavior Research Methods*, *51*(4), 1544–1564. https://doi.org/10.3758/s13428-018-1183-8

van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (1999). *Flexible multivariate imputation by MICE* (Vol. PG/VGZ/99.054). TNO Prevention; Health.

van Buuren, S., & Groothuis-Oudshoorn, K. (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, C., Mora, R., & Parra, G. C. G. (2011). Elite triathlete performance related to age. *Journal of Human Sport and Exercise*, *6*(2 (Suppl.)), 363–373. https://doi.org/10.4100/jhse.2011.62.16

Wakim, A., & Jin, J. (2014). *Functional data analysis of aging curves in sports*. arXiv. https://doi.org/10.48550/arXiv.1403.7548