Performance Persistence in Major League Baseball: Wharton Honors Thesis [*]

Chris Hua *Wharton School, University of Pennsylvania*
Linda Zhao *Wharton School, University of Pennsylvania*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi rhoncus est metus, porttitor scelerisque nisi tincidunt at. Fusce pretium mi nibh, pulvinar hendrerit turpis scelerisque nec. Etiam vitae auctor erat, eget molestie massa. Morbi magna dolor, tincidunt quis iaculis et, suscipit nec leo. Aenean et lectus lorem. Nullam suscipit eros et mi eleifend, id eleifend enim ullamcorper. Aenean molestie vulputate urna, non aliquet mi pellentesque eget.

Keywords: performance persistence, baseball, sports analytics

Introduction

Performance persistence is a well-studied trend in the financial literature, particularly involving mutual funds. In general, researchers aim to determine if there is a cross-period effect where fund returns can be predicted using past-period returns.

The performance of sports teams can be measured analogously to mutual fund returns. Competitive equality is a particularly vibrant field of study in sports.

Literature Review

We examine the literature here from two different perspectives: the performance persistence literature from finance, and the competitive equality literature in sports economics.

Data and Methodology

Calculations and writeups for this paper are done in the R language, using the `RMarkdown` package for typesetting and reproducibility in code (Xie 2014, Allaire et al. (2015)). Full code and writeup will be available on the author's Github.

Data

Major sports leagues have come to realize the importance of comprehensive, open datasets. Major League Baseball in particular has been on the forefront of the data revolution. At a high level, we do not require particularly involved data, though. The most important data that we require is number of games won at a per-team level, which is easily found from a variety of sources, and should be easily available for all major sports leagues.

In particular, we use the Sean Lahman database, and its `R` interface, the package `Lahman` (Friendly, n.d.). This database provides a comprehensive dataset of baseball statistics, and is easily queryable.

Because the methodology we define is fairly generic and extensible, we also identified other sources of data for use if we extend this analysis to other sports. In particular, Sports Reference

[*] Contact: chua@wharton.upenn.edu

provides a number of useful data sources, including Pro Football Reference for the National Football League (NFL) and Basketball Reference for the National Basketball Association (NBA).

Repeat performance methodology

There are several measures through which we measure repeat performance.

First, following (Brown and Goetzmann 1995) we use a nonparametric contingency table-based methodology to measure repeat performance. We define teams as "winners" or "losers" depending on if they win more games than the median number of games won per team for a given year. Then, we measure the behavior of teams in a 2 year period, that is, they are defined as "winner-winner" for 2014 if they are winners for 2014 and also winners in the 2015 season.

Then, we use the cross-product ratio to measure repeat performance.

$$R_{cp} = \frac{WW * LL}{WL * LW}$$

H_0^1: Performance in the first period is unrelated to performance in the second period. That is, $R_{cp} = 1$.
H_1^1: Performance in the first period is related to performance in the second period. That is, $R_{cp} > 1$.

We can approximate the standard error of the natural log of the odds ratio [TODO: Christensen 1990 p40] as the following:

$$\sigma_{\ln R_{cp}} = \sqrt{WW^{-1} + WL^{-1} + LW^{-1} + LL^{-1}}$$

In the above sequence, we consider a team a winner by its performance relative to the median winrate, which should be roughly 0.500, i.e. 50% winning rate. For the sake of comprehensiveness, we will also measure team performance relative to the 0.500 benchmark.

H_0^2: Performance in the first period is unrelated to performance in the second period. That is, $R_{cp} = 1$.
H_1^2: Performance in the first period is related to performance in the second period. That is, $R_{cp} > 1$.

We also consider a performance measure where teams are considered winners if they make the playoffs.

H_0^3: Making the playoffs in the first period is unrelated to making the playoffs in the second period. That is, $R_{cp} = 1$.
H_1^3: Making the playoffs in the first period is related to making the playoffs in the second period. That is, $R_{cp} > 1$.

Finally, to account for the peculiarities of American League vs National League, the wild-card process, or general nonsense, we will also measure winning rates relative to the "worst" team which does make the playoffs, where worst is defined as fewest wins.

H_0^4: Winning enough games to make the playoffs in the first period is unrelated to winning enough games to make the playoffs in the second period. That is, $R_{cp} = 1$.
H_1^4: Winning enough games to make the playoffs in the first period is related to winning enough games to make the playoffs in the second period. That is, $R_{cp} > 1$.

Data Analysis

Extensions

It would be interesting to measure performance persistence at the general manager or otherwise team management level. Brown and Goetzmann (1995) finds support for performance persistence in mutual funds at the fund manager level, which indicates the efficacy of common strategies that managers carry along. With the increased coverage in baseball about the importance of management who [Theo Epstein. . .]

Conclusion

Bibliography

Allaire, J, J Cheng, Yihui Xie, J McPherson, W Chang, Jeff Allen, H Wickham, and R Hyndman. 2015. “rmarkdown: Dynamic Documents for R.” *R Package Version 0.5*.

Brown, S, and William N. Goetzmann. 1995. “Performance Persistence.” *The Journal of Finance* 50 (2): 679. doi:10.2307/2329424.

Friendly, Michael. n.d. “Sean ’Lahman’ Baseball Database.” Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/Lahman/index.html.

Xie, Yihui. 2014. “knitr: A Comprehensive Tool for Reproducible Research in R.” In *Implementing Reproducible Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng, 3–32. CRC Press.