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, especially when determining competitive equality.

Major League Baseball is an often studied sport in the academic literature, in particular due to its wealth of data, mostly individualistic nature, and large sample sizes.

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.

Performance persistence

Performance persistence has long been studied in the context of actively managed funds, which attempt to select equities and other assets, usually following some investment strategy, in order to maximize returns to investors. In large part, the empirical findings have been mixed.

*Contact: chua@wharton.upenn.edu

Performance is typically considered as a fund's ability to generate 'alpha', or more precisely, creating additional returns in excess of the risk free rate, given the fund's market exposure. If we assume that the Capital Asset Pricing Model (CAPM) holds, then alpha is simply given by the slope of the following regression equation:

$$R_{p,t} - R_f = \alpha + \beta(R_{m,t} - R_f) + \epsilon_{p,t}$$

Rewritten in terms of alpha, we have:

$$\alpha = R_p - [R_f + \beta(R_{m,t} - Rf)]$$

However, this definition of alpha accounts only for risk attributable to the market. Within a particular fund and their strategies, there are additional sources of risk. Fama and French (1992) famously found that much of the excess returns that practitioners believed was alpha was actually attributable to exposure to various 'factors,' namely market capitalization (size) and 'value' behavior (book-to-market ratio).

Carlson (1970) was one of the first papers in this area, and found little predictive benefit from using the lagged returns.

Indeed, the importance of mutual funds and other actively managed funds is widely accepted to be decreasing (see, e.g. Benjamin (2016)). The mixed empirical evidence for performance-persistence is a strong driver of this movement, as investors instead choose to invest in cheap exposure to the market or particular strategies, rather than opt for an expensive actively-managed portfolio.

Competitive equality

Significance

Good question

Data

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.

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 2016). 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 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).

Methodology

There are several measures through which we measure repeat performance. We outline:

1. Contingency tables
2. Lagged regression

Contingency tables

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 a given metric. 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.

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

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

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

Then, we have sufficient framework to perform statistical tests of significance.

Hypotheses

First, we define teams as winners if they win more games than the median number of games won per team for a given year.

Hypothesis 1 *Performance in the first period is related to performance in the second period.*

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, $R_{cp} > 1$.

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.

Hypothesis 2 *Performance in the first period is related to performance in the second period.*

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.

Hypothesis 3 *Making the playoffs in the first period is related to making the playoffs in the second period.*

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.

Hypothesis 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.*

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, $R_{cp} > 1$.

Lagged Regressions

A slightly more complex model for measuring effects over time is with a lagged regression. That is, in a linear regression, is the winrate of the year before a significant indicator for the winrate of the year after? This is the approach taken by Carlson (1970).

For any given team i and season s, we can then set up a regression equation:

$$\pi_{i,s} = \beta_0 + \beta\pi_{i,s-1} + \epsilon$$

Here, $\pi_{i,s}$ is the winrate from that season. Our goal is to measure the significance of the β coefficient, which we can do through an Anova test.

The most important benefit of using a regression model is that we can control for exogenous effects, by adding them as additional factors into the regression equation. Our results from a contigency table based method may be better explained by other factors. Some factors which we intend to control for include the following:

- **Player turnover**: We would expect a team with exactly the same players between two seasons to perform approximately the same. However, if teams have massive player turnover, then changes in performance may be due to the turnover rather than performance persistence. We can measure turnover by examining the number of players on the Opening Day roster that are not on the team's ending roster. We can also measure this change relative to other teams' turnover, and use a standardized z-score.
- **Strength of schedule**: It is possible that some deviance in wins can be explained by how strong opponents were. This is typically measured by the aggregate number of games won by opponents. In the context of MLB teams, this is not likely to make a big difference, since over 162 games and many different opponents, it is unlikely to face large year-to-year changes in strength of schedule.
- **Additional lagged years**: Our initial model is defined with 1 year of lagged performance. We can also include more years for a more robust look at performance persistence.

Each of the above models described has been in the form of a linear regression, that is, we have some continuous variable π which we intend to solve for. We can also study this in the context of logistic regression, which has a binary variable y to solve for; in this case, we would use the

previously described winner/loser classifications for teams and then find the effect of past performance.

Hypotheses

We begin with a simple base case.

Hypothesis 5 *Given a regression of form $\pi_{i,s} = \beta_0 + \beta\pi_{i,s-1} + \epsilon$, one year's lagged performance is a significant regression variable.*

H_0^5: Last season's performance is not a significant factor at the $p < 0.05$ level.
H_1^5: Last season's performance is a significant factor at the $p < 0.05$ level.

We can also add more explanatory variables into the regression equation, to control for endogenous effects. We outline a number of such variables above. Let us call each such variable $x_{i,j,s}$, where i is the team, j is the effect number, and s is the season under consideration.

Hypothesis 6 *Given a regression of form $\pi_{i,s} = \beta_0 + \beta_{PP}\pi_{i,s-1} + \sum_j \beta_j x_{i,j,s} + \epsilon$, one year's lagged performance is a significant regression variable.*

H_0: Last season's performance is not a significant factor at the $p < 0.05$ level.
H_1: Last season's performance is a significant factor at the $p < 0.05$ level.

Other adjustment factors

A fact of sport, and life, is that games are never fair. Bill James's Pythagorean expectation for number of games won by a team, and its variants, rely on the assumption that the number of games a team wins is based on its quality, and the assumption that quality is proportional to the ratio of runs scored over runs allowed. Better teams score far more runs than they allow, so they win more games, and so on. In practice, we use this measure to adjust for the effect of luck in win-loss records. An unlucky team would lose in close games and win in blowouts. Conversely, a lucky team would win more close games. Thus, there could be value to using the Pythagorean expected wins rather than realized wins as the dependent variable in a regression.

It is typically given by:

$$\hat{\pi}_{i,s} = \frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2}$$

Note that the above formula is easily generalizable to other competitive sports.

To use the Pythagorean expectation, we will create a "counterfactual universe", where each team's wins are given by their expected wins, and our hypotheses are tested against this set of data.

Data Analysis

Todo in full paper!

Extensions

A core consideration in designing this analysis is how applicable these methods are to other sports. While each sport obviously requires different understandings, such as season length, each of these methods can be applied without many changes. I think it will be within scope of this paper to further incorporate data from other professional leagues and compare the results.

League	Games Played	Number of Teams	Playoff Teams
MLB	162	30	10
NFL	16	32	12
NBA	82	30	16

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

Todo in full paper!

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*.

Benjamin, Jeff. 2016. "What's driving the decade of outflows from actively managed mutual funds." http://www.investmentnews.com/article/20160508/FREE/305089998/whats-driving-the-decade-of-o

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

Carlson, Robert S. 1970. "Aggregate Performance of Mutual Funds, 1948-1967." *The Journal of Financial and Quantitative Analysis* 5 (1): 1–32. doi:10.2307/2979005.

Fama, Eugene F., and Kenneth R. French. 1992. "The cross-section of expected stock returns." *JoF* XLVII (2): 427–66. doi:10.2307/2329112.

Friendly, Michael. 2016. "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.