Gov 52 Final Replication Project

Tucker Boynton

5/7/2021

Introduction

In this report, I replicate the 2006 journal article by Jahn K. Hakes and Raymond D. Sauer titled *An Economic Evaluation of the Moneyball Hypothesis* (Hakes and Sauer 2006). In it, Hakes and Sauer examine the evolution of returns to particular baseball skills relative to how much those skills actually impact winning. They theorize that the publication of *Moneyball: The Art of Winning an Unfair Game* served as a market-correcting force that resulted in higher compensation for the ability to get on base relative to compensation for the ability to hit for power. They also hypothesize that prior to *Moneyball*, teams were overcompensating power and undercompensating getting on base relative to the true impact of those skills on winning baseball games.

Moneyball: The Art of Winning an Unfair Game is a seminal publication by Michael Lewis about the early-2000s Oakland Athletics and their general manager Billy Beane, who used analytics to find unconventional ways to win despite the franchise having one of the lowest payrolls in Major League Baseball (Lewis 2003). Prioritizing getting on base as opposed to hitting for power was central to the Beane's strategy, and Hakes and Sauer test the theory that he discovered a market inefficiency that was then corrected in the mid-2000s. Importantly, on-base percentage measures a player's propensity to reach base safely (more of a finesse skill), whereas slugging percentage measures how well a player hits for power (more of a brute strength skill).

In order to conduct this analysis, the authors first run a series of linear regressions of team win percentage on team/opponent slugging percentage and team/opponent on-base percentage. The ratio of coefficients on slugging and on-base percentage represents - albeit crudely - the relative importance of each of those statistics to winning baseball games. Then, the authors regress log player salary on lagged slugging and on-base percentage. If the market is in equilibrium, the ratio of coefficients on slugging and on-base percentage for those two regression specifications should be equivalent, as players would be rewarded for their true contributions to winning. Hakes and Sauer find that in the early 2000s, the ratio of on-base to slugging coefficients was much larger for the win percentage regression than for the salary regression, which suggests players were paid too much for power and not enough for finesse, but that there was a rise in the salary returns to on-base percentage in the mid-2000s (specifically in 2004).

Replication

In order to reconstruct Table 1, which displays coefficients for the regressions of team win percentage on on-base and slugging percentage, I run four regressions and input the results to {stargazer}. The first two columns use only on-base percentage for/against and slugging percentage for/against respectively, while the third uses all four of the aforementioned variables. In order to "restrict the coefficient [on slugging/on-base] to equal its counterpart" as the authors do in the fourth column of the original paper, I regress win percentage on on-base and slugging percentage differential (team minus opponent), whose coefficients represent the change in win percentage associated with an increase in a team's own on-base/slugging percentage or a decrease in a team's opponent on-base/slugging percentage (p. 176).

Table 1: The Impact of On-Base and Slugging Percentage on Win Percentage

	Model			
	(1)	(2)	(3)	(4)
Constant	0.508 (0.114)	0.612 (0.073)	0.502 (0.099)	0.500 (0.005)
On-Base	3.294 (0.221)		2.141 (0.296)	
On-Base against	−3.317 (0.196)		−1.891 (0.291)	
Slugging		1.731 (0.123)	0.802 (0.149)	
Slugging against		−1.989 (0.112)	−1.005 (0.152)	
Slugging diff				0.900 (0.106)
On-Base diff				2.032 (0.183)
Observations	150	150	150	150
R^2	0.825	0.787	0.885	0.884

Note: Aggregate statistics for teams from 1999-2003.
Models include year indicators. Dependent variable is win pct.

The coefficients in the fourth column suggest that a 1 percentage point increase in on-base percentage differential corresponds to an approximately 2.0 percentage point increase in win percentage, and a 1 percentage point increase in slugging percentage differential corresponds to an approximately 0.9 percentage point increase in win percentage. Notably, these regression results suggest that on-base percentage is more than two times as important to winning baseball games as is slugging percentage.

Table 2: Summary Statistics

Player salaries (in millions)

	2000	2001	2002	2003	2004
N	344	358	347	344	343
Mean	2.60	3.04	3.16	3.48	3.32
10th percentile	0.25	0.26	0.26	0.32	0.32
Median	1.50	1.66	1.75	1.56	1.25
90th percentile	6.35	7.58	8.00	9.07	8.85

	2000		2001		2002		2003		2004	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Catchers	1.92	44	2.17	47	2.13	51	2.68	45	2.47	48
First basemen/DH	3.39	50	4.12	46	4.72	49	4.64	47	3.91	55
HR < 14	1.48	194	1.50	200	1.75	211	1.95	204	1.78	200
HR > 25	5.62	59	6.38	63	7.32	54	8.18	50	8.17	52
Infielders	2.20	123	2.72	129	2.70	126	2.85	119	2.59	115
Outfielders	2.91	127	3.27	136	3.44	121	3.91	133	4.07	125

Note: Summary statistics for salaries of players with more than 130 plate appearances.

Table 2, split into two parts, displays summary statistics for player salaries based on a number of subsections of the data. In order to reconstruct the first part of the the table, I group the full player data set by year, generate summary statistics, transpose the data frame, and display results using the {gt} package. The second part requires a bit more manipulation; because the groupings do not partition the data per se (the categories "HR > 25" and "HR < 14" must be calculated separately from the positional groupings), I calculate mean and number of observations for each position and then for the home run categories. Finally, I bind the two data frames, re-orient the data to reflect the structure of the table in the paper, and again use the {gt} package to display results.

In terms of general trends, it is clear that salaries increased over the years in question, and power hitters (> 25 home runs) as well as positions that tend to correlate with more power (first base, designated hitter) were compensated more highly than non-power hitters (< 14 home runs, infielders).

Table 3: The Baseball Labor Market's Valuation of On-Base and Slugging Percentage

	All years	2000-2003	2000	2001	2002	2003	2004
On-Base	1.400	0.929	1.283	0.489	0.921	1.121	3.559
	(0.607)	(0.658)	(1.197)	(1.190)	(1.451)	(1.552)	(1.566)
Slugging	2.276	2.315	2.828	2.610	2.036	1.951	2.166
	(0.305)	(0.331)	(0.622)	(0.593)	(0.678)	(0.833)	(0.774)
Plate appearances	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0003)
Arbitration eligible	1.236	1.222	1.273	1.081	1.292	1.241	1.325
	(0.047)	(0.051)	(0.101)	(0.099)	(0.098)	(0.111)	(0.114)
Free agency	1.675	1.703	1.739	1.678	1.728	1.665	1.571
	(0.044)	(0.048)	(0.094)	(0.090)	(0.095)	(0.106)	(0.103)
Catcher dummy	0.144	0.173	0.157	0.039	0.225	0.299	0.049
	(0.054)	(0.060)	(0.123)	(0.113)	(0.117)	(0.130)	(0.129)
Infielder dummy	−0.036	−0.014	0.082	0.031	−0.096	−0.074	−0.105
	(0.040)	(0.043)	(0.086)	(0.082)	(0.085)	(0.095)	(0.098)
Constant	10.127	10.227	10.049	10.392	10.526	10.416	9.825
	(0.167)	(0.179)	(0.350)	(0.317)	(0.352)	(0.380)	(0.407)
Observations	1,780	1,436	362	369	355	350	344
R^2	0.675	0.687	0.677	0.725	0.699	0.649	0.638
Value of 1 SD increase (millions)							
On-Base			0.08	0.02	0.05	0.07	0.5
Slugging			0.92	0.81	0.46	0.36	0.47

Note: Dependent variable is ln(salary) in year t. Performance variables in year t-1.
First two models include year indicators. Includes players with 130+ plate appearances.
Final two rows show value of 1 standard deviation increase from average.
Standard errors in parentheses.

Table 3 displays the coefficients from linear regressions of log player salary on various lagged performance statistics. For each column, I simply subset the data by year(s) and run the same regression specification (year is used as a dummy variable in the first two columns). The difficult part of creating this table is generating the last two rows, which show the predicted salary change associated with a one standard deviation increase in slugging and on-base percentage. With a log-linear model, the impact of any increase depends on where you start, and the authors do not specify whether they started at the mean, median, or some other location. As such, I use a for loop to calculate the salary change associated with a one standard deviation increase in slugging/on-base percentage from the mean for each year's model and display the results using the "add.lines" option for {stargazer} tables.

The results in Table 3 show that for most of the period, there was not a statistically significant premium placed on on-base percentage (four out of five seasons), but the opposite is true of slugging percentage, which has a statistically significant coefficient (at the 5% level) in each of the five periods. Moreover, the ratio of point estimates for the coefficients on on-base to slugging percentage is clearly inconsistent with the 2:1 ratio I would expect after seeing the results of the previous regression, but the relative returns appear to approach equilibrium in 2004 (consistent with the *Moneyball* thesis).

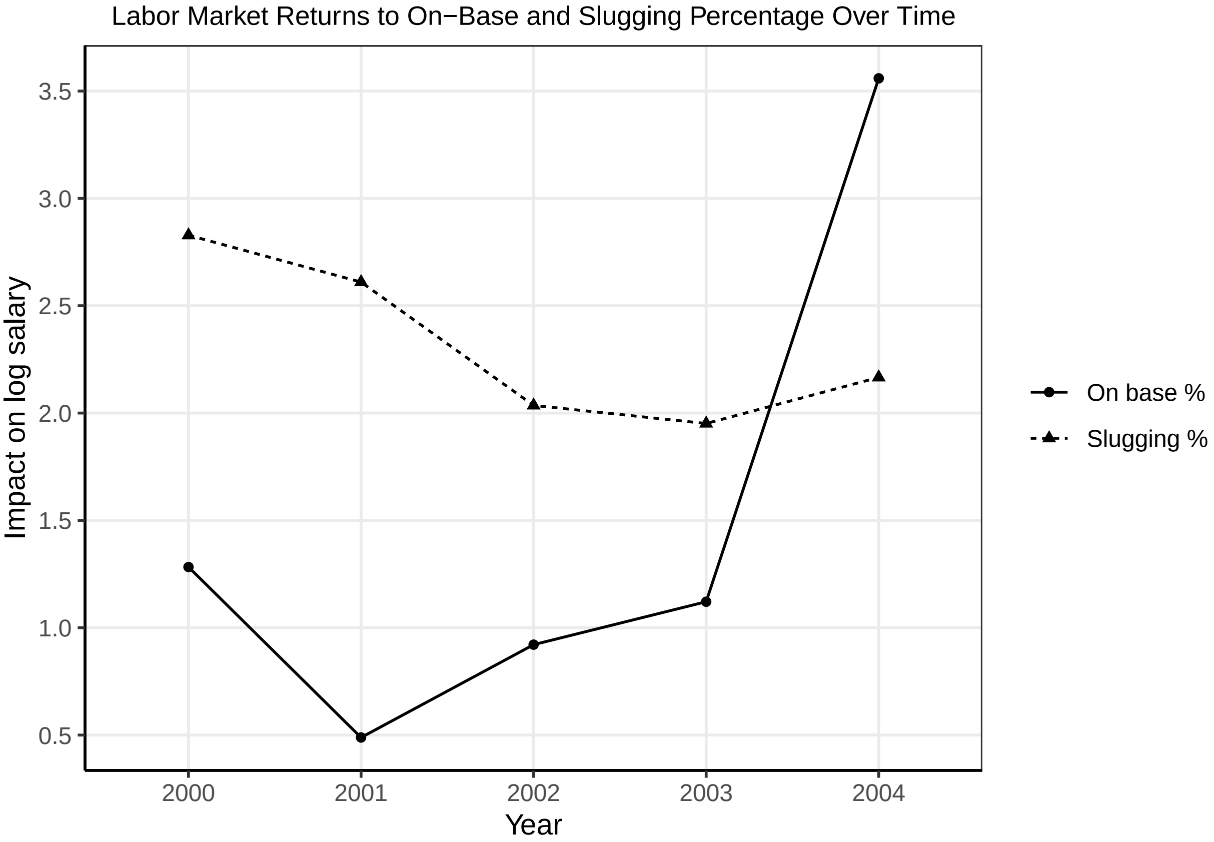

Figure 1: Returns to on-base percentage increased considerably from 2003-04.

In Figure 1, I display the coefficient estimates on slugging and on-base percentage by season. In order to take advantage of {ggplot}'s "group" option, which cleanly separates observations with different values of categorical variables (in this case, on-base vs. slugging percentage coefficients), I need the data in long form. Hence, I create separate data frames with on-base and slugging percentage coefficients, assign each observation a "type" of coefficient (either on-base or slugging %), and bind the rows. Then, I simply use {ggplot} with "group = type" and plot the evolution of the two coefficients separately, also specifying "linetype" and "shape." The visual trend here is stark: for the first four years, slugging percentage dominates, and in 2004, on-base percentage shoots up, signifying the sudden emergence of a salary premium on the ability to get on base.

For Figure 2, the data set provided by the authors includes different observations from the initial paper and seems to have different units on the salary index, but nonetheless, I plot salary index versus win percentage. I also use the "ggrepel" function from the {ggtext} package in order to label the points for which "convex == 1," which represents the team being on the frontier for efficient conversion of salary into wins (hard to tell because data set has many points missing). Although many of the data points are missing, OAK01, or Oakland Athletics 2001, is clearly visible on the frontier of converting salary into win percentage, providing visual support for the fact that Billy Beane made the most of a tough hand.

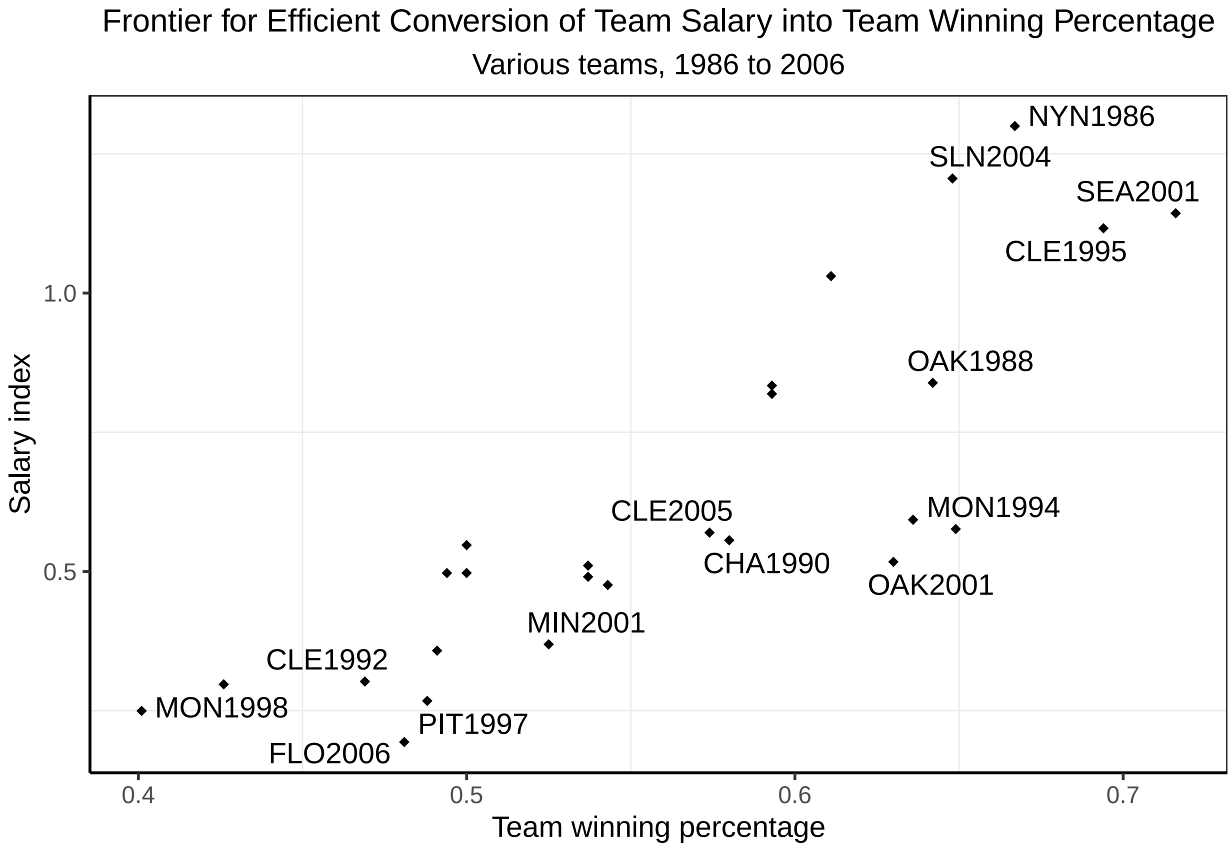

Figure 2: Labeled teams (team abbreviation and year) are on frontier of conversion of salary into wins.

Finally, I am unable to replicate the fourth table because the paper's data repository does not provide ticket price or attendance data for the Oakland Athletics, and the links cited in the paper lead to websites that are no longer operational. Fortunately, this table is not central to the analysis of returns to on-base and slugging percentage.

Overall, the coefficients in my regression tables are extremely similar (if not exactly the same) as Hakes and Sauer's results. The small discrepancies appear to come from differences in the number of observations in each regression (they are off by between five and 10 depending on the regression). Given the fact that the data provided for Figure 2 does not exactly match what is in the paper, it is likely that the other online data sets are not identical to the ones used in the paper itself, so there are likely some observations that appear in my data set but not the original (and vice versa). Overall, however, I am extremely satisfied with how closely my results match the authors' conclusions, and the only table that I was unable to replicate was fortunately not vital to the thesis of the paper.

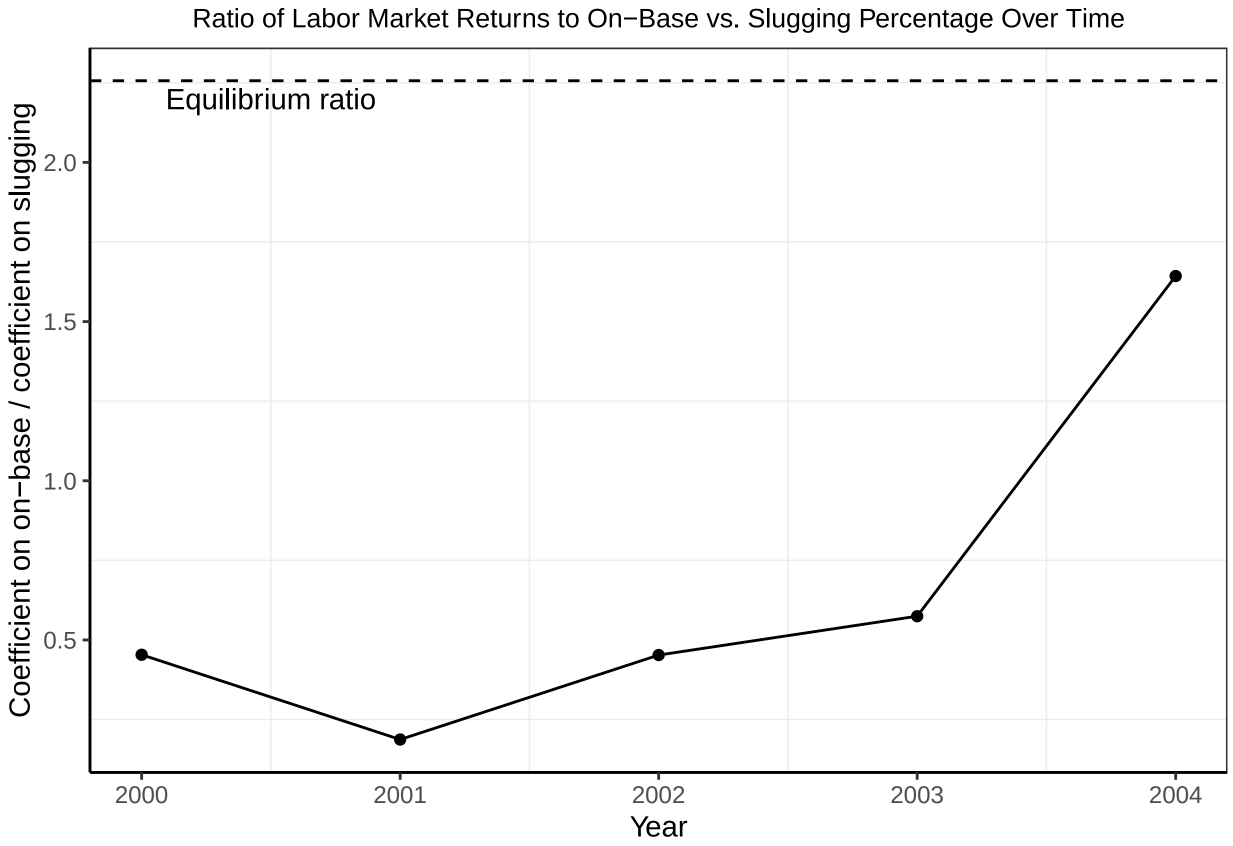

Figure 3: From 2001-04, the player salary market made significant progress in approaching equilibrium.

Extension

One extension I would like to see is a direct visual comparison between the market value and "true" value of slugging and on-base percentage over time. Theoretically, the first regression represents the "true" relative values of slugging and on-base percentage, but the authors never provide a visual aid of how the ratio of coefficients evolves over time in the labor market - we see the individual coefficients in Figure 1 but not how their ratio compares to the theoretical equilibrium ratio. In Figure 3, I plot the ratio of coefficients from the yearly regressions in Table 3 as well as the market equilibrium ratio, which is the ratio of coefficients in the fourth column of Table 1. In theory, as market approaches equilibrium, the ratio of coefficients in the salary regressions will approach the ratio of coefficients in the team win percentage regression; we see that it does, although the ratio does not reach full equilibrium by 2004.

This extension allows the reader to obtain more context in the story. Hakes and Sauer put forward tremendous evidence that there was some correction in the labor market but never explicitly show how quickly the inefficiency disappeared - or if it disappeared entirely at all.

Conclusion

Like Hakes and Sauer, I find that the early 2000s featured a clear trend of overvaluation of slugging percentage and undervaluation of on-base percentage. However, after the publication of *Moneyball: The Art of Winning an Unfair Game*, the market approached outcomes that were closer to equilibrium by providing higher relative returns to on-base percentage, altogether strengthening the central claim of the *Moneyball* hypothesis. In the future, I would be interested in revisiting this analysis with more recent data in order to see whether market efficiency has been reached. In fact, more recent analysis of the *Moneyball* hypothesis was conducted in 2011 by Anthony Farrar and Thomas H. Bruggink in 2011 (Farrar and Bruggink 2011). In closing, I would like to thank Professors Hakes and Sauer for their email correspondence about this paper and the open-source nature of their data[1]. Please reach out to me at tuckerboynton@college.harvard.edu if you have any comments, questions, or concerns.

[1] All data used in this paper can be found at http://media.clemson.edu/economics/data/sports/moneyball/ and all code at https://github.com/tuckerboynton22/gov52_replication

References

Farrar, Anthony, and Thomas H. Bruggink. 2011. “A New Test of the Moneyball Hypothesis.” *The Sport Journal* 14 (1).

Hakes, Jahn K., and Raymond D. Sauer. 2006. “An Economic Evaluation of the Moneyball Hypothesis.” *The Journal of Economic Perspectives* 20 (3): 173–85. https://www.aeaweb.org/articles?id=10.1257/jep.20.3.173.

Lewis, Michael. 2003. *Moneyball: The Art of Winning an Unfair Game*. New York: Norton.