# Pay for Play: Are Baseball Salaries Based on Performance?

Mitchell R. Watnik  
University of Missouri-Rolla

Journal of Statistics Education v.6, n.2 (1998)

Copyright (c) 1998 by Mitchell R. Watnik, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

**Key Words:** Exploratory data analysis; Model selection and validation; Regression; Stepwise model selection.

## Abstract

Well-defined measures of performance are readily available for baseball players, making the modeling of their salaries a popular statistical exercise. In this article, the salaries for non-pitchers for the 1992 Major League Baseball season are provided, along with numerous measures of the players' previous year's performances. Also included are indicators of each player's ability to switch teams. This dataset is useful in upper-division regression analysis courses because it exhibits many "real world" difficulties that can be remedied using techniques outlined in the course.

# The Dataset

The dataset consists of information about Major League Baseball players. The response variable is their 1992 salaries (measured in thousands of dollars and obtained from the New York Times of November 19, 1992). Possible explanatory variables include various measures of the players' 1991 performance. (See the [Appendix](http://ww2.amstat.org/publications/jse/v6n2/datasets.watnik.html#appendix) for descriptions of each variable.) These data were obtained from the Sacramento Bee of October 15, 1991. Students who are not familiar with baseball may be made aware that, with the exception of strike-outs and errors, all of these variables would sensibly be positively correlated with salary.

The last four numeric variables are dummy variables indicating "free agency eligibility," "free agent in 1991/2," "arbitration eligibility," and "arbitration in 1991/2." The special 1991/2 dummy variables are used because the players' union argued that owners colluded to keep the salary of free agents in 1991-2 lower. A list of free agents was obtained from the New York Times of November 13, 1991, and a list of players undergoing arbitration in 1992 was published in the New York Times on February 23, 1992. The reason these variables are important is that, at the time, baseball had rules stating that a player could not go to the team of his choice unless he was "free agent eligible," and he could only be eligible if he had a certain amount of experience. From an economics point of view, it seems reasonable that if a player is not able to market himself to the highest bidder, his salary will not be as high. At the time, "arbitration" was for players who did not have enough experience to be free agents, but had some experience in the league. In this case, the player and his team would go to an appointed "arbitrator" who would choose between the player's suggested salary and the team's suggested salary. Players who were neither "free agent eligible" nor "arbitration eligible" either accepted what their team was willing to pay them or did not play.

## Acknowledgments

I wish to thank Richard Green of the UC Davis Agricultural Economics Department for encouraging me to pursue the use and publication of this study beyond his econometrics course. In addition, I offer thanks to Tom Kirchoff, the anonymous referees employed by the Journal of Statistics Education on my paper, and its section editor and editor, Robin Lock and Jackie Dietz, respectively, for their constructive criticisms and suggestions on the final draft which improved this paper. I take full responsibility for any typos in the dataset and any errors in the text of this paper which may remain.

# Appendix - Key to Variables in baseball.dat.txt

Columns

1 - 4 Salary (in thousands of dollars)

6 - 10 Batting average

12 - 16 On-base percentage (OBP)

18 - 20 Number of runs

22 - 24 Number of hits

26 - 27 Number of doubles

29 - 30 Number of triples

32 - 33 Number of home runs

35 - 37 Number of runs batted in (RBI)

39 - 41 Number of walks

43 - 45 Number of strike-outs

47 - 48 Number of stolen bases

50 - 51 Number of errors

53 Indicator of "free agency eligibility"

55 Indicator of "free agent in 1991/2"

57 Indicator of "arbitration eligibility"

59 Indicator of "arbitration in 1991/2"

61 - 79 Player's name (in quotation marks)

Players' batting averages are calculated as the ratio of number of hits to the number of hits plus the number of outs. On-base percentage is the ratio of number of hits plus the number of walks to the number of hits plus the number of walks plus the number of outs. Therefore, the batting average is less than or equal to the on-base percentage. A batting average above .300 is very good; OBP above .400 is excellent. An RBI is obtained when a runner scores as a direct result of a player's at-bat.

# References

Denby, L. (1988), Dataset from Poster Session sponsored by the Section on Statistical Graphics of the American Statistical Association, on Statlib, ed. Michael Myers. (http://stat.lib.cmu.edu/datasets)

Hoaglin, D., and Velleman, P. (1995), "A Critical Look at Some Analyses of Major League Baseball Salaries," The American Statistician, 49, 277-285.

Mitchell R. Watnik   
Department of Mathematics and Statistics   
University of Missouri-Rolla   
Rolla, MO 65409-0020

mwatnik@umr.edu

# Introduction to Baseball

<https://entertainment.howstuffworks.com/baseball.htm>

<https://entertainment.howstuffworks.com/baseball1.htm>

<https://entertainment.howstuffworks.com/baseball2.htm>

<https://entertainment.howstuffworks.com/baseball3.htm>

<https://entertainment.howstuffworks.com/baseball4.htm>

<https://entertainment.howstuffworks.com/baseball5.htm>

<https://entertainment.howstuffworks.com/baseball6.htm>

<https://entertainment.howstuffworks.com/baseball7.htm>

**Baseball Statistics for Beginners**

<https://www.pbs.org/kenburns/baseball/beginners/stats.html>