

Rod Megraw
Econ 482
3/5/98

How are baseball salaries related to current measures of player performance?

The goal of this research paper is to investigate the relationship between salaries and recent performance metrics of Major League Baseball players. More specifically, a log-linear regression model will be applied to relate player salaries for the 1996 season to player statistics for the 1994, 1995, and 1996 seasons. The sample will be constrained to include only offensive players and statistics. That is, the focus is on batting, as opposed to pitching or fielding. Additionally, only players who competed in each of the 1994, 1995, and 1996 seasons will be included in the sample. (Interestingly, this amounts to only 319 players out of the 770 who played in 1996.)

In keeping with the guidelines of good economic research, this paper relies on a foundation of economic theory. According to basic labor economics, one would expect that baseball players with higher marginal revenue products are paid higher salaries for their services. An important assumption that this paper makes is that the marginal revenue product of a baseball player is determined by the player's key performance statistics. In most cases, this seems to be a reasonable assumption. Players who hit more home runs, bat in more runs, or get more hits, make a greater contribution to the success of the team. As both team success and great individual performance are valued by fans, it is reasonable to assume that players with better statistics attract more fans to games and help sell more team merchandise. As such, team owners are willing to pay higher salaries to players who perform particularly well, because they tend to bring in more revenue. This paper seeks to model the particular way that player salaries relate to individual performance statistics. Additionally, the model will include the league of the particular player (American League or National League), and the size of the market of the player's team. Done successfully, such a model will allow us answer several interesting questions about the relative economic value

of particular measurable skills possessed by baseball players. Additionally, the model will allow us to test the hypothesis that teams in big cities tend to pay higher salaries.

Of course, it is also important to point out that the proposed model may not include several factors that could explain player salaries. For example, player age, years of experience, lifetime performance statistics, marketability of personality, and perceived potential for future success are all factors that one could argue for inclusion in the model. While adding these factors to the model may improve its accuracy and believability, practical reasons (such as time constraints, availability of data, etc.) suggest that inclusion of additional factors can wait until the author is a highly motivated graduate student.

Before presenting the statistical analysis of the player salary model, a few details about the data used for the regression should be given. Most important, perhaps, is the source of the data. All of the data used in this project was taken from a Web site called *Sean Lahman's Baseball Archive*, which offers a wealth of historical baseball information free to the public. The URL is <http://www.baseball1.com/>. While one must always question the credibility of free data garnered over the internet, it seems that in this particular case “Sean” has had his site up since 1994 (an eternity in internet years), and has been able to withstand the scrutiny of countless baseball statistical-junkies. Furthermore, there is little reason to suspect that he would either maliciously or negligently distribute tainted baseball data to the masses. The particular statistical data used for this project were taken from the “Player Statistics” section, for the years 1994, 1995, and 1996. Salary data was obtained from the “Sabermetrics” section of the site. (As an aside, sabermetrics is defined as “the search for objective knowledge about baseball”. Additionally, the “Sabermetric Manifesto” at <http://www.baseball1.com/bb-data/grabiner/manifesto.txt> is most interesting.)

The following section provides details on each of the data series used in the model. The series names are those used in the Eviews workfile.

Salary – player salary in 1996.

AL – a Boolean variable that equals 1 if the player plays for an American League team, and equals 0 otherwise.

LG_MKT – a Boolean variable that equals 1 if the player plays for one of the following teams: Chicago Cubs, Chicago White Sox, Detroit Tigers, Houston Astros, Los Angeles Dodgers, New York Mets, New York Yankees, Philadelphia Phillies, San Diego Padres, Texas Rangers, or Toronto Bluejays.

AB_94/AB_95/AB_96 – total number of at bats in 1994, 1995, and 1996, respectively.

BB_94/BB_95/BB_96 – total number of bases on balls in 1994, 1995, and 1996, respectively.

DOUB_94/DOUB_95/DOUB_96 – total number of doubles hit 1994, 1995, and 1996, respectively.

G_94/G_95/G_96 – total number of games played in 1994, 1995, and 1996, respectively.

HR_94/HR_95/HR_96 – total number of home runs hit in 1994, 1995, and 1996, respectively.

H_94/H_95/H_96 – total number of hits in 1994, 1995, and 1996, respectively.

RBI_94/RBI_95/RBI_96 – total number of runs-batted-in in 1994, 1995, and 1996, respectively.

R_94/R_95/R_96 – total number of runs scored in 1994, 1995, and 1996, respectively.

SB_94/SB_95/SB_96 – total number of stolen bases in 1994, 1995, and 1996, respectively.

SO_94/SO_95/SO_96 – total number of strike outs in 1994, 1995, and 1996, respectively.

TRIP_94/TRIP_95/TRIP_96 – total number of triples hit in 1994, 1995, and 1996, respectively.

AVG_94/AVG_95/AVG_96 – batting average for 1994, 1995, and 1996, respectively.

For this project, a log-linear regression model was used to relate the above variables to the natural-log of player salary.

The first regression included all of the above variables, and produced some unintuitive results. First of all, the R-squared statistic was surprisingly high at .71. In the absence of other confounding results, one would be quite happy with a model that appeared to explain 71% of the variation in log salary. Of course, many of the coefficients were puzzling, to say the least. Most surprising, perhaps, was that the coefficient for *AL* was -.47 and had a t-statistic of -5.01! It is hard to imagine that just by playing in the American League a player's salary is lowered by nearly 50%! The fact that average player salaries in the American and National Leagues differ by only \$100,000 (\$1.6M in the National League vs. \$1.5M in the American League) indicates that something is wrong with the model.

Another interesting result was that team market size was not significant in the model. There was a .47 p-value for the *LG_MKT* variable, indicating that we cannot refute the hypothesis that its value is zero. Other variables that appeared to be of questionable significance were *BB_95*, *DOUB_95*, *DOUB_96*, *HR_95*, *RBI_96*, *R_95*, *R_96*, *SB_95*, *SB_96*, *TRIP_96*, *AVG_95*, and *AVG_96*. Also of interest is that fact that at a 90% significance level, both total hits and doubles in 1995 and 1996 are significantly *negative!* Clearly this model is not conforming to the economic theory presented earlier in this paper! Similarly unintuitive results were achieved in numerous additional regression models. In these additional models attempts were made to achieve more intuitive results by using only the 1995 statistics, using only the 1996 statistics, removing variables that appeared to be of questionable significance, allowing for differential returns to statistics in the two leagues, and a good deal of trial and error.

From among all the regression models run, one stood out as being particularly appealing.

This model explained log salary in terms of differential returns to hits and home runs in 1996 for the two leagues. Of particular interest in this model is the fact that the hypothesis that the coefficient of *AL* equals zero *cannot* be rejected. This may indicate that most all of the salary difference between the two leagues is explained by the differential returns to hits and home runs. While we cannot reject the hypothesis that the return to home runs is the same in both leagues, with a t-statistic of 1.28, the return to hits is significantly greater in the National League than in the American League (t-statistic = 3.06). Additionally, this model achieved an R-squared statistic of .45, which is still quite good. Perhaps most importantly, this model indicates that the inclusion of differential returns to performance statistics between the two leagues can improve the quality and believability of the model.

After achieving some modest success with a model allowing for differential returns to statistics between leagues, much thought was put into applying this idea to a new and improved model. Going back to economic theory, it occurred to the author that it is foolish to assume that baseball salaries should be based on any one particular year of performance statistics. Rather, it seems more reasonable that a player's salary is based on their record of success over a number of years. This revelation suggested that a more realistic approach would be to examine the relationship between player salary and *average* performance statistics over the past few years.

The model that made the most sense according to the (new) theory, was one that related 1996 log player salary to performance statistics averaged over 1994, 1995, and 1996, allowing for differential returns between leagues and including the "large market" dummy variable. The results of this improved model were surprisingly satisfying:

```

=====
Dependent Variable: LOG(SALARY)
Method: Least Squares
Date: 03/04/98   Time: 19:33
Sample: 1 319
Included observations: 319
=====

```

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.96744	0.185051	64.67109	0.0000
AL	0.026372	0.235372	0.112044	0.9109
LG_MKT	0.135208	0.094998	1.423268	0.1557
AL*(H_3YA)	0.003867	0.004064	0.951602	0.3421
AL*(RBI_3YA)	0.019336	0.010084	1.917365	0.0561
AL*(HR_3YA)	0.026834	0.018956	1.415588	0.1579
AL*(BB_3YA)	0.012498	0.004161	3.003691	0.0029
AL*(SO_3YA)	-0.015450	0.003190	-4.843263	0.0000
AL*(SB_3YA)	0.011304	0.008099	1.395612	0.1638
(1-AL)*(H_3YA)	0.014607	0.005229	2.793203	0.0055
(1-AL)*(RBI_3YA)	-0.004240	0.013304	-0.318694	0.7502
(1-AL)*(HR_3YA)	0.039100	0.026937	1.451523	0.1477
(1-AL)*(BB_3YA)	0.020277	0.005399	3.755381	0.0002
(1-AL)*(SO_3YA)	-0.009526	0.004004	-2.379060	0.0180
(1-AL)*(SB_3YA)	-0.002992	0.010531	-0.284149	0.7765

```

=====
R-squared          0.603395      Mean dependent var 13.48841
Adjusted R-squared 0.585131      S.D. dependent var 1.261126
S.E. of regression 0.812296      Akaike info criter 2.467976
Sum squared resid  200.5866      Schwarz criterion  2.645023
Log likelihood     -378.6422      F-statistic        33.03615
Durbin-Watson stat 2.258865      Prob(F-statistic) 0.000000
=====

```

In this model, the affect of being in the American League is expressed almost entirely through differential returns to performance statistics. Furthermore, this model suggests that playing for a large market team boosts a player's salary by 13 percent (though significant only at the 85 percent level). Examining the coefficients on the player statistics also proves to be most interesting. For example, the model indicates that National League players are rewarded with a 2 percent salary increase for bases-on-balls (averaged over 3 years), while American League players receive only a 1.2 percent gain. Consistent with economic theory, it appears that in both leagues home-runs are highly rewarded. National League players get a 3.9 percent salary boost for each HR (averaged over 3 years), while their American League counterparts enjoy a 2.6 percent increase (both

statistics significant at the 85 percent level). The R-squared statistic for the model is quite good at .60.

Residual plots against each of the explanatory variables did not indicate the presence of heteroskedasticity in the model. Furthermore, it is not clear how economic theory would explain different variances in salary for varying levels of particular performance statistics. The Durbin-Watson statistic for the model, at 2.26, does not indicate the presence of autocorrelation. (Of course, it would be somewhat odd to detect autocorrelation in this model, as it is not a time-series model.)

If a sports agent were to use the above model to advise their baseball player clients, their advice would be as follows: “Play for a large market team. If you play in the American League, concentrate on hitting home-runs, batting in runs, getting walked, and stealing bases. Do not strike out! If you play in the National League, concentrate heavily on hitting home-runs, getting walked, and getting hits. Try not to strike out. (And keep my 5 percent commission coming!)”

In conclusion, it would be an exaggeration to state that the initial objective of explaining how player salaries relate to performance statistics has been met without error. However, a model was developed that seems to yield some intuitive results and conform to economic theory. Most significantly, this project demonstrates the importance of applying sound economic theory to developing and supporting a meaningful econometric model.