Investigating Effects of Player Performance on Average Home Game Attendance in Major League Baseball

BU 495Q

Nikita Vikhliaev 100269590

Charlie Ma 100174130

Andrew Koo 203793010

Problem and Motivation

Major League Baseball (MLB) is one of the most popular sports leagues in the world. Over 70 million fans attend games annually, and many more tune-in through television and radio broadcasts. With team payrolls regularly surpassing $100 million, baseball is also a serious business; MLB teams constantly aim to maximize revenue and to increase their fan base. A common way of gauging a team's popularity and support in the community is through attendance numbers at home games. One of the ways MLB teams try to increase this metric is through acquiring and trading for new players, with the idea that better players leads to better performance, which leads to more ticket sales.

With this said, we ask the question "does having star players increase attendance numbers?" In the following analysis, we define what it means to be a "star player", tabulate the number of star players each team has had from 1995 – 2010, and evaluate whether the number of star players in a team has a statistically significant effect attendance numbers. Teams commonly break the bank to sign outstanding free-agent players; we want to see if these decisions ultimately have an effect on attendance figures

The Dataset

The data for this analysis is made up of the following:

Performance figures of MLB players: contains data from 1995 - 2010 for every player that played for a MLB team during this time. Includes data on their homeruns, stolen bases, strikeouts, etc.

Attendance figures: contains data from 1995 - 2010 of the average attendance to home games for every MLB team that played during this timeframe

Stadium capacity: contains the maximum capacity for each MLB team's home stadium from 1995 – 2010. Note that a team's home stadium sometimes changes, and this is reflected in the maximum capacity data

City population: the population of each home city that had a MLB team from 1995 - 2010.

Preparing the Analysis

We first define a "performance metric" based on which we will decide whether a player is considered a "star". The formula for the metric for each player is as follows:

$$pmetric = home\ runs + bases\ stolen$$

Though the metric appears simplistic, we argue it is a good metric for three reasons. First, it allows us to account for both "power players" and "speed players". Secondly, the metric consists of home runs and stolen bases, which are arguably some of the most attention-grabbing events in baseball – this makes sense in the context of our analysis, which aims to see if "star players" (i.e. players that draw a lot of attention) can improve attendance figures. Lastly, the metric is computationally efficient and easy to interpret for individuals of various technical backgrounds.

Next, we calculate our *pmetric* for every player that participated in MLB from 1995 - 2010 and append it to the Performance Figures dataset defined above. We then take this new Performance Figures dataset and partition it into smaller datasets, specifically, one dataset for every year 1995 - 2010 – this creates 16 data sets.

For each of these datasets we then filter it so that it contains only the players that in the 90^{th} quantile or above, based on our *pmetric*. We define these to be the "star players" for each year. For each of these 16 datasets we then sum the number of occurrences of every MLB team ID (e.g. a MLB team ID of "TOR" represents the Toronto Blue Jays, "BOS" represents the Red Sox, etc.), in effect we

count how many "star players" each team had for every year 1995 – 2010. Lastly, we re-combine these 16 datasets into one dataset. For an example of how the resulting dataset looks, see Figure 1.

As a last step, we also make the following adjustments to this dataset:

- We say that any data which contains team ID of 'CAL', 'ANA', or 'LAA' will simply represent data for 'ANA' , i.e. we say that from 1995 – 2010 , the Anaheim Angels, Los Angeles Angels of Anaheim, and California Angels were the same team, which is accurate as this team simply underwent several name changes during this timeframe
- Any data which contains team ID of 'ML4' or 'MIL' will represent data for 'MIL' – in fact ML4 and MIL represent the same team, the Milwaukee Brewers. The difference in ID codes comes from the fact that in 2000 the Brewers changed their uniform, and their representative code in the data was also changed.

Does having more "Star" players affect average attendance?

Now that we have all our data, we can test whether the number of "star players" that a team has affects the team's average attendance to home games. To answer this question, we perform the following linear regression:

Response: Average attendance to home games

Explanatory variables:

- # of "Star" Players
- Population Size of Home City
- Stadium Capacity of Home
- Year

The result of fitting this model is shown in Figure 2. Even though the model summary seems to indicate that the *Number of Star Player* has a significant effect on *average attendance*, looking at the R^2 value of this model we see it is too low to make any meaningful interpretation of the significance of the variables.

Instead, we try running 31 isolated regressions, that is, we run a regression with the same Response and Explanatory Variables as defined above, but instead of using data points for all teams at once, we run it for each team individual. The chart in Figure 3 summarizes the results. The chart is interpreted as follows:

- The *TeamID* column represents the team for which we have run the isolated regression
- The *Adj_R2* column is the adjusted R^2 we get for the isolated regression
- The *p_value* column is the p-value of the coefficient of "number of start players" for each regression
- The *sig* column is binary and represents whether the p value is significant (we say a p value =< 0.05 is significant)

We see that the adjusted R^2 we get for each team from running the isolated regressions is generally much higher than that we got from running full regression, this now allows us to make meaningful interpretations of variable significance. We see that the Number of Star Player was only found to have a statistically significant effect on average attendance for 2 teams.

Adding Star Players improves attendance, sometimes

From the preceding analysis, we've found that, for certain MLB teams, adding more "Star Player" may indeed have an effect on average home game attendance. To investigate why this is only the case for certain teams, or what qualities a team must have before it can benefit from having more "star players" would take further investigation and is beyond the scope of this analysis. Indeed, Figure 4 shows that generally, the correlation between simply the *number of star players* and *average attendance* is quite weak, suggesting that there are other factors at play. Our recommendation from this analysis is that it is more important to look at game attendance from a holistic approach, including marketing, promotions, and pricing, and not necessarily chasing after high-budget performance players in the hope that they alone will boost a team's attendance figures.

Appendix

Figure 1 – Excerpt of table showing number of start players per team per year

Team	yearID	Stadium	Capacity	City	City.Population	Avg.Attendance	star_players
ANA	1995	Angel Stadium	45050	Anaheim	285064	24287	3
ANA	1996	Angel Stadium	45050	Anaheim	287789	22476	0
ANA	1997	Angel Stadium	45050	Anaheim	292989	21553	2
ANA	1998	Angel Stadium	45050	Anaheim	297741	31102	1
ANA	1999	Angel Stadium	45050	Anaheim	300650	27816	0
ANA	2000	Angel Stadium	45050	Anaheim	328611	25518	4
ANA	2001	Angel Stadium	45050	Anaheim	331120	24703	2
ANA	2002	Angel Stadium	45050	Anaheim	332672	28464	3
ANA	2003	Angel Stadium	45050	Anaheim	334029	37330	2
ANA	2004	Angel Stadium	45050	Anaheim	334392	41675	2
ANA	2005	Angel Stadium	45050	Anaheim	333165	42033	2
ANA	2006	Angel Stadium	45050	Anaheim	331261	42059	3
ANA	2007	Angel Stadium	45050	Anaheim	330178	41551	2
ANA	2008	Angel Stadium	45050	Anaheim	332102	41194	3
ANA	2009	Angel Stadium	45050	Anaheim	334429	40005	4
ANA	2010	Angel Stadium	45050	Anaheim	336265	40134	3
ARI	1998	Chase Field	49033	Phoenix	1199173	44571	1
ARI	1999	Chase Field	49033	Phoenix	1211466	37280	3
ARI	2000	Chase Field	49033	Phoenix	1326682	36324	2
ARI	2001	Chase Field	49033	Phoenix	1342908	33783	2

Figure 2 – model summary of running regression on the full dataset for every team for every year

```
Call:
lm(formula = Avg.Attendance ~ star_players + City.Population +
    Capacity + yearID)

Residuals:
     Min       1Q   Median       3Q      Max
-23049.0  -6010.7   -634.5   6809.0  19224.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.383e+05  1.790e+05  -4.684 3.69e-06 ***
star_players    1.738e+03  2.957e+02   5.878 7.87e-09 ***
City.Population 6.490e-04  2.080e-04   3.120  0.00192 **
Capacity        2.134e-01  6.664e-02   3.202  0.00146 **
yearID          4.258e+02  8.891e+01   4.789 2.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8447 on 469 degrees of freedom
Multiple R-squared:  0.1647, Adjusted R-squared:  0.1576
F-statistic: 23.12 on 4 and 469 DF,  p-value: < 2.2e-16
```

Figure 3 – summary table of running an isolated regression for every team individually

TeamID	Adj_R2	p_value	sig
ANA	0.739093292293119	0.307451384674189	0
ARI	0.711234738605669	0.432754288907916	0
ATL	0.748583810595503	0.047982987671391	1
BAL	0.944853197961429	0.230267586353273	0
BOS	0.930417910764093	0.968799679217509	0
CHA	0.544379575665202	0.187679702395409	0
CHN	0.813870057818904	0.521145738304707	0
CIN	-0.198138192824111	0.591562876537305	0
CLE	0.795355112312532	0.0836651963086782	0
COL	0.598653499722958	0.186179744584815	0
DET	0.516093852499723	0.38655936178241	0
FLO	0.490729383378451	0.254423344964133	0
HOU	0.288089957816893	0.992586743554874	0
KCA	-0.0223870527171444	0.943308500919987	0
LAN	0.684390007129126	0.621771440829144	0
MIL	0.68496031913629	0.943018502504159	0
MIN	0.846046395610354	0.767082692688641	0
MON	0.635413380411384	0.753197777564774	0
NYA	0.936945831808643	0.211752281597304	0
NYN	0.654102653925208	0.479065779329892	0
OAK	0.415095623132918	0.284770573503231	0
PHI	0.803424597387642	0.588241032741585	0
PIT	0.660026206989979	0.00205816782282386	1
SDN	0.228829435351754	0.984425257813314	0
SEA	0.321410138026212	0.662509923648817	0
SFN	0.91537317568908	0.17151771348579	0
SLN	0.614742413022133	0.600850326827932	0
TBA	0.482332795897678	0.637237184307781	0
TEX	0.240645175217821	0.366728295801052	0
TOR	0.500828938842499	0.148388053442333	0
WAS	0.74622607912454	0.439822855444553	0

Figure 4 – 'Number of star players' vs. Average Attendance per team per year