A STUDY OF HOME RUNS IN THE MAJOR LEAGUES

KEITH KNIGHT
ALEX SCHUSTER[*]
Department of Statistics
University of Toronto

May, 1992

Abstract

It is well-known that the rate of home runs varies significantly among different major league stadiums. Using data from 1979 to 1991, an attempt is made to rank these stadiums by means of an additive Poisson regression model. The model also yields estimates of the home run hitting proficiency of the 26 major league teams over the period 1979 to 1991. We also use the stadium estimates to rank home run hitters in the American and National Leagues for the 1991 season.

KEY WORDS: Baseball, home runs, generalized additive models, Poisson regression.

1 Introduction

Since the end of the "dead ball" era of baseball around 1920, home runs and home run hitters have provided baseball with much of its intrigue. Some of baseball's most dramatic and remembered moments have resulted from home runs (for example, Bobby Thomson's home run to win the National League championship in 1951 or Kirk Gibson's pinch hit home run to win the first game of the 1988 World Series) and traditionally, most of its best known and highest paid players have been home run hitters.

[*] This project was initiated while the second author held an NSERC Undergraduate Summer Research Fellowship in the summer of 1990 at the University of Toronto.

A feature which distinguishes major league baseball from other North American professional team sports is the total absence of uniformity in the dimensions of its playing surfaces. The rules of baseball essentially dictate only that the distance between the bases be 90 feet and that the distance from the pitcher's rubber to home plate be 60 feet 6 inches; the distance from home plate to the outfield fences as well as the height of the outfield fences are not specified. If one considers other factors which vary significantly between cities such as weather (for example, temperature and wind) and altitude, it is reasonable to expect sizable differences in the number of home runs hit in different stadiums. (Strangely, there does not seem to be any strong relationship between the size of the playing field and the number of home runs hit; for example, the playing field at Wrigley Field in Chicago is larger than at Houston's Astrodome despite the fact that the Astrodome is generally acknowledged to be the most difficult stadium in which to hit a home run and Wrigley Field among the easiest.)

Determining the most and least favourable stadiums for home run hitters seems to be a fairly complicated problem. The naive solution to this problem is to simply count the number of home runs hit in each stadium over a period of time. The main problem with this approach is obvious: doing so does not account for the home team's ability to hit home runs. For example, Fulton County Stadium in Atlanta has long had a reputation of being a haven for power hitters but this may, in part, be attributed to the fact that the Atlanta Braves have traditionally been a power-hitting team. (In fact, it seems reasonable to assume that a team will try to obtain players whose abilities are suited to its home field.) Other stadiums (for example, Seattle's Kingdome) are reputed to be more favourable to power hitters despite the fact that the home team has not hit an extraordinary number of home runs. The Elias Sports Bureau (Siwoff *et al*, 1992) ranks the stadiums by looking at the ratio of the number of home runs hit in home games (by the home team and its opponents) to the number hit in away games (over a period of five years). This ratio, while giving a useful measure of the effect of stadiums, is flawed because of the fact that teams do not play the same number of games in each stadium. This "home team" factor is complicated by the fact that the strength of a team varies from year to year. Because a small fraction of players account for the majority of home runs on any given team, it would not be unexpected for a team to greatly increase (or decrease) its home run production over a short period of time as a result of adding (or losing) good home run hitters. However, with very few exceptions, the composition of most teams does not change greatly from year to year.

In this article, we will come up with a plausible probability model for the number of home runs hit under various combinations of factors. There seems to be a significant year-to-year variation in the number of home runs; for example, many players greatly exceeded their previous home run production in 1987 and then returned to form subsequently. A classic, but not atypical, example of this phenomenon was Wade Boggs of the Boston Red Sox whose career high prior to 1987 was 8 home runs; in 1987, he hit 24 home runs and followed up with 3 in 1988 and 5 in 1989. (The reason for the increase in home runs during the 1987 season has not been adequately explained although some experts feel that the baseballs were extraordinarily lively that year.) We should also expect there to be substantial differences in the number of home runs hit by American league and National league teams because of the designated hitter rule in the American league. This rule allows teams to replace a field player (almost always the pitcher) by the so-called designated hitter in the batting order; the designated hitter is usually a power hitter. Another factor which may influence the number of home runs hit by a given team in a game is whether or not the team is the home team or the visiting team; the visiting team will sometimes have one more opportunity to bat in a game (if the home team is ahead after the top of the ninth inning) and therefore may exceed their home field output.

Obviously, these are not the only factors which account for the number of home runs in a given game; weather, pitching and length of game are some obvious factors which may affect the number of home runs hit in any given game. However, since game-by-game data are difficult to obtain, we are forced to work with aggregate totals for each season; accordingly, these four factors would seem to be the most important. Another point to keep in mind when attempting to rank stadiums is the fact that any given stadium may favour a certain type of hitter (for example, right handed) over others for a number of reasons; for example, a stadium may have its right-field fence closer to the plate than its left-field fence or the prevailing wind may blow out to a certain field.

To carry out our analysis, we obtained data on the number of home runs hit by each of the 26 major league baseball teams in each stadium from 1979 to 1991 as well as the number of games played by each team in each stadium during a given year. The data were obtained from *The Baseball Guide* (1980-1992) which is published annually by *The Sporting News.*

2 A Model for Home Runs

Since the number of home runs hit by a team in a game is typically a small number (in almost all cases less than 5), it seems reasonable to model the number of home runs hit by a team by Poisson distribution (or some related distribution) whose mean depends on a number of factors. For example, a Poisson random effects model was used by Albert (1992) to estimate the home run ability of 12 great major league players.

Assuming that the only factors to be considered are season, league, team and stadium, a simple model that we might consider is a Poisson log-linear regression model (McCullagh and Nelder, 1989). More precisely, if x_{ijkl} is the number of home runs hit by team k of league j playing in stadium l in a single game of season i, then x_{ijkl} has a Poisson distribution with mean μ_{ijkl} where

$$\log(\mu_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_{ik} + \delta_l;$$

the parameters α_i, β_j, γ_{ik} and δ_l indicate the effects of season, league, team (depending on season) and stadium. To use this model, we would require game-by-game data; in fact, we have only yearly totals. Suppose that n_{ijkl} is the number of games played by team k of league j playing in stadium l during season i and let y_{ijkl} be the sum of x_{ijkl} over the n_{ijkl} games. Then if the x_{ijkl}'s are independent, we get that y_{ijkl} has a Poisson distribution with mean $\phi_{ijkl} = n_{ijkl}\mu_{ijkl}$ and hence

$$\log(\phi_{ijkl}) = \log(n_{ijkl}) + \mu + \alpha_i + \beta_j + \gamma_{ik} + \delta_l.$$

While the assumptions used to get to the model are not likely to be valid (for example, the game-by-game home runs x_{ijkl} are probably not independent), the model is nonetheless still useful and also conceptually appealing. Moreover, we can weaken the assumption that the yearly totals y_{ijkl} are Poisson to assume only that $\mathrm{Var}(y_{ijkl})$ is proportional to $E(y_{ijkl})$. Such a model is called a quasi-likelihood model (McCullagh and Nelder, 1989). Note also that we could easily add a parameter to account for the home/away factor.

We shall explore a variation of the model given above. Note that this model allows for arbitrary season effects α_i and "team" effects γ_{ik} as i, the season, varies from 1979 to 1991. We will consider a model where α_i and γ_{ik} are smooth functions of i; that is, we assume that the differences $|\alpha_i - \alpha_{i+1}|$ and $|\gamma_{ik} - \gamma_{i+1,k}|$ are small in some vague sense. This variation falls into the class of generalized additive models (Hastie and Tibshirani, 1990). It is useful to write the model as

$$\log(\phi_{ijkl}) = \log(n_{ijkl}) + \mu + \alpha(i) + \beta_j + \gamma_k(i) + \delta_l.$$

to indicate that $\alpha(i)$ and $\gamma_k(i)$ are smooth functions of i.

As previously mentioned, we have data from the 1979 to 1991 seasons. Also during the time period studied, the Minnesota Twins, Toronto Blue Jays and Chicago White Sox switched stadiums and several other stadiums were modified. These modifications typically involve raising or lowering all or part of the outfield fence, or altering the dimensions of the stadium somehow; thus it seems reasonable to believe that the rate of home runs would be affected in these stadiums. However, due to the complications involved in obtaining data on when modifications were made, we will assume for simplicity that any modifications would have no effect; we will discuss the validity of this assumption later. The Minnesota Twins moved from Metropolitan Stadium to the Metrodome in 1982 and the Toronto Blue Jays moved from Exhibition Stadium to the SkyDome after one third of the 1989 season. We will incorporate these stadium changes into our analysis. (We will pretend that all home runs hit in Toronto in 1989 were hit at the SkyDome.) The Chicago White Sox moved into a new stadium at the beginning of the 1991 season; we will estimate an effect for the new Comiskey Park based on only one year (1991) of data. Hence we need to estimate effects for 17 stadiums in the American League rather than 14.

Tables 1 and 2 give the average number of home runs per game hit in American and National League stadiums from 1979 to 1991 while Table 3 gives the average number of home runs hit per game in the American and National leagues from 1979 to 1991.

3 Results of Data Analysis

As mentioned previously, our analysis involved data from the 1979 to 1991 seasons. Thus for each National League season, we should have $12^2 = 144$ observations and $14^2 = 196$ observations for each American League season. However, the 1981 season was affected by a players strike which wiped out roughly one-third of the season. Because of this strike, there were seven cases where one team did not play in another team's stadium. Therefore, we have a total of $13 \times (144 + 196) - 7 = 4413$ observations.

Analysis of the data indicated that the home/away factor was insignificant (practically and statistically) in explaining the number of home runs hit. The parameters in the model were estimated by using the backfitting algorithm (Hastie and Tibshirani, 1990) to maximize the Poisson likelihood function. The season effects $\alpha(i)$ and team effects $\gamma_k(i)$ were estimated using the loess smoother (Cleveland and Devlin, 1988); to handle the anomalous year 1987

RANK	STADIUM	HR/GAME
1	Tiger Stadium (Detroit)	1.09
2	Kingdome (Seattle)	1.01
3	Memorial Stadium (Baltimore)	1.00
4	Metrodome (Minnesota)	0.98
5	New Comiskey Park	0.94
6	Anaheim Stadium (California)	0.94
7	Exhibition Stadium (Toronto)	0.93
8	Yankee Stadium (New York)	0.91
9	Fenway Park (Boston)	0.90
10	SkyDome (Toronto)	0.90
11	County Stadium (Milwaukee)	0.84
12	Alameda County Coliseum (Oakland)	0.83
13	Arlington Stadium (Texas)	0.82
14	Cleveland Stadium (Cleveland)	0.79
15	Old Comiskey Park (Chicago)	0.76
16	Metropolitan Stadium (Minnesota)	0.71
17	Royals Stadium (Kansas City)	0.64

Table 1: Home runs per game in American League stadiums 1979-1991.

RANK	STADIUM	HR/GAME
1	Wrigley Field (Chicago)	0.93
2	Fulton County Stadium (Atlanta)	0.91
3	Riverfront Stadium (Cincinnati)	0.84
4	Veterans Stadium (Philadelphia)	0.76
5	Jack Murphy Stadium (San Diego)	0.75
6	Candlestick Park (San Francisco)	0.74
7	Three Rivers Stadium (Pittsburgh)	0.73
8	Shea Stadium (New York)	0.72
9	Dodger Stadium (Los Angeles)	0.68
10	Olympic Stadium (Montreal)	0.65
11	Busch Stadium (St. Louis)	0.51
12	Astrodome (Houston)	0.44

Table 2: Home runs per game in National League stadiums 1979-1991.

YEAR	AL	NL
1979	0.89	0.74
1980	0.82	0.64
1981	0.71	0.56
1982	0.92	0.67
1983	0.84	0.72
1984	0.87	0.66
1985	0.96	0.73
1986	1.01	0.78
1987	1.16	0.94
1988	0.84	0.66
1989	0.76	0.70
1990	0.79	0.78
1991	0.86	0.73

Table 3: Home runs per game in American and National Leagues 1979-1991.

in the estimation of the season effect, a separate effect was estimated for 1987. All of the computations were carried out using the S language (Becker *et al*, 1988).

The estimates of δ_l from our model allow us to rank the stadiums in each league; Table 4 contains the rankings for the American League while Table 5 contains the rankings for the National League. The columns headed RATE give the adjusted home run rate (actually estimates of $\exp(\delta_l)$) for each stadium with the league average set to 1.000; for example, the rate for the Kingdome is 1.229 meaning that about 22.9% more home runs would be hit at the Kingdome than the American League average under similar conditions. Because there is no interleague play during the regular season, we are unable to compare stadiums from the two leagues. The standard error of the difference between the estimated rates for two stadiums is approximately 0.05; hence the certainty of the rankings must be taken with a grain of salt. There is much less certainty in the estimates for Metropolitan Stadium, the SkyDome and the new Comiskey Park as their estimates are based on fewer years of data.

An interesting by-product of our analysis are the estimates of the year and team effects. The estimates of the team effects are particularly interesting as they give some indication of the evolution of a given team's home run productivity from 1979 to 1991. Figures 1 through 6 display these estimates for Houston, Kansas City, Los Angeles, Milwaukee, New York Mets and Toronto; the league average for each year is set to 1.000 and a rate of 1.11 would mean that a team's home run production was 11% greater than the league average for that year (after adjusting for stadium effects). With the exception of Kansas City, these plots indicate great changes in the home run production of the teams over this 13 year period; in many ways, these changes seem to reflect the fortunes of the teams. (For example, note the steady rise of the Toronto Blue Jays from 1979 to 1989.) In contrast, the plot for Kansas City seems to indicate an amazing consistency in home run production over this period.

4 Discussion

We assumed in our analysis that the effect of each stadium remains constant from season to season. This assumption may be reasonable (at least for the sake of approximation) if the configuration of each stadium were unaltered; however, the reality is that many of the major league stadiums were modified in some way. These modifications can involve moving or changing the height of the outfield fence or adding a structure to the existing stadium. The Chicago Cubs began playing night games at Wrigley Field in 1990; this may affect the

RANK	STADIUM	RATE
1	Kingdome (Seattle)	1.229
2	Tiger Stadium (Detroit)	1.181
3	Metrodome (Minnesota)	1.127
4	New Comiskey Park (Chicago)	1.122
5	Memorial Stadium (Baltimore)	1.066
6	Anaheim Stadium (California)	1.053
7	SkyDome (Toronto)	1.045
8	Exhibition Stadium (Toronto)	1.036
9	Fenway Park (Boston)	1.014
10	Yankee Stadium (New York)	0.974
11	Cleveland Stadium (Cleveland)	0.959
12	Metropolitan Stadium (Minnesota)	0.959
13	Arlington Stadium (Texas)	0.942
14	County Stadium (Milwaukee)	0.936
15	Alameda County Coliseum (Oakland)	0.898
16	Old Comiskey Park (Chicago)	0.867
17	Royals Stadium (Kansas City)	0.731

Table 4: Rankings of American League stadiums.

RANK	STADIUM	RATE
1	Wrigley Field (Chicago)	1.253
2	Fulton County Stadium (Atlanta)	1.250
3	Riverfront Stadium (Cincinnati)	1.163
4	Jack Murphy Stadium (San Diego)	1.126
5	Three Rivers Stadium (Pittsburgh)	1.029
6	Veterans Stadium (Philadelphia)	1.027
7	Candlestick Park (San Francisco)	0.982
8	Shea Stadium (New York)	0.970
9	Olympic Stadium (Montreal)	0.896
10	Dodger Stadium (Los Angeles)	0.879
11	Busch Stadium (St. Louis)	0.798
12	Astrodome (Houston)	0.630

Table 5: Rankings of National League stadiums.

number of home runs hit at Wrigley Field. An analysis of the data from 1988 to 1991 shows that, in the National League, Wrigley Field drops to third place behind Riverfront Stadium (Cincinnati) and Jack Murphy Stadium (San Diego). Likewise, the analysis of the 1988 to 1991 data indicates that Seattle's Kingdome has dropped from first to third place behind Tiger Stadium (Detroit) and Yankee Stadium (New York). The Metrodome (Minnesota) has dropped from third to seventh, perhaps due to raising its rightfield wall. However, given the uncertainty in the results, it is not unexpected to see teams move up or down several positions. Regarding the effect of a given stadium as fixed (in any way) is certainly unrealistic. For some stadiums, the number of home runs hit seems to depend strongly on the weather; the classic example of this phenomenon seems to be Wrigley Field. The number of home runs hit in Toronto's SkyDome seems to increase when the roof is closed.

It may be argued our stadium rankings are biased somewhat by the quality of the home team's pitching; for example, a consistently poor team like the Seattle Mariners might surrender more home runs over a season than would a better team and this would naturally imply more home runs hit in its home stadium. While this argument has some merit, there is some debate as to whether poor pitchers give up more home runs than do good pitchers. In fact, there have been many excellent pitchers (for example, Ferguson Jenkins and Bert

Blyleven) who gave up an inordinate number of home runs over their careers in spite of their overall success. To get some idea of the relationship between pitching quality and susceptibility to home runs, we looked at some data on pitchers from the 1991 season using earned run average (ERA) as a measure of pitching quality. Figures 7 and 8 give plots of home runs (per nine innings) versus earned run average (ERA) for pitchers in the National and American Leagues who pitched 50 or more innings during the 1991 season; the solid lines in each plot are non-parametric estimates of mean home run rate given the ERA. While the plots seem to indicate an association between ERA and the home run rate for a given pitcher, there is substantial variation in the home run rate for any given ERA. Indeed, one might expect to see a greater association given that home runs contribute directly to the ERA. It seems difficult to assess the effect of pitching without more detailed data which would allow a "pitching" effect to be included in the model. (In fact, it would be possible to fit such a model to our data using the EM algorithm although this will not be pursued here.)

Despite its gross simplicity, our model seems to fit the data remarkably well as demonstrated by a careful examination of various residual plots. Of course, it almost goes without saying that our model is by no means the best model. We would certainly encourage other interested statisticians to look at this or other related data sets.

(The data set analyzed in this article can be obtained from the first author by sending electronic mail to `keith@utstat.toronto.edu`.)

5 Application: Ranking Home Run Hitters

One possible application for the rates given in Tables 4 and 5 would be in ranking home run hitters on the basis of the number of home runs hit in each stadium. We will adjust the total number of home runs hit by a given player by dividing the number of home runs hit in each stadium by the corresponding rate for each stadium and then summing these quotients. For example, if a player hit 15 home runs, 10 at the Kingdome and 5 at Tiger Stadium, his adjusted total would be $10/1.229 + 5/1.181 = 12.4$. This approach estimates the number of home runs a player might have hit had he played all his games at the non-existent "average" stadium.

As an illustration of the simple procedure described above, we will attempt to rank the top home run hitters for the 1991 season. Tables 6 and 7 list the top home run hitters in the

PLAYER	RAW	ADJ
Jose Canseco (Oakland)	44	45.6
Cecil Fielder (Detroit)	44	39.2
Cal Ripken (Baltimore)	34	32.6
Joe Carter (Toronto)	33	32.3
Frank Thomas (Chicago)	32	29.2
Danny Tartabull (Kansas City)	31	35.3
Mickey Tettleton (Detroit)	31	29.9

Table 6: 1991 American League home run leaders with “adjusted” home runs (ADJ).

PLAYER	RAW	ADJ
Howard Johnson (New York)	38	39.1
Matt Williams (San Francisco)	34	34.6
Ron Gant (Atlanta)	32	29.8
Andre Dawson (Chicago)	31	27.3
Fred McGriff (San Diego)	31	29.6
Will Clark (San Francisco)	29	30.2
Paul O’Neill (Cincinnati)	28	25.0
Darryl Strawberry (Los Angeles)	28	33.2

Table 7: 1991 National League home run leaders with “adjusted” home runs (ADJ).

American and National Leagues for the 1991 season and give the adjusted home run total (ADJ) for each player as well as the unadjusted number of home runs (RAW). (The data were obtained from *The Baseball Guide* (1992).) In the American League, note that Jose Canseco is clearly in first place in the adjusted standings despite being tied with Cecil Fielder in the unadjusted standings. Danny Tartabull moves from sixth place in the unadjusted standings to third in the adjusted standings. Likewise, in the National League, Darryl Strawberry moves from seventh place to third in the adjusted standings.

References

Albert, J. (1992) A Bayesian analysis of a Poisson random effects model for home run hitters. *American Statistician.* **46**, 246-253.

Becker, R.A., Chambers, J.A. and Wilks, A.R. (1988) *The New S Language.* Pacific Grove, CA: Wadsworth.

Cleveland, W.S. and Devlin, S.J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association.* **83**, 596-610.

Hastie, T. and Tibshirani, R.J. (1990) *Generalized Additive Models.* New York: Chapman and Hall.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models (2nd edition).* New York: Chapman and Hall.

Siwoff, S., Hirdt, S., Hirdt, T. and Hirdt, P. (1992) *The 1992 Elias Baseball Analyst.* New York: Fireside.

The Baseball Guide. (1980-1992) St. Louis: The Sporting News.