

A Model for Attendance at Major League Baseball Games

Brian Stoner
Executive Summary

This paper builds a model for predicting the average per game home attendance in Major League Baseball. The model does a remarkable job at “explaining” attendance in that it accounts for over 98% of what determines a team’s attendance. This does not appear to simply be a quirk in the data. A resampling analysis of the regression showed that the model itself is quite stable.

The results of the generalized least squares regression reveal several interesting properties. First, the most important factor in attendance is the previous year’s attendance (both the absolute attendance and the percentage it was filled). What this seems to say is that, for better or worse, there is a certain consistency in the attendance for a certain team. If a team can do something to break this pattern (i.e. winning a lot of games or building a new stadium), it should carry over into the following season and become a new pattern.

The second revelation is that ticket prices have a much smaller effect on attendance than many believe. The sample only consisted of actual data, so there is no way of knowing with complete certainty how an extremely high or low ticket price would affect attendance. However, the regression suggests that a team which performed well and had high attendance figures for the previous season can raise ticket prices and (while accounting for the drop in predicted attendance) increase ticket revenues.

The model performs reasonably well at predicting the future attendance with a one year horizon. Its main fault is how much it relies on the previous year’s attendance in its predictions. This reliance generally causes underestimation of attendance in the first year of a new stadium and overestimation in the second year. Some of these errors should be cleared away in its estimation of the effect that a new stadium has on attendance, but this estimate appears to be too low. A team’s stadium is only new in one year, so the number of data points for new stadiums is rather small.

The Model

The goal of this paper is to build a model that predicts average home attendance at Major League Baseball games and test its effectiveness in predicting. Specifically, this model is constructed using data from 1998 to 2001 and will be tested on how well it predicts the year to date average attendance for 2002.

There are a number of expectations that shape the questions for this paper. The first expectation is that if the price of a ticket goes up (holding all else constant), the quantity of tickets demanded will decrease. Second, if a team performs better and price is held constant, it is reasonable to assume the quantity of tickets demanded will increase. The third expectation is that if a team carries a high payroll (indicating a higher ‘quality’ product on the field), quantity of tickets demanded will be greater.

The model that is used to estimate attendance is:

$$ATTEND = b_0 + b_1TICKET + b_2TICKET^2 + b_3SEATS + b_4PREVATTEND + b_5\%CAPACITY + b_6NEWSTAD + b_7SALARY + b_8WINS + b_9MKTSIZE$$

The first two variables of the model are the current year’s ticket price and the square of this ticket price. These will give an answer the question regarding the first expectation about ticket prices and quantity demanded. The rationale is that teams respond to the previous years outcomes by raising or lowering ticket prices. Fans, in turn, respond by choosing to attend games in greater or less frequency over the course of a season.

The third, fourth and fifth variables are the number of seats in a stadium (SEATS), the previous year’s attendance (PREVATTEND), the capacity the stadium was filled the previous year (%CAPACITY, expressed as a percentage). These reflect a number of somewhat abstract ideas. If more seats are built, will more seats be sold? If we can sell more tickets this year, will it be able to carry over to next year? This is a question of what kind of customer base a team has and if it is a dependable one.

The final three variables reflect the “quality” component of a team’s product and another demand component. NEWSTAD is a binary variable for whether or not they play in a brand new stadium (opened in the

current season). SAL is the current year's payroll (in millions) and WINS is the number of wins for the season. MKTSIZE is the size of team's home television market.

The data that was used for this paper was gathered from several sources. Ticket prices were collected from components that went into Team Marketing Report's Fan Cost Indices for the year's 1998 to 2002. Attendance data and number of wins came from BaseballReference.com, the product of Sean Forman, a math professor at St. Joseph's University. Salary data was taken from USA Today. Estimated television market sizes were gathered from a study done using Nielsen television market sizes that was posted on rec.sport.baseball.

Regression Results

The generalized least squares regression produces some very remarkable results. First, the R-squared value is an astonishing 0.9838. Second, of the 10 regression variables (including the constant), only 3 coefficients were not significant at the 95% level. The following is the S-Plus printout for the results of the regression.

	Value	Std.Error	t-value	p-value
(Intercept)	-25425.74	1656.928	-15.34511	<.0001
ticket	125.53	156.127	0.80401	0.4232
ticket.2	-7.89	3.849	-2.04885	0.0429
seats	0.47	0.022	21.27954	<.0001
prev	0.16	0.029	5.45411	<.0001
cap	40019.56	1364.734	29.32406	<.0001
newstad	2238.94	640.765	3.49417	0.0007
sal	9.70	7.645	1.26861	0.2073
wins	19.06	11.504	1.65650	0.1005
mkt.size	7.25	2.741	2.64355	0.0094

This regression answers the initial questions I posed quite well. The first question, regarding ticket prices and attendance can be answered by taking the derivative of attendance with respect to ticket prices. It yields:

$$\frac{\partial ATTEND}{\partial TICKET} = 125.53 + (2) \times (-7.89) \times TICKET$$

Solving this for 0, TICKET is \$7.95. This means if ticket prices are at or above \$7.95, an increase will decrease attendance. Of the 118 ticket prices in the survey, none are below \$7.95 and the mean ticket price over the four years of data is \$15.79. It should be noted that the TICKET coefficient is not significant, but the TICKET² coefficient is significant at the 95% level.

The second expectation was if a team improved, so would attendance. This is true – as the WINS coefficient is positive. However, it is only 19.06, meaning that about 19 fans will come for every win. Going

from 81 wins (a .500 season) to 116 wins brings an additional 53,865 fans to the ball park over the course of a season or 665 per game.

The final expectation was that if there are better, higher paid players on a team, attendance will increase. The SAL coefficient is 9.7, meaning 10 additional fans will show up to each game for every extra million a team spends on its players. Over a season, this only amounts an extra 810 fans. The SAL coefficient is not significant.

The market size coefficient is rather interesting. The data took values ranging from a low of 38 (Kansas City Royals) to a high of 262 (New York Yankees). If the Royals were to move to Yankee Stadium, it estimates an additional 1,600 fans would come to the games. For the Royals, this is an estimated increase of over 9%. If the Yankees were to move to Kansas City, their attendance would drop by the same 1,600, but the percentage change is only 4%. This lends some additional empirical weight to the argument regarding the structural disparity that small market teams face.¹

Extensions

The results were highly significant and had an R-squared very close to one, but some of the results seemed a little strange. Because of this I became curious about the stability of this model and if it was being heavily influenced by a few extreme observations. To evaluate this, I turned to the bootstrap to compare the observed values against the estimates that are built empirically.

To estimate the stability of the model, I turned to the jackknife. That is, I took sample we have and systematically removed one observation. The coefficients for the generalized least squares regression were calculated from the remaining 117 observations. The results of the jackknife are listed below

	Observed	Bias	Mean	SE
(Intercept)	-25425.7412	-2.313e+002	-25427.7178	1742.71063
ticket	125.5281	4.733e+001	125.9327	168.43828
ticket.2	-7.8861	-1.373e+000	-7.8979	4.08225
seats	0.4702	-2.068e-003	0.4702	0.02753
prev	0.1557	-3.438e-004	0.1557	0.03467
cap	40019.5569	-6.621e+001	40018.9910	1581.89130
newstad	2238.9394	-1.072e+001	2238.8477	777.80141
sal	9.6991	-4.937e-002	9.6987	6.68751
wins	19.0570	2.754e-001	19.0594	11.22739
mkt.size	7.2466	-1.320e-001	7.2454	2.48840

¹ For a further discussion on this topic, I recommend the report issued by the Commissioner's Blue Ribbon Panel on Baseball Economics, which is available at http://www.mlb.com/mlb/downloads/blue_ribbon.pdf

The column named Observed is the coefficients for the initial sample. The bias is the difference between the Mean (average of jackknife sample coefficients) and the Observed. The standard error is the standard deviation of the distribution of the jackknife coefficients.

The most surprising part of the jackknife was how small the biases were. I expected them to be a little larger than they are because of outlying values in some of the data (one team has extremely high ticket prices and others have extremely high or low salary). Because the biases are so small (the observed values are close to the mean of the empirical estimate) it indicates that the model and its coefficients are quite stable. Because they are stable, I felt confident in proceeding to the forecasting and decision analysis portion of the paper.

Predictions and Results

To forecast the attendance for the 2002 season, we have to make a certain number of assumptions. The most careful one is in the number of wins for the season. Since we don't actually know what this number will be, we can simply guess what it will be by using last year's win total or use a formula. One method is the Pythagorean winning percentage, described by Bill James (the father of sabermetrics – "the scientific research of all things baseball"). The most common method of finding the Pythagorean winning percentage is taking the total runs scored squared divided by the sum of runs scored squared and runs allowed squared. This is an estimate of a winning percentage, so multiplying by 162 it will yield an estimated number of games that will be won for an entire season. This is a simple, yet remarkably effective predictor. To predict the 2002 games won, I used the Pythagorean wins based on the year to date runs scored and runs allowed.

The predictions themselves had a standard error of 3,901, which is over three times that of the standard error of the regression's residuals (1,234.14). The errors on the prediction ranged from a very low 291 seats to an extremely high 12,663 seats. After considering the errors, I noticed that of the three teams with errors above 10,000 – one opened a new stadium in 2000 and another opened a stadium in 2001. The NEWSTAD coefficient is only 2,239, indicating that 2,239 extra fans should show up to each game if a new stadium is built. I (very unscientifically) hypothesize that the new stadiums drew fans in much higher numbers, but this honeymoon ended soon after because of their poor performance or press coverage.

Implications for Ticket Prices and Attendance

One of the most interesting implications of the estimated model is that ticket prices have a surprisingly small effect on the predicted attendance. For example, the model predicts that the Montreal Expos will draw 6,659 fans per game at their current US\$9 ticket price. If ticket prices were doubled to US\$18, the predicted seats sold per game drops to 5,873 – a difference of only 785 seats. Another scenario is one where the Seattle Mariners raise their ticket prices from \$24.60 (second highest) to \$39.68 (the Boston Red Sox MLB leading 2002 ticket price). The predicted attendance drops from 42,100 to 36,300, but the \$15 increase would increase predicted yearly ticket revenue by over \$33 million. If attendance were to drop to 32,400 (one standard error lower), revenue is still increased by over \$20 million. It needs to be reiterated that the quoted figures are influenced by the TICKET coefficient, which is not significant at the 90% level.

Making these assumptions is rather simplistic and ignores other factors that go into determining the price of a ticket. For example, a team might charge more for concessions inside the stadium, so bringing in the additional fans at the lower ticket price could lead to more concessions purchases. Evidenced in the movie theater industry, there is plenty of money to be made from selling concessions at inflated prices. An important point is that this is not an empirical study of the individual ticket demand curves each team faces. The multiple regression pools all the teams together over a four year period, which makes extracting truly meaningful information about a specific team difficult.

This model provides a useful framework for forecasting attendance, but it certainly has its pitfalls and is clearly missing something. There are coefficient results that appear dubious – particularly NEWSTAD. The model does not include variables that may have an effect on attendance, such as the number of activities that compete with professional baseball or qualitative variables, such as the feeling a city has toward a team or if the city is a “baseball town.” I certainly think that including some of these might result in a better predictor of attendance, but the obvious questions are “What exactly competes with baseball” and “How are feelings measured?”

Appendix - Complete S-Plus printout for the regression and jackknife

*** Generalized Least Squares ***

Generalized least squares fit by REML

Model: attend ~ ticket + ticket.2 + seats + prev + cap + post90 + sal + wins + market.size

Data: mlb.attend

	AIC	BIC	logLik
	1968.564	1998.067	-973.2818

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	-25425.74	1656.928	-15.34511	<.0001
ticket	125.53	156.127	0.80401	0.4232
ticket.2	-7.89	3.849	-2.04885	0.0429
seats	0.47	0.022	21.27954	<.0001
prev	0.16	0.029	5.45411	<.0001
cap	40019.56	1364.734	29.32406	<.0001
newstad	2238.94	640.765	3.49417	0.0007
sal	9.70	7.645	1.26861	0.2073
wins	19.06	11.504	1.65650	0.1005
mkt.size	7.25	2.741	2.64355	0.0094

Correlation:

	(Intr)	ticket	tckt.2	seats	prev	cap	post90	sal	wins
ticket	-0.652								
ticket.2	0.602	-0.955							
seats	-0.629	0.082	-0.021						
prev	0.467	-0.348	0.341	-0.565					
cap	-0.330	-0.008	-0.053	0.612	-0.761				
newstad	0.306	-0.176	0.164	-0.311	0.410	-0.386			
sal	0.214	-0.197	0.044	-0.050	-0.123	0.054	0.067		
wins	-0.276	-0.059	0.078	-0.171	0.187	-0.297	-0.017	-0.240	
mkt.size	0.183	-0.114	-0.009	-0.241	0.026	0.079	0.121	-0.029	-0.119

Standardized residuals:

	Min	Q1	Med	Q3	Max
	-2.446542	-0.6542217	0.2492332	0.7287676	2.170384

Residual standard error: 1234.138

Degrees of freedom: 118 total; 108 residual

Note: The jackknife was used because of the large number of zeros in the new stadium column. It is very likely that a bootstrap resample would have all zeros in that column, making it a singular matrix. A jackknife, removing only 1 observation at a time, does not suffer this fate.

*** Jackknife Results ***

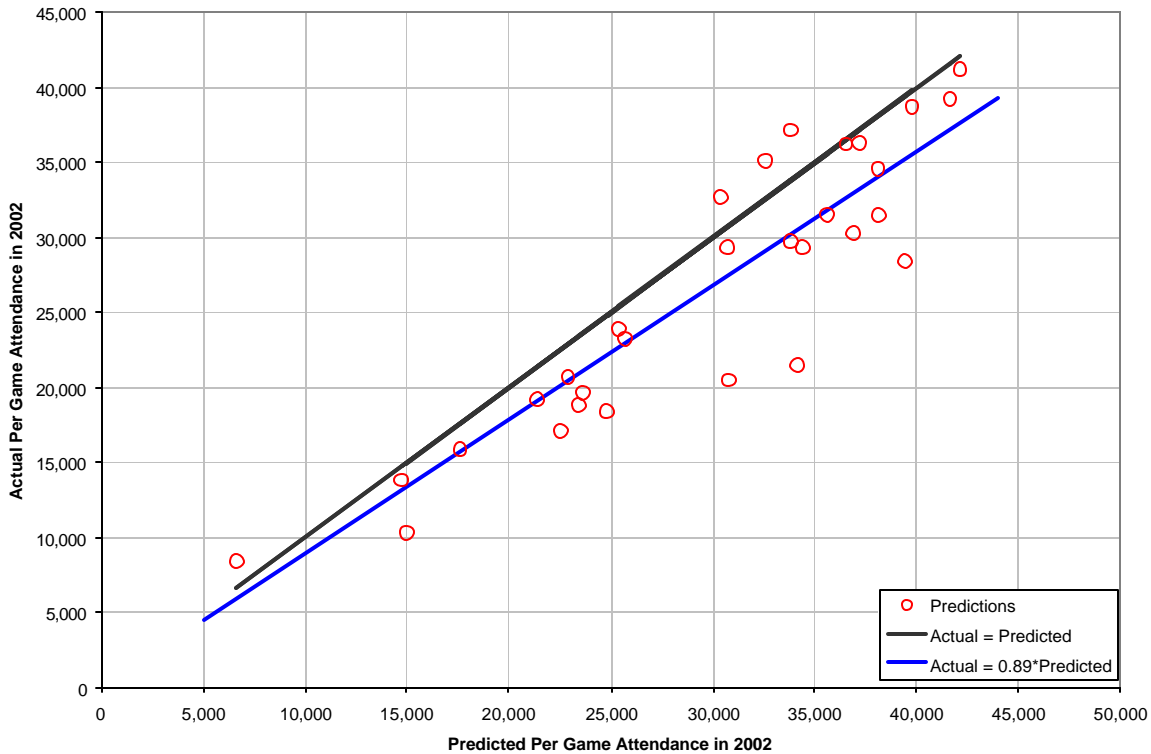
Call:

```
jackknife(data = mlb.attend, statistic = coef(gls(data = mlbbest, attend ~ ticket +
ticket.2 + seats + prev + cap + post90 + sal + wins + market.size)),
assign.frame1 = F)
```

Number of Replications: 118

Summary Statistics:

	Observed	Bias	Mean	SE
(Intercept)	-25425.7412	-2.313e+002	-25427.7178	1742.71063
ticket	125.5281	4.733e+001	125.9327	168.43828
ticket.2	-7.8861	-1.373e+000	-7.8979	4.08225
seats	0.4702	-2.068e-003	0.4702	0.02753
prev	0.1557	-3.438e-004	0.1557	0.03467
cap	40019.5569	-6.621e+001	40018.9910	1581.89130
newstad	2238.9394	-1.072e+001	2238.8477	777.80141
sal	9.6991	-4.937e-002	9.6987	6.68751
wins	19.0570	2.754e-001	19.0594	11.22739
mkt.size	7.2466	-1.320e-001	7.2454	2.48840



If the predictions were absolutely correct, the slope of the line regressing the predictions on the actual attendance should be 1.0 – indicated by the black 45° line. The actual slope of this regression is 0.89 and is shown in blue. With the slope equal to 1.0, the standard error is 3,901. With the slope at 0.89, it drops to 2,852.

Complete data was only available from 1998 forward. Additionally, I think it would be irresponsible to group attendance figures from 1994 to 1996 seasons with that from 1997 to 2002 without including a “fan sentiment” variable. I hypothesize that the 1994 players strike adversely affected attendance for most teams in 1994, 1995 and probably 1996. If this model had included a measure of fan sentiment, it would be very valuable to also include the earlier figures – since there were a lot of negative attitudes towards baseball in 1994 and 1995. The wide range of fan sentiment values would provide a more powerful regression.