Root for the Home Team?

Sierra Doherty

December 2, 2015

1 Introduction

"Root, Root, Root, for the home team. If they don't win it's a shame. 'Cause it's..." the home team that is supposed to have the advantage right?

In Major League Baseball, it would seem that home teams seem to win more often than the visiting teams. The main contributor to this is that the home team has the last at bat- the last chance to put runs on the board. But, there are other factors that could affect a team's chance of winning. Attendance, for one, can drive a team's momentum. There's also the type of stadium. Are they playing in a dome or outdoors? What's the distance from home plate to the left foul pole? Are the players used to having shorter hits?

What about factors that are involved with the game play itself? Is the team playing with a forty man roster or a twenty-five man roster? Is it an American League versus National league game, or even an AL East versus AL Central game? Some teams could do poorly playing out of their leagues. Is the team on a winning streak or a losing streak? Are they ranked high or low in their division?

So, with all of these factors, we can use logistic regression and backwards elimination to come up with a model to determine if a home-field advantage actually exists. The logistic regression is model is give by

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta_1 X_1 + \beta_2 + X_2 + ... + \beta_n X_n,$$

which will be denoted by Z_i. To find the probability of x, the model would be solved for $p(x)$ to obtain

$$p(x) = \frac{1}{1+e^{-Z}}.$$

2 Building the Model

For this model, data from the 2011-2015 Major League Baseball seasons will be used. The response variable is a binary variable taking the value of 1 if the team in question wins and 0 if the team loses. The predictor variables being considered are:

- Home- a binary variable taking the value of 1 if the game is played at home and 0 if played elsewhere.
- Sep- a binary variable taking the value of 1 if the game is played in September or October and 0 if else. This is to account for the 40 man roster being allowed for playoffs.
- C_att- a ranking of attendance within the team's respective league.
- League/Away_League- binary variables that take the value of 1 for National League and 2 for American League.

- Runs- the number of runs scored during the game by the team in question.
- Diff- the absolute value of the difference between runs scored and runs allowed.
- Rank- the ranking of the team within their division.
- Streak- the number of games a team has won or lost in a row.
- Att- the attendance at the game.
- Time- a variable that takes the value of 1 if the game is played during the day and 2 if it is played at night.
- Day- a variable that takes the value of 1 if the game is played on Sunday, 2 on Monday, 3 on Tuesday, etc.
- Old- a binary variable that takes the value of 1 if the team in question is an older team and a 0 if else. The older teams are going to be the ones that were around in the 30's, such as the Boston Red Sox, New York Yankees, and even the Los Angeles Dodgers, though they used to be the Brooklyn Dodgers (the reason for this will be explained later).

To come up with the model, there are four steps:

1. Fit all variables to the one-variable regression model

   $$Z_i = \alpha + \beta X.$$

   Since there are twelve variables being considered, there will be twelve individual models in this step. Using 0.5 as the cutoff for p-values (a higher p-value than normal is used because this is just kind of a preliminary step), any insignificant variables are eliminated from the model.

2. Fit the remaining variables to the multiple variable model

   $$Z_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

   and perform backwards elimination. This simply means take out the variable with the highest p-value and then re-fit the model. This can be done as many times as necessary.

3. Check to see if quadratic terms are needed.

4. Check to see if interaction terms are needed. A significant interaction, $X_i X_j$, means that the effect of X_i on the response is different at different values of X_j.

After going through this process, the model for this data is

$$Z_i = \underset{\alpha}{-2.689} + \underset{Home}{0.717X_1} + \underset{Runs}{0.807X_2} - 0.014X_2^2 - 0.047X_1 \times X_2 - \underset{Diff}{0.088X_3}$$

$$-0.008X_3^2 - 0.125X_1 \times X_3 - \underset{Streak}{0.056X_4} + 0.069X_1 \times X_4 + \underset{Old}{0.113X_5}.$$

3 Application

There are a couple of different uses for this model. For the first application, remember that $p(x) = 1/(1+e^{-Z})$ and consider a team at home on a one game streak. So, $X_1 = X_4 = 1$ is being held constant. We can look at different combinations of runs and run differentials to see how the probability changes.

	R=2, D=1	R=5, D=2	R=10, D=5
Older Team	0.3537	0.7588	0.9564
Newer Team	0.3283	0.7375	0.9515

Table 1: Probability that a home team on a one game streak will win.

From looking at Table 1, we can see that older teams have a slightly greater probability of winning in the given situations than newer teams. While this is interesting to look at, it doesn't give a lot of info, which leads us to why we would need to look at a second use for the model.

Instead of looking at the probability of a win, it may be more helpful to look at the odds ratio. To find the odds ratio, denoted OR, of an older team winning to a newer team winning, take e raised to the model plugging in the necessary values:

$$OR = \frac{odds_{older}}{odds_newer} = \frac{e^{\alpha+\beta_1(1)+\beta_2(5)+\ldots+\beta_9(1)(1)+\beta_{10}(1)}}{e^{\alpha+\beta_1(1)+\beta_2(5)+\ldots+\beta_9(1)(1)+\beta_{10}(0)}}.$$

Looking at the values, we can see that the model is the same for the older and newer teams with the exception of the last term. So, we can cancel everything else.

$$OR = \frac{e^{\beta_{10}(1)}}{e^{\beta_{10}(0)}},$$

and, of course any thing to the zero power is one, so we can reduce even further to obtain

$$OR = e^{\beta_{10}}.$$

From the model, the β_{10} coefficient was 0.113, so

$$OR = e^{0.113} = 1.119.$$

This means that an older team is 1.119 times more likely to win than a newer team. This is a lot more helpful than just trying to compare probability as in Table 1 above.

3.1 Home Field Advantage Through the Decades

Recall that the "old" variable is teams that were around in the thirties. So, what was the point of including this variable? The older teams are the more established teams. Then, why not include teams from later decades, like the sixties? The reason for this is that baseball was a completely different world in the thirties. A few examples of differences are that baseball was not integrated, they traveled by bus, and all games were played during the day.

So the model was refit for the 2011-15 data, throwing out the “old” variable, and then a model was fit for the 1931-35 data using the same variables for comparison. To measure strength of home-field advantage and to separate its effect from other variables, two other models for each era were also used.

1931-1935				2011-2015		
3	2	1	Models	1	2	3
-2.312	-2.587	-0.236	Intercept	-0.141	-2.738	-2.629
0.243	0.750	0.456	Home	0.271	0.857	0.720
0.624	0.625		Runs		0.808	0.808
-0.013	-0.013		Runs^2		-0.014	-0.014
0.026	0.028		Home*Runs		-0.047	-0.047
-0.036	-0.038		Diff		-0.091	-0.089
-0.006	-0.006		Diff^2		-0.008	-0.008
-0.125	-0.122		Home*Diff		-0.124	-0.125
-0.110			Str			-0.056
0.249			Home*Str			0.069

Table 2: Coefficients for the models that were constructed.

From Table 2, we can start to draw conclusions about our data. For both eras, the “Home” coefficient stays positive through all three models. Since it is positive in model 1, we can say that the home team has a greater probability of winning than the away team. Since it is positive in models 2 and 3, we can say that the probability of the home team winning increases when other factors are held constant.

Since the coefficient for “Runs” is positive in models 2 and 3, we can say that the probability of a team winning increases as the number of runs scored increases when other factors are held constant. Since the “Runs^2” coefficient is negative, this increase happens at a decreasing rate. A difference occurs between the two eras when it comes to the interaction between Home and Runs. For the thirties, the coefficient is positive, meaning the probability of winning will be greater when a team plays at home as runs increase versus when they are the visiting team. For the 2011-15 data, the coefficient is negative, so the opposite of the thirties is true. The team’s probability of winning will be greater when a team is away as runs increase versus when they are home. This does not mean that scoring more runs hurts their chance of winning when they play at home- just that runs are more important while away.

	Runs			
Era	1	2	3	4
1931-1935	1.309	1.343	1.379	1.415
2011-2015	1.96	1.87	1.784	1.702

Table 3: Odds Ratio of Home Team Winning to Away Team Winning based on Runs Scored.

Table 3 demonstrates the difference between the positive and negative "Home*Runs" interaction coefficients. With the thirties data, as runs increase, the odds ratio increase, while the odds ratio for 2011-2015 decreases when runs increase.

The coefficient for "Diff" is negative, meaning the probability will increase as the run differential decreases. As with runs, the "Diff^2" coefficient is also negative, so the decrease happens at a decreasing rate. The coefficient of the interaction between Home and Diff is negative, so The team's probability of winning will be greater when a team is away as the run differential increases versus when they are at home.

"Streak" is negative, so the probability will increase as the streak decreases. The interaction coefficient is positive, so at home, the probability will increase as streak increases. The thirties coefficient is considerably larger than the 2011-15 coefficient, so it's safe to say that streak affected baseball games in the thirties more than it does today. This is probably because teams are more evened out today. So, if a team was on a seven game losing streak and they came home to play for their next game, the probability of winning would be greater than if they were away. This is can be demonstrated with the odds ratio as well.

	Streak							
Era	1	2	3	4	5	6	7	8
1931-35	1.6	2.1	2.7	3.5	4.4	5.7	7.3	9.3
2011-15	2.2	2.4	2.5	2.7	2.9	3.1	3.3	3.6

Table 4: Odds Ratio of Home Team Winning to Away Team Winning based on Streak Length.

In Table 4, the odds ratio was calculated for a team winning based on streak length and whether they were home versus away. For both eras, the odds ratio increases as streak increases, but where the thirties odds ratio is 9.3 when streak length is 8, the 2011-15 odds ratio is only 3.6. This would read that a team is 9.3 times more likely to win at home if they were on an eight game streak than if they were away.

4 Conclusion

From the models constructed, it can be said that a home-field advantage exists, however, it seems that playing at home is not nearly as important as it was in the thirties. Further research would be of interest to see if any other variables that may not have been considered here have an effect or if there is a trend in how the home-field advantage has decreased over the decades.

References

[1] William Levernier and Anthony G. Barilla, An Analysis of the Home-Field Advantage in Major League Baseball Using Logit Models: Evidence from the 2004 and 2005 Seasons, Journal of Quantitative Analysis in Sports 3.1 (2007), Web.

[2] Baseball-Reference.com, 2015. Sports Reference, LLC. (`http://www.baseball-reference.com`).

[3] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (`http://www.R-project.org/.`)