STATISTICAL MODELS FOR STUDENT PROJECTS WITH SPORTS THEMES

Robin H. Lock
Department of Mathematics, Computer Science and Statistics, St. Lawrence University,
United States of America
rlock@stlawu.edu

We describe several types of student project assignments that involve applications of statistical models to address questions arising from sports data. Although we illustrate these ideas with examples from specific sports, our goal is to provide sufficiently general guidelines to allow instructors to adapt and extend the topics to different sports, teams, leagues or levels of play. Some of the projects are accessible to students at the introductory levels while others are more appropriate for a second course or even an undergraduate capstone/thesis. Topics include Bill James' so-called "Pythagorean law" for estimating team winning percentages, investigations of home field advantage, logistic regressions on the chance of winning a match based on boxscore statistics, the use of empirical Bayesian Stein estimators to project player performance over a full season based on early season results, and methods for modeling outcomes in seeded tournaments.

INTRODUCTION

Some students are avid sports fans and/or active participants on athletic teams. Instructors can find lots of questions that are of interest to sports enthusiasts and also serve to illustrate important concepts about how we use techniques of statistics to address practical issues. Our goal in this paper is to identify some common questions and templates of projects and activities that appeal to students with interests in sports. In most of our examples (for this paper and in class) we use data from professional sports that are popular in the United States: Major League Baseball (MLB), the National Basketball Association (NBA), the National Football League (NFL) and the National Hockey League (NHL) as well as various college/university level sports sponsored by National Collegiate Athletic Association (NCAA). Obviously, these can be adjusted to sports, teams and leagues that are more relevant to your own country and students.

HOMEFIELD ADVANTAGE

The concept of an advantage for the home team is well established among sports fans. But how big an effect is it? This question provides lots of avenues for student investigations involving inference for one or two means or proportions. For example, Figure 1 shows the difference between points scored by the home and away teams for all 256 games from the NFL's 2009 regular season. One popular rule of thumb is that home field in (American) football is worth about an extra field goal (3 points) for the home team. For this season the average margin was slightly less than that, +2.21 points with a standard deviation of 16.48 points. A 95% confidence interval for the mean size of the homefield advantage in the NFL would be between 0.2 and 4.2 points.

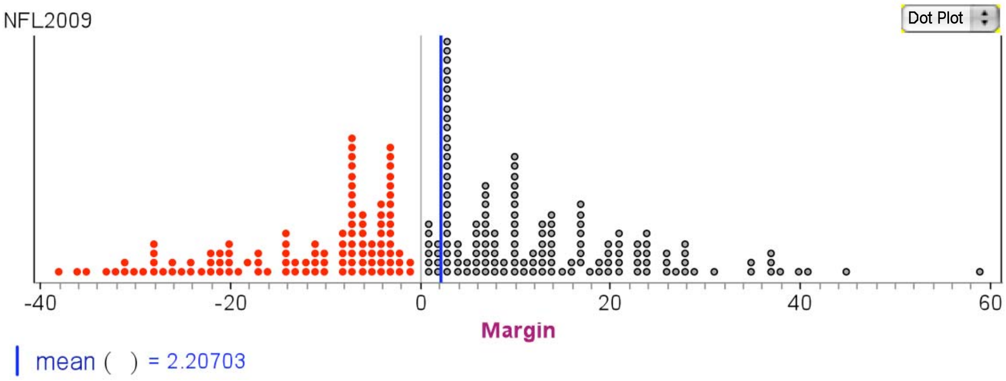


Figure 1. Homefield margins for n=256 NFL games in the 2009 regular season

Figure 1 also shows that the home team won 146 of the 256 games in the 2009 NFL regular season. Treating this as a sample of all NFL games, we would estimate the proportion of times the home team wins to be $\hat{p} = 146/256 = 0.57$ and a 95% confidence interval for the

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

proportion of home winners would go from 51% to 63%. Note that the absence of any homefield advantage would mean an average margin of zero and winning proportion of 50%, both of which lie just outside of the respective confidence bounds. Thus one might also use the data to test (as a mean or a proportion) whether a homefield advantage exists at all. For the 2009 NFL data the respective p-values for these (one tail) tests would be 0.017 for mean margin and 0.012 for the proportion of home wins - both relatively significant and consistent with each other.

Another way to explore possible advantages to playing at home is to compare team or individual performance statistics. Table 1 shows shooting results for the 2009 MVP of the NBA, LeBron James of the Cleveland Cavaliers, broken down between home and road games. Although he had a higher proportion of field goals and free throws made at home, his three point shooting was better on the road and none of these proportions is significantly different between home and away (as shown by the p-values for a two sided test for each type of shot). One curious observation is that he had quite a few more attempts in each of these categories in road games. So his average number of field goals attempted per game at home $(n_H = 40, \bar{x}_H = 17.63, s_H = 4.99)$ is significantly less (p-value<0.0001) than his mean attempts when playing away from Cleveland $(n_A = 41, \bar{x}_A = 22.15, s_A = 4.63)$.

Table 1. Home vs. road shooting proportions for LeBron James in 2008-2009

	Field Goals		3 Point Shots		Free Throws	
	Home	Road	Home	Road	Home	Road
Made	354	435	49	83	257	337
Attempts	705	908	148	236	323	439
Attempts	0.502	0.479	0.331	0.352	0.796	0.768
P-value	*0.36*		*0.68*		*0.36*	

The sort of analysis is the previous paragraph can be turned into a homework assignment, small project or even an in-class activity (assuming an internet connection to find the data) for sports-minded students. Have them pick a favorite team or player and statistic and test for a difference in performance at home and on the road. Fortunately the data, including the home/road splits, are often easily accessible on the web.

MODELING WINNING PERCENTAGE

Offense or defense?

One of the age-old debates among fans of many sports is relative merits of a prolific offense vs. a stingy defense. While obviously a team would like to excel in both areas, does one aspect of the game have a stronger influence on winning than the other? Figure 2 shows the relationships between goals scored (per game) and goals allowed (per game) and the standings points earned by 30 teams in the National Hockey League (NHL) during the 2008-2009 regular season. We use standings points rather than straight winning percentage since ties are common in ice hockey and the NHL uses an overtime and then sudden death shootout to award either one or two points in tie games. Although the correlations and percentages of variability explained by each predictor on its own are similar (r^2=0.58 for goals scored, r^2=0.51 for goals allowed) we might give a slight nod in favor of offense as the slightly stronger predictor of team success for these data.

Of course, most sports fans will agree that performing well on *both* offense and defense is the best way to gain a high winning percentage. If we use a multiple regression model for the 2008-2009 NHL data we obtain the following fitted model

$$\widehat{Points} = 86.1 + 30.2 GoalsFor - 28.4 GoalsAgainst$$

and R^2=91.7%, indicating that a significantly greater proportion of the variability in standings points can be explained by accounting for both offensive and defensive performances of NHL teams in the same model. But multiple regression might often be beyond the scope of an introductory statistics course. No problem – one can easily capture a similar effect creating a new predictor *Diff* as the difference between offensive and defensive scoring rates for each team and use a simple linear regression model. For the 2008-9 NHL data the prediction equation becomes

$$\widehat{Points} = 91.4 + 29.34Diff$$

and the R^2 value remains at 91.7%. Might the ratio (*GoalsFor/GoalsAgainst*) be a better predictor? It would be relatively easy for your students to check this or other possibilities.

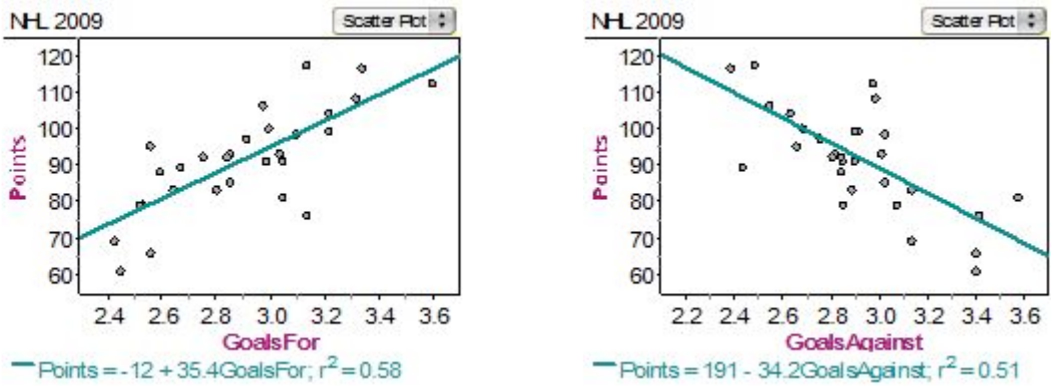


Figure 2. Predicting NHL standings points based on goals scored or allowed

"Pythagorean" Theorems

Here's another example of predicting winning percentage based on scoring/defending abilities. Bill James (1981) introduced a formula in his *Baseball Abstract* for estimating the winning percentage of a team based on its runs scored (RS) and runs allowed (RA).

$$WinPct = \frac{RS^2}{RS^2 + RA^2}$$

Given the use of squares and sums of squares this was dubbed the "Pythagorean" theorem of baseball. Later analysis showed that a better exponent for baseball was closer to 1.83 and other sports might have quite different exponents depending on the nature of scores in the sport. If we assume a general formula of the form

$$WinPct = \frac{RS^\alpha}{RS^\alpha + RA^\alpha} = \frac{1}{1 + \left(\frac{RA}{RS}\right)^\alpha}$$

the question of interest is how do we estimate a value for the exponent, α? Although this is a non-linear relationship, it is not difficult for a student with a spreadsheet to start with data on winning percentage along with points scored and allowed (either average per game or total for a season), then do a least squares analysis by "brute force". They can compute a column of "predicted" winning percentages using the formula above for a specific value of α stored in some cell, put the difference from the actual winning percentages in another column and add a formula to compute the sum of squared errors. Once this is all set up in a spreadsheet, it doesn't take long to do a "trial-and-error" search to find an optimal exponent. For example, using data for the 2008-2009 regular season for the 30 teams in the NBA, an exponent of $\hat{\alpha} = 15.72$ minimizes the sum of squared errors between estimated and actual winning percentages at 0.0134. For the 2008-2009 NHL data of the earlier example $\hat{\alpha} = 2.44$.

WHO SHOULD WIN?

Binary Logistic Regression

A project for a second statistics courses asks students to find their own data to use for fitting a binary logistic regression model. Once again, sports are a popular choice and a standard template for a project is to use information from a sample of boxscores from a favorite sport to model the probability of winning a contest (avoiding the obvious predictors of the amount each team scores!). For example, how is the probability of winning a baseball game related to the number of hits a team gets in the game? Figure 3 shows a fitted logistic regression curve for the probability of a win based on the number of hits using boxscores from each of the 2009 World

Series Champion New York Yankees regular season games. The actual game results (0=loss, 1=win) are shown with jittering on the plot. The equation of the estimated logistic curve is

$$\hat{\pi}(Hits) = \frac{e^{-1.51+0.218\cdot Hits}}{1+e^{-1.51+0.218\cdot Hits}}.$$

which produces an odds ratio of $e^{0.218} = 1.24$. Thus we see roughly a 24% improvement in the odds of winning a game for every extra hit the Yankees get.

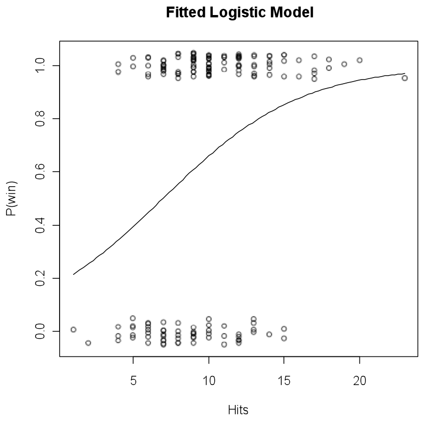


Figure 3. Logistic curve for Yankee wins based on hits

Seeded Tournaments

Many elimination tournaments (e.g. Grand Slam Tennis, World Matchplay Golf, many NCAA Championships) are organized with a seeding system where the entrants (players or teams) are ranked by perceived ability and a draw pits the top seeds against the lowest seeds in the early rounds. Assuming the seeding is an accurate reflection of relative ability we should be able to estimate a probability of either team/player winning a particular match, based on the relative seeds.

For example, the NCAA's Men's Basketball Tournament (dubbed "March Madness") generates much interest among fans (including many college students) who attempt to forecast the bracket, i.e. predict the results for all 64 teams in the tournament. The teams are separated into four regions with teams being seeded 1 to 16 within each region. The number #1 (top) seed plays the #16 (lowest) seed in the first game of each region. In the 24 years since this format was instituted the #16 seed has yet to win a game (although some have come very close to an upset). Thus the NCAA tournament might appear to be fairly predictable, but most fans will tell you that the predictability wanes as other seeds are compared. So suppose that Team A seeded #i plays Team B seeded #j–how might we estimate the probability that Team A wins based only on the seeds? Here's one easy method

$$P(\#i \quad seed \quad beats \quad \#j \quad seed) = \frac{j}{i+j}$$

but we might prefer to use results from past tournament play to obtain more accurate estimates. Berry (2000) provides such estimates based on NCAA Men's Basketball tournament results from 1986-2000. A former student, Jared Fostveit (2008), developed his own method as part of a senior honors project. He assumed that the strength of teams would decrease by a factor of 0<α<1 as the seed increased, $S_{i+1} = \alpha S_i$ and that the probability of winning would be based on relative strengths

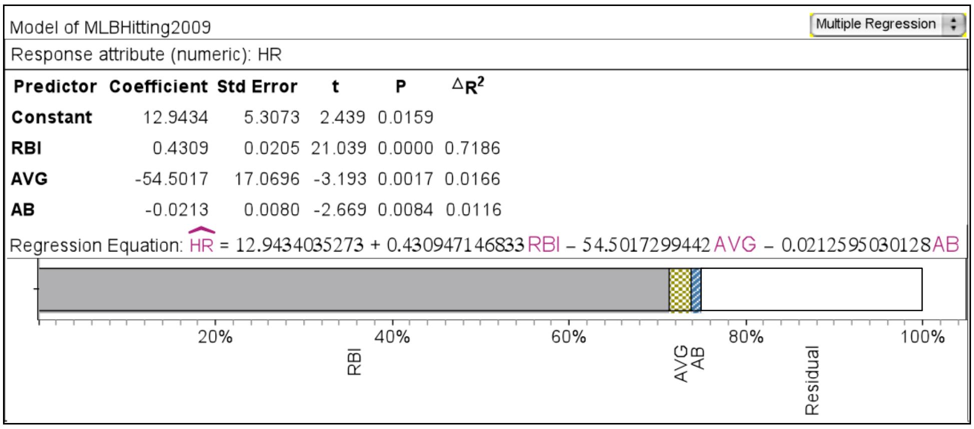
$$P(\#i \quad seed \quad beats \quad \#j \quad seed) = \frac{S_i}{S_i+S_j} = \frac{1}{1+{S_j}/{S_i}} = \frac{1}{1+\alpha^{(j-i)}}$$

By translating this into a logistic regression model with the difference in seeds as the predictor for who wins the game, Jared came up with a way to estimate α based on past tournament results. When fit to the past NCAA basketball tournaments, the estimate was $\hat{\alpha} = 0.8416$, so a #1 seed should beat the #16 seed about 93% of the time, but only have 67% chance of beating a #4 seed. He also looked at 20 years of data from the World Matchplay Golf Championships which has a similar structure of 64 players seeded into four 16-player brackets. For the golf data the estimated $\hat{\alpha} = 0.9541$, so he concluded that the NCAA March Madness was much more predictable than matchplay golf.

ASSESSING PLAYER PERFORMANCE

A Multiple Regression Project

Students looking for their own data for a multiple regression project are often drawn to sports examples. A common theme is to try to model how one statistic (e.g., home runs hit by MLB players in a season) might depend on other statistics (batting average, RBI, games played, age, weight, position,...). These projects often introduce interesting discussions, for example on the effects of multicollinearity when many of the predictors are strongly related to each other. The regression output in Figure 4 was produced from data on all (*n*=157) MLB players in the 2009 season with at least 500 plate appearances (Note: Students often restrict their "sample" to a single favorite team in which case playing time becomes a dominating factor in many models). We observe the curious fact that in this model both batting average (AVG) and number of at bats (AB) have significant *negative* coefficients in the model when runs batted in (RBI) is included. On their own AVG has little linear relationship with HR (*r*=-0.0008) and AB is mildly significant and positively correlated with HR (*r*=0.173, p-value=0.03).



Model of MLBHitting2009

Response attribute (numeric): HR

Predictor	Coefficient	Std Error	t	P	ΔR^2
Constant	12.9434	5.3073	2.439	0.0159	
RBI	0.4309	0.0205	21.039	0.0000	0.7186
AVG	-54.5017	17.0696	-3.193	0.0017	0.0166
AB	-0.0213	0.0080	-2.669	0.0084	0.0116

Regression Equation: $\widehat{HR} = 12.9434035273 + 0.430947146833\,RBI - 54.5017299442\,AVG - 0.0212595030128\,AB$

Figure 4. Multiple regression to predict home runs for MLB players

Empirical Bayes – Stein Estimators

Efron & Morris (1975, 1977) and Everson (2007) have written about the use of James-Stein estimators in an empirical Bayesian approach to parameter estimation. Each illustrates the method with a sports-based example - using early season player/team statistics (MLB player batting averages for Efron & Morris, NBA team scoring averages for Everson) to predict full season results. The gist of the method is to develop a posterior distribution for a parameter (batting average or team scoring average) based on a weighted combination of the early season performance of the individual player/team and the distribution of the other players/teams in the league. An undergraduate senior honors student, Joe Cleary (2008), adapted this method to predict save percentages for NHL goalies. He started with the observed save percentage for each regular NHL goalie over the first ten games. The basic idea of the Stein estimator is to adjust the individual save proportions from the initial games ($\hat{p}_i$) to a weighted average that takes into account the distribution of all the goalies ($\overline{p}$) over a full season (for example, the previous NHL season).

$$\hat{\theta}_i = B_i\overline{p} + (1 - B_i)\hat{p}_i$$

Thus a goalie who starts the season "hot" with several strong games will find his estimated full season proportion scaled back a bit towards the rest of the league and one who struggles in the first few outings will be helped. The weighting factor (B_i) is different for each goalie and depends on the number of shots faced and results in the early ten games. Figure 5 shows a plot of initial estimates and confidence bounds based on the Stein estimates along with traditional proportion confidence intervals and the final season save proportions for each of the goalies. The Stein estimates tended to be closer to the full season save proportions (sum of squared errors less than half the size of the errors based on sample proportions over the first ten games) and the Stein intervals are narrower but still as accurate.

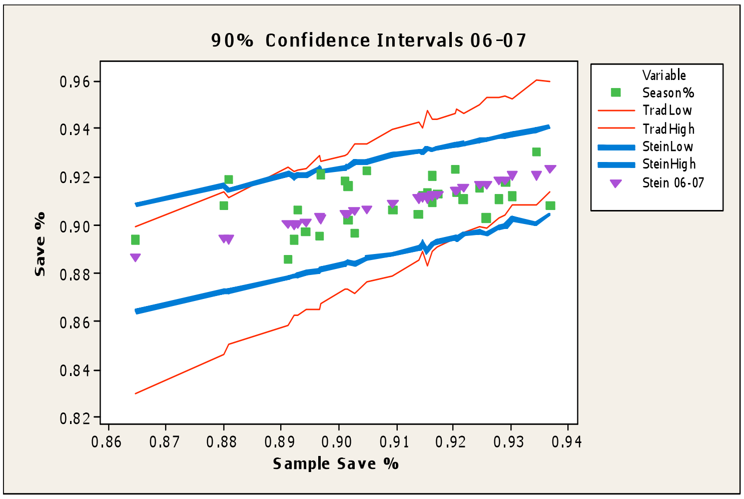


Figure 5. Stein and traditional estimates and confidence bounds for NHL save proportions

CONCLUSION

While recognizing that all students are not equally enthusiastic about sports-based examples, projects and activities, they do provide a good opportunity for motivating some students and giving them real experience in collecting data and performing statistical analyses to address questions of interest to them. The widespread availability of sports data on the web makes it feasible to carry out projects such as those outlined in this paper and many more that would be suggested by students/sports fans in these and similar areas.

REFERENCES

Berry, S. (2000) My Triple Crown. *Chance*, *13*(3).
Cleary, J. (2008). Using Stein Estimation to Predict Performance. Senior thesis.
Efron, B., & Morris, C. (1975). Data Analysis Using Stein's Estimator. *JASA*, *70*, 311-319.
Efron, B., & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific America*, *236*(5), 119-127.
Everson, P. (2007). Stein's Paradox Revisited. *Chance*, *20*(3), 49-56.
Fostveit, J. (2008). Madness of March: More Predictable than Golf. Senior thesis.
James, B. (1981). *Baseball Abstract 1981*. Self published.

WEB SOURCES

Many links related to this paper can be found at *it.stlawu.edu/~rlock/icots8sports.html*
2008-9 NBA data: *espn.go.com/nba/standings?year=2009*
2009 NFL data: *www.pro-football-reference.com/years/2009/games.htm*
2008-9 NHL data: *espn.go.com/nhl/standings?year=2009*
2009 MLB batting: *www.cbssports.com/mlb/stats/playersort/regularseason/yearly/MLB/ALL*
LeBron James' 2009 splits: *www.basketball-reference.com/players/j/jamesle01/splits/2009/*
LeBron James' 2009 games: *www.basketball-reference.com/players/j/jamesle01/gamelog/2009/*
2009 NY Yankees data: *www.baseball-reference.com/teams/tgl.cgi?team=NYY&t=b&year=2009*