

Teaching Statistics with Sports Examples

Paul H. Kvam and Joel Sokol
School of Industrial & Systems Engineering

pkvam@isye.gatech.edu
jsokol@isye.gatech.edu

1. Introduction

Modern statistics education has emphasized the application of tangible and interesting examples to motivate students learning about statistical concepts. Introductory texts aimed at special audiences (e.g., business students, epidemiology students, engineering students) can feature problems and illustrations in their field of application, complementing course material from related classes. The current textbook (Hayter, 2002) used for Georgia Tech's introductory statistics course in the School of Industrial and Systems Engineering includes a strong emphasis in science and engineering; more than half of the homework problems stress simple and illustrative examples that relate to engineering undergraduates.

So why should one consider teaching statistics using sports examples? Clearly, an introductory course that is dominated with such examples is inappropriate for students who will apply statistical methods in business, science or engineering. Most sports examples found in the literature are based on well known western sport games, the most popular being baseball. American males, who make up a majority of statistics instructors around the country, are generally familiar with such sports examples, while an increasing proportion of the students in the class have little or no experience with national sports in the United States.

In our own experiences, however, when it comes to choosing projects for various data analyses (regression, contingency tables, analysis of variance), the most popular theme, year after year, are sports related. We're surprised to find students from China or India eager to analyze attendance data for Atlanta Braves home games or goodness-of-fit tests applied to NCAA college basketball outcomes. While engineering examples have a clear purpose in teaching students in our college of engineering, sports examples seem to bring an extra level of excitement to the classroom experience.

Introductory statistics lends itself to endless applications of sports examples, especially baseball, where statistics are core to almost all aspects of player performance. Albert (2002), a professor at Bowling Green State University, outlines a basic statistics course that can be taught entirely through baseball examples alone. Simonoff (1998) focused on the home run race between Sammy Sosa and Mark McGwire during the 1998 baseball season, and utilized both introductory statistics (graphs, categorical data analysis, analysis of variance) along with more advanced methods (logistic regression, and smoothing methods) indifferent examples.

The statistics literature features several more sports examples, but in general, they are used to motivate or illustrate new and advanced methods of statistical inference, e.g., Cochran (2002), Samaniego and Watnik (1997), Harville and Smith (1994), Crowder, et al. (2002), Gill (2000). For its eight most published sports topics, the *Current Index to Statistics* (CIS) lists 230 articles that in appeared in statistics-related journals between 1960 and 2002. Figure 1 charts the frequency of the eight sports in the database; although many of the international journals in the CIS are published outside the United States, baseball still dominates the list. This is partly due to baseball's close affinity with statistics and statistical analysis, and because so much statistical information about baseball is readily available on the Internet.

Another reason American sports dominate the literature is because mostly American authors are submitting sports-related research papers to refereed journals (case in point: peruse the author list of this special issue!). The modest goal of this article is to show different ways sports examples can be used to illustrate simple statistical methods or to motivate project work in an introductory class. Examples are limited to the popular sports seen in Figure 1, notably baseball, football and basketball.

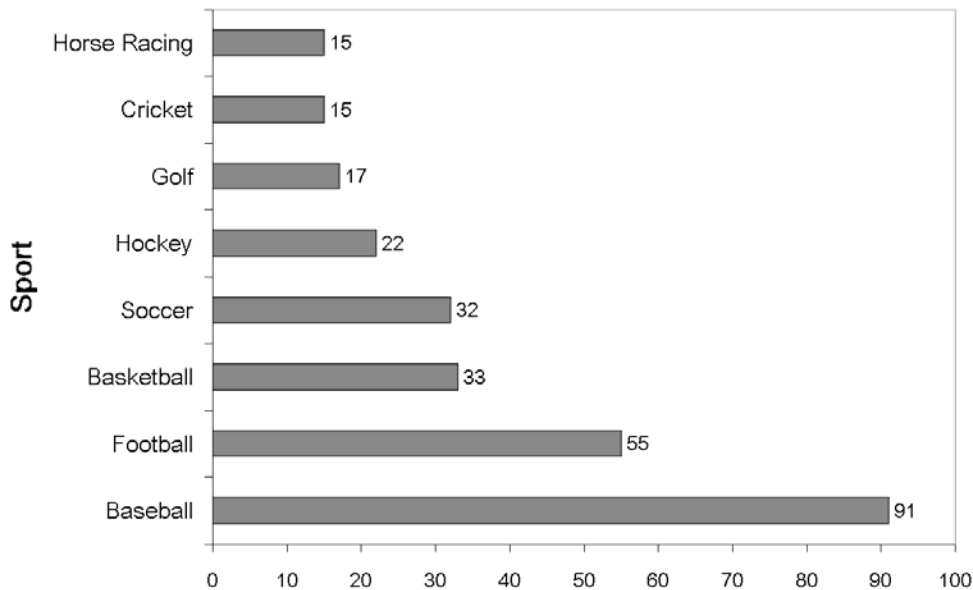


Figure 1: No. Articles in CIS

2. NBA Draft Lottery: A Tiring Exercise in Probability

For teaching elementary probability, a colorful substitute to the standard ball-and-urn examples can be found in the NBA draft-order determination held in spring before the summer draft. A lottery system started in 1985 which prevented the team with the worst record from automatically receiving the first pick, so that teams would not intentionally lose games to ensure that top draft pick. The first year proved to be memorable as the New York Knicks received the first pick (with a one in seven chance) and selected Patrick Ewing weeks later on draft day.

After a few seasons, critics pointed out that the first selection in the draft generally had not gone to the worst or even second worst team in the league. In response, the draft lottery changed in 1990 to a weighted probability system. Since then, the NBA draft lottery has provided probability and statistics instructors with non-trivial alternative to the bland ball-in-urn homework problems seen in most introductory textbooks.

In the 1990 draft, the eleven worst teams participated in the lottery and the i th “best” team (of the 11) would receive a weight of $w_i = i$. Although this change made the worst team eleven times more likely to receive the number one pick than the 11th worst team, luck came to the Orlando Magic in 1993 (the 11th worst team) when they received the first pick with the highly unlikely chance of $1/(1+2+\dots+11)=$

0.0152.

Critics again demanded a change in the system, perhaps not fully understanding the rarity of occurrence for the 1993 draft outcome, and this “catastrophic error” rate changed from 0.015 to 0.005. The prerequisite for understanding the draft lottery probabilities evolved even more in subsequent years. Fourteen numbered balls were placed in a drum, and four were chosen without replacement (14-choose-4 = 1001 ways). 1000 combinations were assigned to the 11 lottery teams, with 250 of the combinations belonging to the worst team and 5 to the best (one combination was left over; drawing it would lead to a re-drawing).

In 1995, the lottery brought in two more teams and reassigned some of the 1000 combinations, keeping 250 for the worst team and reducing the chances for the 2nd to 6th worst teams. Each augmentation provides different probability distributions for the lottery teams, and each one offers interesting insights to probability students computing and comparing the probabilities associated with lottery ranking. The NBA has posted several web pages associated with the draft lottery and the history of lottery picks and probabilities (see references for hyper-links).

3. Graphical Statistics

Graphical statistics, including bar charts (e.g., Figure 1), pie charts and histograms, represent the broadest interface between statistics and the general public. Graphical statistics are mandatory in the print media, and it is now commonplace to see a political candidate use statistical charts to support their point of view, especially in debates. Ross Perot used charts in his presidential bid in 1992. Dennis J. Kucinich, during his 2004 campaign for the Democratic presidential nomination, actually came to a National Public *Radio* debate prepared with a pie chart to argue his point about the Pentagon budget (to show the other candidates, he claimed).

Every sector of news presentation relies on charts of data summaries to communicate statistics on television, magazines and daily papers. The USA Today relies on charts to communicate anything from national trends to entertaining trivia. Occasionally, bar charts are used in the sports page. While sports examples are easily used to motivate bar charts, there are less common sports examples that show more powerfully how graphical statistics can communicate information to its audience.

3.1 Uses and Mis-uses of Graphical Statistics

Statistical lies are most frequently committed in graphical form, where the eyes can be more easily deceived by spurious trends suggested in a picture. A common infringement is manipulating scales to charts and graphs by truncating, censoring or transforming the axis values. Figure 2 below shows two different charts showing an increase in average attendance in NCAA Women’s Soccer between 1998 and 2003. The (blue) chart on the right is the default Excel chart; many statistical software packages, in fact, will restrict both axes to the smallest set of values that contains the data, which helps the reader focus on chart differences more clearly. However, it also removes the scale of difference from the picture, which has potential to mislead readers who pay little attention to the axes labels.

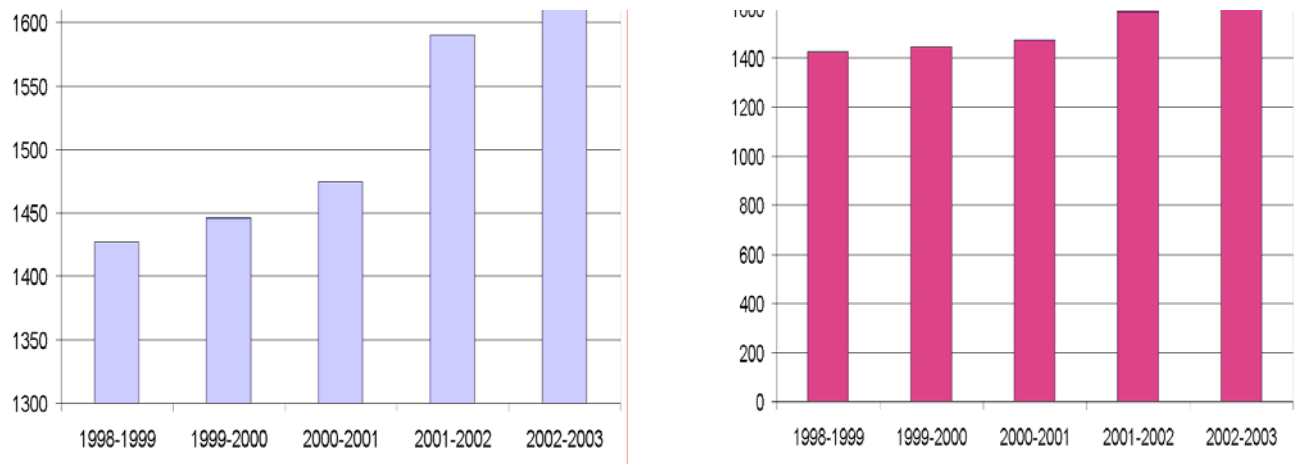


Figure 2: Two different charts showing average attendance at NCAA Women's Soccer (season) matches

The reader's sense of proportion can be manipulated further with image-based charts, which are standard in publications such as USA Today. As an example, Figure 3 below graphs the season wins for the New England Patriots using clip-art in place of the standard bar in the chart. While the height of the football icons correspond to the information the graph is meant to communicate, the *size* of the footballs do not; the Patriots improved 56% in wins between 2002 and 2003, but the increase in *area* of the football icons is over 150%.

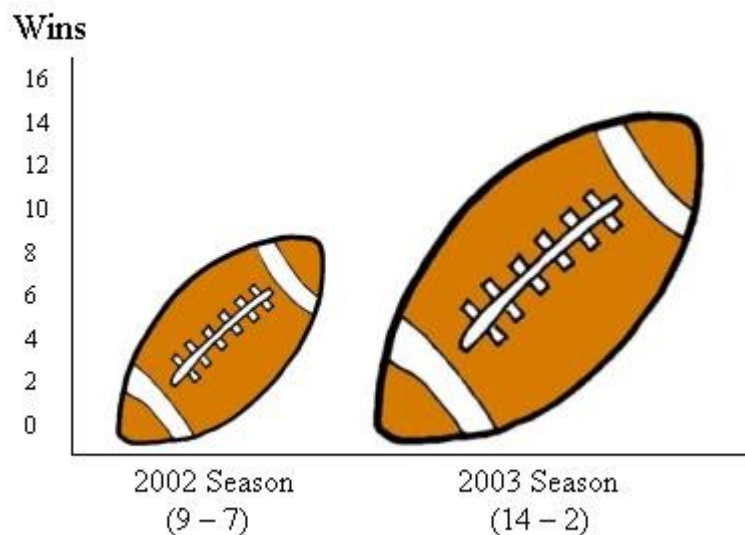


Figure 3: Regular season wins for the New England Patriots, 2002-2003.

3.2 Boxplots

Sports provide numerous examples for illustrating statistics with pie charts, scatter plots, Pareto charts,

Bubble charts, Surface plots and Box plots. Below is an example of how a box plot can summarize salary differences in Major League Baseball for the 2003 season. In this case, outlying data points (Alex Rodriguez – Texas, Carlos Delgado – Toronto) draw attention away from the bars, and a plot without plotted outliers (an option in most statistical packages) can show more with respect to team salary quartiles.

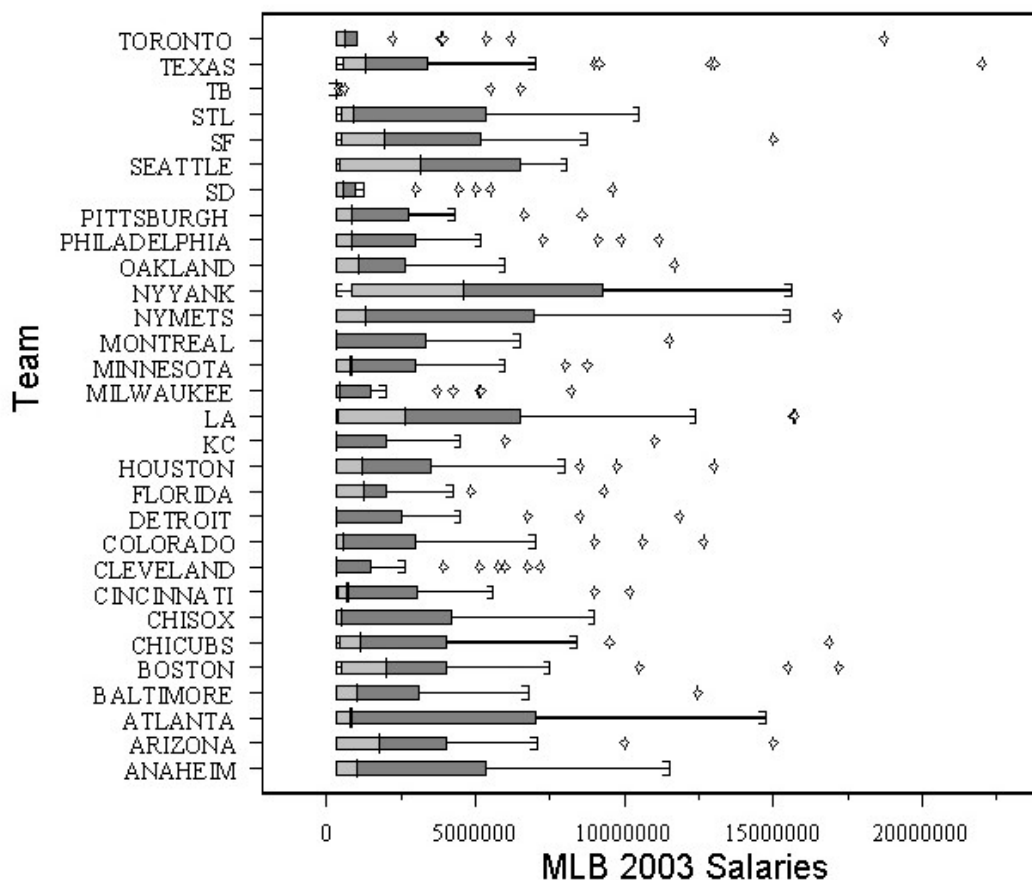


Figure 4: Box plot for player salaries of MLB teams in 2003.

3.3 Graphical Summary for Basketball Games

Innovative plots have been developed for special sets of data. Westfall (1990) presented a simple, unique way of illustrating a revealing summary of a basketball game with a chart showing the point difference between the two teams graphed across time (e.g., 48 minutes for an NBA game or 40 minutes for an NCAA game). In basketball, perhaps more than any other of the mainstream American sports, the outcome is difficult to summarize in a simple box score. Figure 5 below shows the summary of the February 1, 2004 NBA game between the Minnesota Timberwolves and the Philadelphia 76ers. Minnesota won the game 106-101. The box score, shown in Table 1, fails to summarize what happened in the game: Minnesota overcame an 18-point deficit and pulled ahead for the first time late in the game.

Students can learn about the power of graphical statistics through such novel uses of charts.

	1	2	3	4	Total
76ers	36	27	17	21	101
Timberwolves	23	29	27	27	106

Table 1: Box score for NBA game between Minnesota and Philadelphia, 2/1/2004

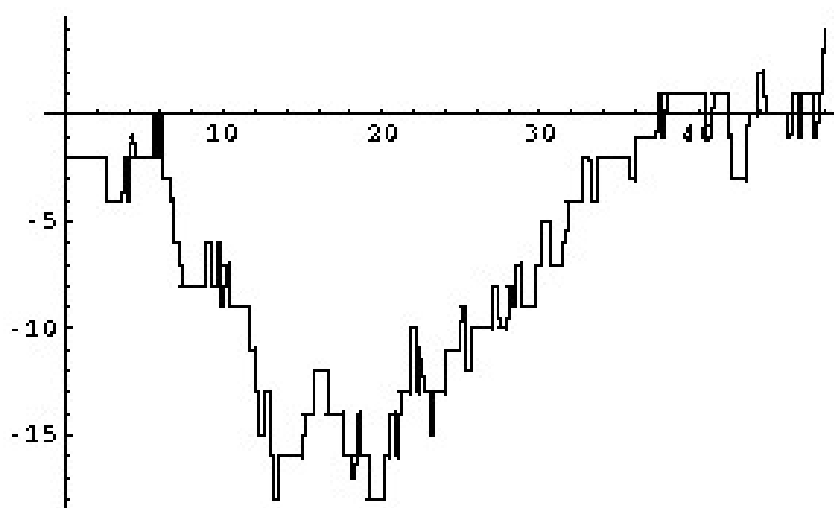


Figure 5: Point difference in NBA game between Minnesota and Philadelphia, 2/1/2004

4. Teaching Simpson's Paradox with Sports Statistics

Simpson's paradox occurs with categorical data that has three variables when an association between two of the variables is consistent across all of the levels of the third variable, but is completely different if the third variable is eliminated. It essentially shows that this association is spurious when averaged over the other factor.

The paradox is best described using a pair of two-by-two contingency tables, and baseball presents many examples of Simpson's paradox. The three variables, each at two levels, are player (two batters), batting outcome (hit or out), and batting situation (runners in scoring position or not). Table 2 below shows one of 56 pairings in which this paradox took place in the 2003 MLB season. It shows how Dustan Mohr (Minnesota Twins and San Francisco Giants) had a higher batting average (hits per at-bat) than Darin Erstad (Anaheim Angels) in both batting situations when examined separately, but overall Erstad had a

higher batting average than Mohr. The key to the paradox, of course, is that the proportions being compared are based on different sample sizes. In this case, Erstad appeared with runners in scoring position a smaller proportion of the time (20%) than did Mohr (28%). Other pairings that illustrate Simpson's paradox include Carl Everett vs. Hideki Matsui, Jose Reyes vs. Carlos Beltran, and Frank Thomas vs. Josh Phelps.

	Runners in Scoring Position		No Runners in Scoring Pos.		Overall	
	Mohr	Erstad	Mohr	Erstad	Mohr	Erstad
Hits	19	9	68	56	87	65
At Bats	97	50	251	208	348	258
PCT	0.196	0.180	0.271	0.269	0.250	0.252

Table 2: Simpson's Paradox in MLB batting averages

5. Regression Analysis

Student projects involving large sets of real data are a center point to effective statistics classes. Projects are ideal for teaching linear regression because students achieve the highest amount of freedom to select their own models to characterize the relationship between the response and the regressors.

One of the richest examples we have found for use in a statistics class is the problem of modeling a baseball player's value based on their individual statistics. For each player, batter or pitcher, there are dozens of potential regressor statistics to consider in the model; the *Microsoft Excel Worksheet* below contains the 2003 MLB batting statistics for 336 major league batters has 23 basic statistics listed (more refined databases have many more statistics to consider). We used a "fantasy league value" as the response of interest. This fantasy league value, from The Sporting News 2003 Fantasy Players Guide, is related to player performance via statistics such as hits, RBI, runs, homeruns, stolen bases, but the functional link cannot easily be characterized in a linear or nonlinear regression because many other variables influence the response. Other variables that influence fantasy value are age, team, position, injury history, and consensus findings from scouting reports.



Microsoft Excel
Worksheet

2003 MLB Batting Statistics

Students usually work in pairs, and with so many possible regressions, it is possible that no two groups arrive at the same model. As instructors, we could not help but notice that students who knew the most about baseball did not derive the best fitting model. Often, a pair of students knowing little about the nuances of the game would garner the best model relying entirely on empirical results of the data to guide their model selection. Baseball fans, on the other hand, tended to interject regressors they subjectively preferred but were not optimal variables to add into the regression model. More advanced students can consider categorical (or nominal) inputs (e.g., player's team) to form general linear models, regression diagnostics, and variable transformations to improve model fit.

6. Logistic Regression Analysis

Examples from sports can also be used to teach more advanced regression models such as logistic regression. Examples for logistic regressions are usually limited to biostatistics and other life sciences, but the following example, which examines the effects of home court advantages in college basketball, shows how sports can be used to provide students with new insights into a familiar problem.

Many NCAA basketball conferences play full or partial home-and-home round-robin schedules, so that the conference teams play each other twice during the season, once at each school. Using data collected from the 1999-2000 season through the 2002-2003 season, we seek to answer the question “Given that team A beat team B at home (or on the road) by X points, how likely are they to win the return match on the road (or at home)?”

College students, especially those at a school like Georgia Tech with a big-time basketball program, often give a question like this much more passionate thought than it might deserve (especially when asked close to NCAA tournament selection time), so it might make capturing their attention an easier task. However, answering the question might not be as easy as they would expect, because the model is more complex than they first imagine – in addition to modeling binomial data by linking the success probability to the observed point difference, students observe grossly unequal sample sizes; that is, there are very few observations of extreme cases because few teams ever win or lose a game by more than 40 points. Figure 6 shows the observed probability of winning a road game given the previously observed point spread in the home game (blue bars) along with the estimated probability based on the logistic regression model (white bars) with

$$P(\text{Win} \mid \text{point spread} = x) = \frac{e^{-(ax+b)}}{1 + e^{-(ax+b)}}$$

where (a,b) are estimated as $(-0.6228, 0.0292)$ with standard errors $(0.0231, 0.0017)$. Figure 7 charts the number of observations collected at the respective point spreads.

Figure 6: Observed win probabilities (blue) and logistic regression estimates (white) for home games at a given point spread

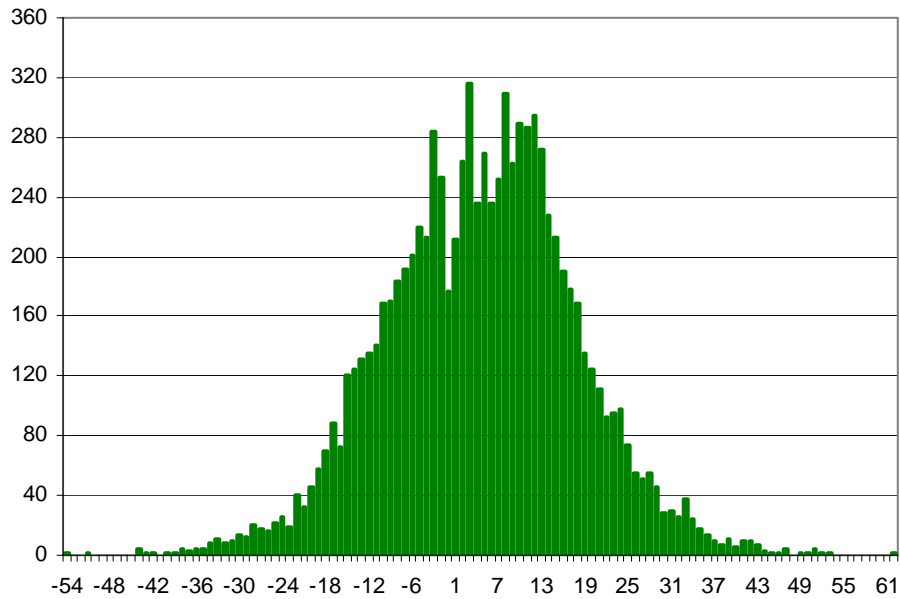


Figure 7: Number of games at various point spreads

A benefit of using this example to teach statistics is that, in addition to learning more about statistics, students also see how properly applied statistical methods can give the sports fan a new understanding of an old problem. In this case, they can see for themselves that home-court advantage, usually valued at 3-5 points (see, for example, Sagarin (2004)), the true value is probably more on the order of 10-15 points. The moral of the story? In addition to having learned (and gained an appreciation for) statistical methods, they now know it's worthwhile walking across campus to the basketball arena when that "unbeatable" opponent comes to play; with a 15-point advantage, who knows what might happen!

7. Regression Vs. Linear Programming

Statistical regression methods are also often used to obtain relative ratings of sports teams. In statistics classes (and in optimization classes), power-rating examples from sports can help teach students this use of regression as well. For example, in college football many conferences are too large for full round-robin play. The conference winner is still determined by won-lost record within the conference, but some teams play more difficult schedules than others. In the 1999 Big Ten example below, for example, students might wonder whether Wisconsin's easier schedule led to their finishing with a better record than Michigan and/or Michigan State.

	Wisconsin	Michigan	Michigan State	Penn State	Minnesota	Illinois	Purdue
Loser →							
Winner							

Wisconsin	---	---	40-10	---	20-17	---	28-21
Michigan	21-16	---	---	31-27	---	---	38-12
Michigan State	---	34-31	---	35-28	---	27-10	---
Penn State	---	---	---	---	---	27-7	31-25
Minnesota	---	---	---	24-23	---	37-7	---
Illinois	---	35-29	---	---	---	---	---
Purdue	---	---	52-28	---	33-28	---	---
Ohio State	---	---	---	---	20-17	---	25-22
Indiana	---	---	---	---	---	34-31	---
Northwestern	---	---	---	---	---	---	---
Iowa	---	---	---	---	---	---	---

Table 3. Results of play in the Big Ten Conference, 1999 (winner's score is listed first).

In fact, this is an interesting example that makes students think about the relative benefits of different statistical models. If the power ratings are defined using a linear programming approach (where the error in a prediction is defined as its absolute difference from the observation), Michigan and Michigan State are much closer to Wisconsin. On the other hand, if the power ratings are defined using a linear regression with the error defined as the squared difference, then Wisconsin has a much larger advantage.

Team (Record)	Linear Programming Power Rating	Team (Record)	Regression Power Rating
Wisconsin (6-2)	17.2	Wisconsin (7-1)	19.6
Michigan State (6-2)	14.2	Michigan (6-2)	9.4
Michigan (6-2)	11.2	Michigan State (6-2)	8.2
Minnesota (5-3)	8.2	Penn State (5-3)	7.8
Penn State (5-3)	7.2	Minnesota (5-3)	5.5
Purdue (4-4)	1.2	Purdue (4-4)	2.6
Ohio State (3-5)	-1.8	Illinois (4-4)	-2.4
Illinois (4-4)	-2.8	Ohio State (3-5)	-2.6
Indiana (3-5)	-13.8	Indiana (3-5)	-10.2
Northwestern (1-7)	-19.8	Northwestern (1-7)	-17.0
Iowa (0-8)	-20.8	Iowa (0-8)	-19.9

Table 4. Power ratings calculated using two simple regression models.

8. References

- Albert, J. (2002), "A Baseball Statistics Course," *Journal of Statistics Education*, Volume 10, Number 2.
- Cochran, J. (2002), "Data Management, Exploratory Data Analysis, and Regression Analysis with 1969-

2000 Major League Baseball Attendance,” *Journal of Statistics Education*, Volume 10, Number 2.

Crowder, M., Dixon, M., Ledford, A. and Robinson, M. (2002) “Dynamic modelling and prediction of English league football matches for betting”. *The Statistician*, Volume 51, No. 2, 157 -- 168.

Gill, P.S. (2000), “Late-game reversals in professional basketball, hockey and football”. *The American Statistician*, Volume 54, No. 2, 94 -- 99.

Harville, D. A. and Smith, M. H. (1994), “The home court advantage: how large is it, and does it vary from team to team?” *The American Statistician*, Volume 48, No. 1, 22 -- 28.

National Basketball Association (2003), Evolution of the Draft Lottery.
http://www.nba.com/history/draft_evolution.html

National Basketball Association (2003), Year by Year Lottery Picks.
http://www.nba.com/history/lottery_picks.html

National Basketball Association (2003), Year by Year Lottery Probabilities.
http://www.nba.com/history/lottery_probabilities.html

Sagarin, J. (2004), Jeff Sagarin NCAA Basketball Ratings.
<http://www.usatoday.com/sports/sagarin/bkt0304.htm>

Samaniego, F.J. and Watnik, M.R. (1997), “The separation principle in linear regression” *Journal of Statistics Education*, Volume 5, Number 3.

Simonoff, J.S. (1998), “Move Over, Roger Maris: Breaking Baseball's Most Famous Record ,” *Journal of Statistics Education*, Volume 6, Number 3.

Westfall, P. H. (1990) “Graphical presentation of a basketball game”. *The American Statistician*, Volume 44, No. 1, 35 -- 38.