# *Moneyball* in the Classroom
## Using Baseball to Teach Statistics

NCTM Annual Conference
Denver, CO
April, 2013

Josh Tabor
Canyon del Oro High School
joshtabor@hotmail.com

**Description**
As illustrated in the movie *Moneyball*, understanding the power of statistical analysis can be very rewarding. Using a formula from the movie, we will learn how to make predictions, calculate residuals, and develop the concept of least-squares. We will also use activities to explore regression to the mean and "least-squares regression" lines.

**Objectives**
At the end of the session, participants will:

- Obtain several classroom-tested examples that promote the real-world applications of mathematics and help students meet the Common Core State Standards
- Understand that the goal of a model should be minimize the size of prediction errors
- Understand the properties of least-squares regression lines and how to interpret the slope and intercept
- Understand the concept of regression to the mean and what it reveals about future performances

**Focus on Math**
This session is intended to help teachers address the Common Core State Standards which require students to "represent data on two quantitative variables on a scatterplot; fit a function to the data; and interpret the slope and the intercept of a linear model in the context of the data" (S-ID.6, S-ID.7). In addition, participants will learn how to use activities to develop the concepts of least squares and regression to the mean with their students.

Materials adapted from *Statistical Reasoning in Sports* by Josh Tabor and Christine Franklin, published by W.H. Freeman. For electronic copies of materials, including Fathom files, email me at joshtabor@hotmail.com.

**Part 1: Pythagorean Winning Percentage**

Bill James, one of the leading figures in sabermetrics, proposed that a team's winning percentage could be well modeled by the following formula, where *RS* = runs scored and *RA* = runs allowed. He called it the "Pythagorean" winning percentage formula because the denominator reminded him of the Pythagorean theorem.

$$\text{Predicted winning percentage} = \frac{RS^2}{RS^2 + RA^2}$$

How does it work? Why did he use an exponent of 2? Let's find out using data from the 2012 Major League Baseball season.

| Team | Runs Scored | Runs Allowed | Wins |
|---|---|---|---|
| ARI | 734 | 688 | 81 |
| ATL | 700 | 600 | 94 |
| BAL | 712 | 705 | 93 |
| BOS | 734 | 806 | 69 |
| CHC | 613 | 759 | 61 |
| CHW | 748 | 676 | 85 |
| CIN | 669 | 588 | 97 |
| CLE | 667 | 845 | 68 |
| COL | 758 | 890 | 64 |
| DET | 726 | 670 | 88 |
| HOU | 583 | 794 | 55 |
| KCR | 676 | 746 | 72 |
| LAA | 767 | 699 | 89 |
| LAD | 637 | 597 | 86 |
| MIA | 609 | 724 | 69 |
| MIL | 776 | 733 | 83 |
| MIN | 701 | 832 | 66 |
| NYM | 650 | 709 | 74 |
| NYY | 804 | 668 | 95 |
| OAK | 713 | 614 | 94 |
| PHI | 684 | 680 | 81 |
| PIT | 651 | 674 | 79 |
| SDP | 651 | 710 | 76 |
| SEA | 619 | 651 | 75 |
| SFG | 718 | 649 | 94 |
| STL | 765 | 648 | 88 |
| TBR | 697 | 577 | 90 |
| TEX | 808 | 707 | 93 |
| TOR | 716 | 784 | 73 |
| WSN | 731 | 594 | 98 |

**Part 2: Modeling Runs Scored**

Knowing how to predict winning percentage using runs scored and runs allowed is great. But, how can we predict runs scored? Let's look at more data from 2012.

| Team | Runs scored | Hits | Home runs | On-base | Slugging | OPS (On-base Plus Slugging) |
|---|---|---|---|---|---|---|
| ARI | 734 | 1416 | 165 | 0.328 | 0.418 | 0.746 |
| ATL | 700 | 1341 | 149 | 0.320 | 0.389 | 0.709 |
| BAL | 712 | 1375 | 214 | 0.311 | 0.417 | 0.728 |
| BOS | 734 | 1459 | 165 | 0.315 | 0.415 | 0.730 |
| CHC | 613 | 1297 | 137 | 0.302 | 0.378 | 0.680 |
| CHW | 748 | 1409 | 211 | 0.318 | 0.422 | 0.740 |
| CIN | 669 | 1377 | 172 | 0.315 | 0.411 | 0.726 |
| CLE | 667 | 1385 | 136 | 0.324 | 0.381 | 0.705 |
| COL | 758 | 1526 | 166 | 0.330 | 0.436 | 0.766 |
| DET | 726 | 1467 | 163 | 0.335 | 0.422 | 0.757 |
| HOU | 583 | 1276 | 146 | 0.302 | 0.371 | 0.673 |
| KCR | 676 | 1492 | 131 | 0.317 | 0.400 | 0.716 |
| LAA | 767 | 1518 | 187 | 0.332 | 0.433 | 0.764 |
| LAD | 637 | 1369 | 116 | 0.317 | 0.374 | 0.690 |
| MIA | 609 | 1327 | 137 | 0.308 | 0.382 | 0.690 |
| MIL | 776 | 1442 | 202 | 0.325 | 0.437 | 0.762 |
| MIN | 701 | 1448 | 131 | 0.325 | 0.390 | 0.715 |
| NYM | 650 | 1357 | 139 | 0.316 | 0.386 | 0.701 |
| NYY | 804 | 1462 | 245 | 0.337 | 0.453 | 0.790 |
| OAK | 713 | 1315 | 195 | 0.310 | 0.404 | 0.714 |
| PHI | 684 | 1414 | 158 | 0.317 | 0.400 | 0.716 |
| PIT | 651 | 1313 | 170 | 0.304 | 0.395 | 0.699 |
| SDP | 651 | 1339 | 121 | 0.319 | 0.380 | 0.699 |
| SEA | 619 | 1285 | 149 | 0.296 | 0.369 | 0.665 |
| SFG | 718 | 1495 | 103 | 0.327 | 0.397 | 0.724 |
| STL | 765 | 1526 | 159 | 0.338 | 0.421 | 0.759 |
| TBR | 697 | 1293 | 175 | 0.317 | 0.394 | 0.711 |
| TEX | 808 | 1526 | 200 | 0.334 | 0.446 | 0.780 |
| TOR | 716 | 1346 | 198 | 0.309 | 0.407 | 0.716 |
| WSN | 731 | 1468 | 194 | 0.322 | 0.428 | 0.750 |

**Part 3: Modeling Runs Allowed**

Modeling runs allowed is even more challenging than modeling runs scored. Fortunately, there has been much progress in the last 10 years. Here are some data from 2012:

| Team | Runs allowed | Home runs | Walks | Strikeouts | Strikeout/ Walk |
|---|---|---|---|---|---|
| ARI | 688 | 155 | 417 | 1200 | 2.88 |
| ATL | 600 | 145 | 464 | 1232 | 2.66 |
| BAL | 705 | 184 | 481 | 1177 | 2.45 |
| BOS | 806 | 190 | 529 | 1176 | 2.22 |
| CHC | 759 | 175 | 573 | 1128 | 1.97 |
| CHW | 676 | 186 | 503 | 1246 | 2.48 |
| CIN | 588 | 152 | 427 | 1248 | 2.92 |
| CLE | 845 | 174 | 543 | 1086 | 2 |
| COL | 890 | 198 | 566 | 1144 | 2.02 |
| DET | 670 | 151 | 438 | 1318 | 3.01 |
| HOU | 794 | 173 | 540 | 1170 | 2.17 |
| KCR | 746 | 163 | 542 | 1177 | 2.17 |
| LAA | 699 | 186 | 483 | 1157 | 2.4 |
| LAD | 597 | 122 | 539 | 1276 | 2.37 |
| MIA | 724 | 133 | 495 | 1113 | 2.25 |
| MIL | 733 | 169 | 525 | 1402 | 2.67 |
| MIN | 832 | 198 | 465 | 943 | 2.03 |
| NYM | 709 | 161 | 488 | 1240 | 2.54 |
| NYY | 668 | 190 | 431 | 1318 | 3.06 |
| OAK | 614 | 147 | 462 | 1136 | 2.46 |
| PHI | 680 | 178 | 409 | 1385 | 3.39 |
| PIT | 674 | 153 | 490 | 1192 | 2.43 |
| SDP | 710 | 162 | 539 | 1205 | 2.24 |
| SEA | 651 | 166 | 449 | 1166 | 2.6 |
| SFG | 649 | 142 | 489 | 1237 | 2.53 |
| STL | 648 | 134 | 436 | 1218 | 2.79 |
| TBR | 577 | 139 | 469 | 1383 | 2.95 |
| TEX | 707 | 175 | 446 | 1286 | 2.88 |
| TOR | 784 | 204 | 574 | 1142 | 1.99 |
| WSN | 594 | 129 | 497 | 1325 | 2.67 |

**Part 4: Regression to the Mean**

*It's difficult to make predictions, especially about the future.*
–Yogi Berra

We now have a better understanding of how to model runs scored, model runs allowed, and use these values to model winning percentage. Of course, all of our "predictions" have been for values in the past. What does the concept of "regression to the mean" tell us about future performance?