Moneyball-inspired Data Analysis Project

Proposal/First Draft Due: May 16, 2012
Final Draft Due: May 28, 2012

In our discussions of Moneyball, we’ve encountered many claims concerning the relevance of various metrics by which baseball players and teams are judged. Digging beneath the surface of Michael Lewis’s exposition, we find these claims supported by a vast foundation of statistical theory, using technical jargon such as ANOVA, factor analysis, central limit theorem, correlations, and regression. In this project your goal is to investigate the topmost layers of theory and analysis that underlie a particular statistical hypothesis. By “topmost layers of theory” we mean the high-level tools that are included in all common statistical research software. You don’t have to delve into the proofs of the various theorems that go into these tools, but you must do enough research to inform yourself of their limitations. The hypothesis you research can come from Moneyball or any other quantifiable subject (sports or otherwise) that happens to interest you. Examples of possible hypotheses include:

- On-base percentage is a better predictor, compared to other available metrics, of games won by a given team throughout the baseball season.

- Pitching and fielding contribute at most ten percent of a team’s success; offense is by far the most beneficial facet of your team to invest in.

- People born in the early months of the year are overrepresented among high-performing athletes [2].

- People graduating in the middle of a recession have an average career income of up to 25% less than people who graduate in a booming economy [4].

- Children who learn to read before the age of 6 are, on average, no better off in middle age than their peers who learned to read at age 10 [1].

- Differences in academic achievement are almost entirely explained by family background, especially as it manifests in individual aptitude. Hardly any explanatory role can be attributed to differences in teacher credentials, curriculum, physical facilities, or money spent per student [3].

Your task is to investigate these claims quantitatively. Finding the original studies (if available) or similar numerical data is a good starting point. You also want to see which statistical tools are being used by the researchers. For a more rounded perspective, also read some of the critical reviews by the research community. Reading criticisms by the author’s
peers, competitors, and others who know the limitations of statistical analysis is the best way to develop your own skeptical ear for such claims.

Write a report summarizing (i) the research in support of the hypothesis, (ii) the mathematical theory that supports the chosen research methodology, and (iii) possible objections to the interpretation of the statistical results. Using 3—5 pages (double-spaced, 12-point font) of prose, tables, and equations, discuss these three aspects of your chosen hypothesis. Your work will be graded on mathematical accuracy, sophistication of language and ideas, and appropriate essay structure.

References


