DECISION SCIENCES INSTITUTE
Experiential Exercises in Applied Data Analytics: Baseball

ABSTRACT

This paper presents a series of experiential exercises using Major League Baseball data and the book, Moneyball by Michael Lewis. The exercises progress from simple descriptive analytics using scatter plots, to an evaluation of a predictive model, and ending with an application of regression analysis.

KEYWORDS: Experiential, Learning, Analytics, Sports

INTRODUCTION

The idea of this paper began with my reading of Moneyball, by Michael Lewis. Beginning in the preface and continuing throughout the book, Lewis describes the application of data processing, statistical application, and analytics to tell a story about decision-making in baseball. With the availability of Major League Baseball (MLB) data via the internet, there is an opportunity to use the book's story with the MLB data to create a series of experiential/hands-on exercises that enhance and expand the learning opportunities of an introductory statistics course. This paper summarizes these exercises.

PEDAGOGY

My process of teaching a "soft" approach to analytical methods in a first statistics course has several learning outcomes. Listing them:

- Students need to know the basic elements of data sets, ignoring the values and focusing on the variables. A data set with a large number of observations is not complex.
- Students need to know that they can use simple descriptive analytics to summarize data.
- Students need to develop the skills to connect the results of their analysis with clear, concise logical written interpretation
- Students need to experience the process of testing and evaluating alternative models using data and analytical methods.
- Students need to know about the software available to support analytics.

EXERCISES

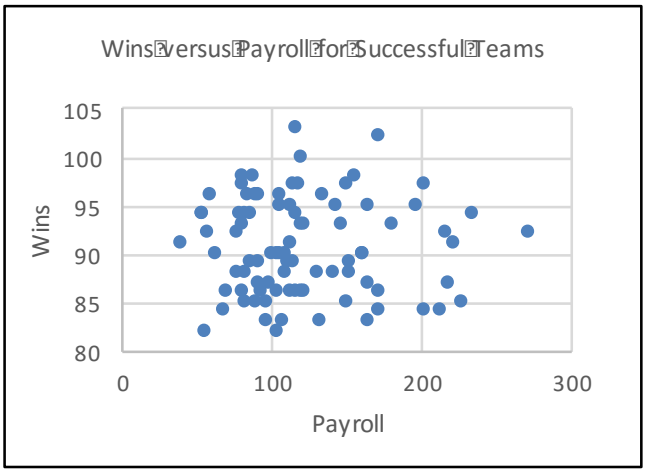
All exercises described in this paper use the baseball team in any year as the observation or unit of analysis. All data and variables referenced, except for team payroll, are accessible on the MLB website, www.mlb.com. Many hitting, pitching, and fielding variables are available going back to the 1876 season.

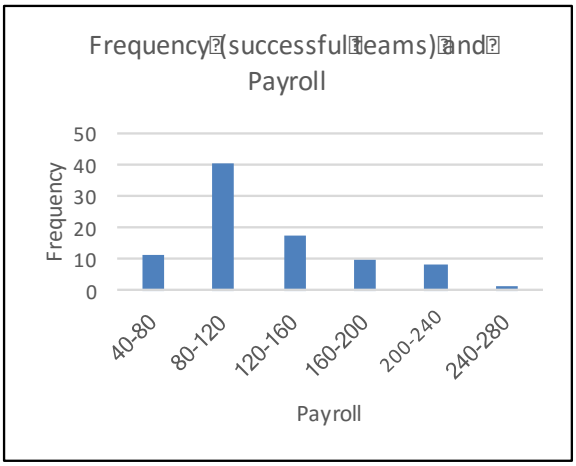
Payroll and Success in Baseball

The first in the series of exercises uses with the short preface of Moneyball and the beginning of Chapter 6, "The science of winning an unfair game." The story begins (circa 1980) with a belief or claim that MLB teams who can afford to buy the best, and most expensive, players were more likely to have a successful season. The commissioner of Major League Baseball, Bud Selig,

assembled a Blue-Ribbon Panel on Baseball Economics to investigate the claim. One of the four members was Paul Volcker, a former head of the Federal Reserve Bank. In retrospect, it is surprising that a simple, descriptive analytic summary could have shortened the controversy.

The relevant variables in this exercise are team wins and payrolls. Team payroll data is available at www.stevetheump.com. The data should include several years. The following example uses data from the 2013 through the 2016 seasons. The nuance in this exercise is defining a successful season. A successful season could be conservatively defined as winning a majority of the 162 games with a record of 82 wins and 80 losses. A class discussion motivates students to consider the definition of a successful team. There is no best decision. For example, Moneyball suggests that a team's goal is to win at least ninety games to assure play in the post-season. Students are also asked to consider a way to evaluate the presumed relationship between payroll and number of wins. With this guidance about defining a successful team and the data, students can create a scatter plot of wins versus payroll, compute a correlation between payroll and wins for successful teams, or construct a frequency table and histogram of successful teams by payroll classes. The following are the results for the selected years.

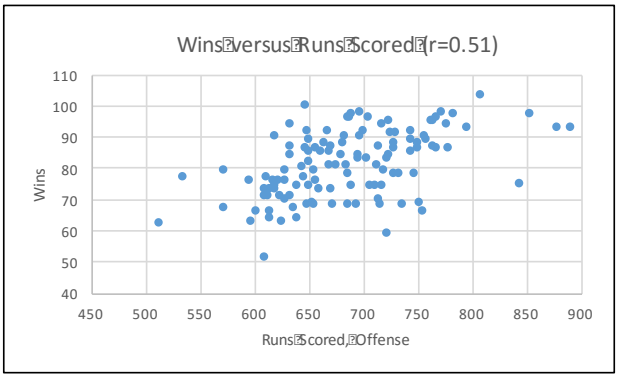


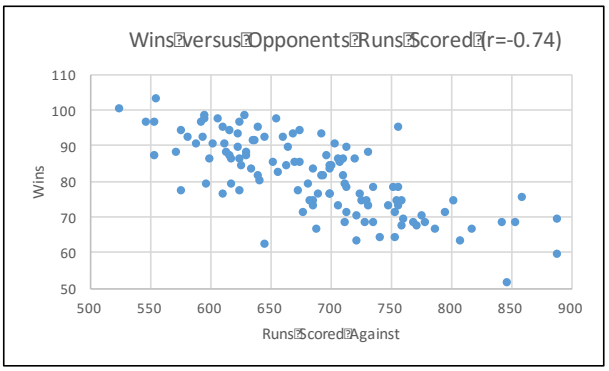


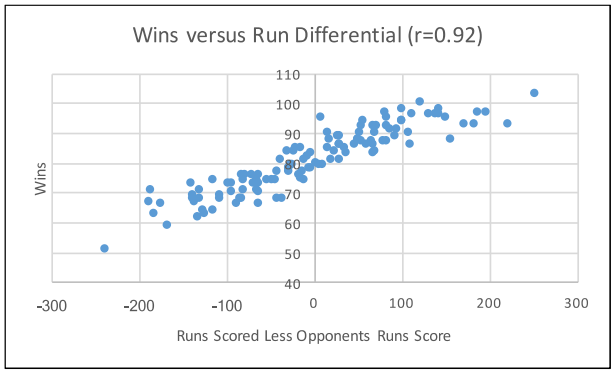
With these simple descriptive analytics, what statements can be made? The scatter plot shows no indication that successful teams are more likely to have higher payrolls. The histogram shows that most successful teams are paid in the $80 to $120 million range which is more than $100 million less than the highest payroll teams. The correlation, r = -0.019, is nearly zero. So, payroll is not related to a successful baseball season.

Getting back to basics: so, what wins games?

The result of every baseball game is that the winning team scores more runs than the losing team. In preparation for this exercise, a class discussion motivates students to evaluate the relationship between wins and runs, wins and opposing team runs, and wins and a computed run differential (runs scored – opposing runs scored). The assignment is to evaluate these relationships with scatter plots and use the plots to make statements about the relationships. Correlations can also be computed and interpreted. Examples using the data for the 2013 through 2016 seasons (n=120) are shown here.







The data confirms the nature of baseball as a competitive game. The scatter graphs and correlations show that wins are positively related to scoring more runs. However, the relationship is not strong with a modest correlation (r=.51); wins are inversely related to more opposing team runs scored with a fairly strong correlation (r= -0.74); wins are positively related to the run differential with a strong correlation (r=0.92). The interpretation of the data should be straightforward: successful baseball teams score more runs than their opponents. And, based on the results from the first exercise, a team can be successful with a modest payroll. So, what is the key to scoring runs?

Selecting baseball players

In chapter two, Lewis describes a group discussion about selecting players for the team. The group includes baseball scouts, the team's general manager (Billy Bean), and the team's newly hired sabermetrician. Chapter two describes the nature of baseball scouting and how scouts use their years of experience interviewing and watching hundreds of potential players to assess a player's talent for hitting, fielding, and pitching. The criteria used by the scouts is contrasted with the player performance statistics collected and analyzed by the sabermetrician.

Using this situation, the next exercise for students is to explore the set of variables discussed in the meeting to select players. The assignment is to list all the variables discussed by the group and then classify the variables as qualitative or quantitative. There are at least thirty variables. The follow-up class discussion reviews the list of variables and the type of variable with the additional question about the variable's reliability to assess a player's ability to score runs. Comparing the reliability of the variables should result in a preference for quantitative performance statistics over the qualitative assessments. The results of this discussion lead into

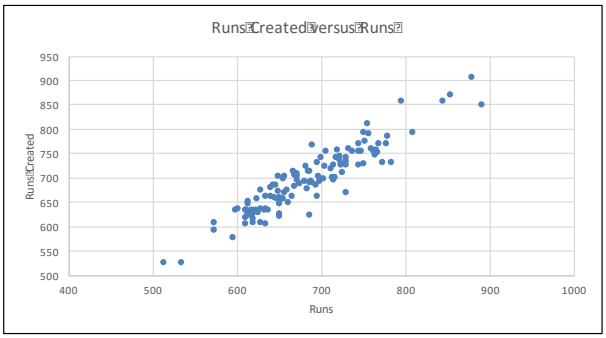
the next exercise. It focuses on creating a model using variables that are related to scoring runs.

What scores runs?

In chapter four of Moneyball, Bill James, the father of sabermetrics, is introduced. He has a fascination with baseball and a focus on using data to answer questions and test hypotheses. One of the key questions was: how are runs created? Using limited data and computing resources, he created and tested the following equation or model referenced in Moneyball:

$$Runs\ Created = \frac{(hits + walks)}{(at\ bats + walks)}(Total\ Bases)$$

The next exercise is to use the MLB data to test and validate the James model with MLB data. The assignment is to use the variables: runs, hits, walks, total bases, and at-bats) to calculate the predicted "runs created" and compare the actual season runs to the predicted season runs. As with the previous exercises, a scatter plot and simple computations of summary statistics (percent differences, means, minimum and maximum values) show the effectiveness of the equation to model how runs are created.



Comparison of Runs versus Runs Created	
Average Difference	-8.8
Average Percent Difference	-1.3%
Minimum Percent Difference	-11.1%
Maximum Percent Difference	9.4%
Correlation	0.93

The results show that the James formula is a fairly good predictor of runs. It overestimates the seasonal run total by an average of 8.8 runs or a 1.3 average percent difference. The graph shows a clear, tightly clustered association between runs and runs created. The interpretation is that the James Formula has identified several key variables associated with scoring runs: hits, walks, and total bases. Further, these three variables should be used to select players: players who get on base (hits and walks) and players who get hits for multiple bases (total bases).

Predictive analytics applied to James formula

After teaching the concepts of regression analysis, the third assignment is to replicate and extend James' analysis with regression analysis. Regression analysis provides additional information such as the R^2, standard error of the estimate, and the values of the regression coefficients. All are useful in understanding the relationships hidden in the data.

The first part of the assignment is to produce scatter plots of the three relationships implied in the James formula: runs and hits, runs and walks, and runs and total bases. Then, apply simple regression analysis to evaluate each relationship. A summary of the results follows:

Simple Linear Regression Analysis of Runs versus Hits, Walks, and Total Bases				
Independent variable	Coefficient	P-value	Adjusted R^2	Standard error
Hits	0.66	0.000	49%	47
Walks	0.46	0.000	16%	61
Total Bases	0.41	0.000	80%	30

Students should verify that each variable is significantly related to scoring runs. This is information that supports James' formula. The scatter plots and regression results also show the differences in the strength of the relationships with the R^2. This analysis shows that each of the individual independent variables is, as James thought, significantly related to scoring runs. However, each has a very different R^2 and standard error showing the differences in the ability of each variable to predict runs.

The next part of the assignment, is to use multiple regression analysis to create a model that includes all the variables from the James formula. The results follow:

Multiple Linear Regression Analysis of Runs versus Hits, Walks, and Total Bases				
Independent variable	Coefficient	P-value	Adjusted R^2	Standard error
Hits	0.14	0.005	85%	25
Walks	0.27	0.000		
Total Bases	0.33	0.000		

The results of multiple regression illustrate the ability to add variables to a linear model that will account for additional variance in runs. A comparison of the single variable models with the multiple variable model shows the combination of the three variables in a model results in a higher R^2, a lower standard error, and a better prediction of runs scored.

The final part of the assignment is to experiment with the variables to create a model that is more descriptive of how runs are scored or a model that results in a higher R^2 and lower standard error. The variables offered to students are: at-bats, runs, hits, singles, doubles, triples, home runs, walks, total bases, and strikeouts. Students are encouraged to play with different models. Sometimes the results provide opportunities to talk about autocorrelation. For example, a student may decide to model runs in relation to hits, singles, and doubles. The counterintuitive results follow:

Multiple Linear Regression Analysis of Runs versus Hits, Singles, and Doubles				
Independent variable	Coefficient	P-value	Adjusted R^2	Standard error
Hits	1.41	0.000	.81	29
Singles	-1.07	0.000		
Doubles	-0.19	0.277		

Clearly, there is something wrong. How could hits that are singles result in fewer runs? The answer is in the question and the statistical reason is in the correlations:

Correlations				
	Runs	Hits	Singles	Doubles
Runs	1.000			
Hits	0.702	1.000		
Singles	0.207	0.788	1.000	
Doubles	0.657	0.648	0.261	1.000

The correlation matrix shows very high correlations between hits and singles, and hits and doubles. The practical reason for the inverse relationship between runs and singles and doubles is that singles and doubles are hits, so these variables are correlated. And when correlated variables are included as independent variables, the results do not represent the real relationships. These correlations result in regression results that are unexpected and not supported by observation or logic. This analysis should be ignored.

A final analysis

If James had the technology, what would he have analyzed? Students are encouraged to further explore the data with multiple regression analysis. For example, what if a model to predict runs includes every type of hit and walks? The results follow:

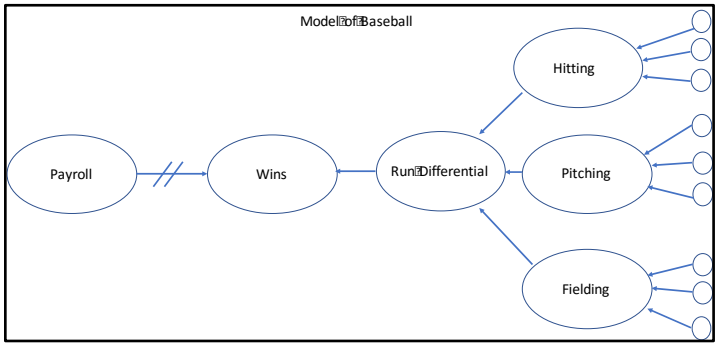
Multiple Linear Regression Analysis of Runs versus Hits, Singles, Doubles, Triples, Home Runs, and Walks				
Independent variable	Coefficient	P-value	Adjusted R^2	Standard error
Singles	0.43	0.000	86%	25
Doubles	1.01	0.000		
Triples	0.83	0.002		
Home runs	1.39	0.000		
Walks	0.24	0.000		

The results show that each way to get on base is significantly related to runs. None should be discounted. The model's R^2 and standard error do not change from the model of hits, walks, and total bases. However, this model has interesting regression coefficients. For example, for every single, 0.43 runs will be scored. So, on average, three singles are needed to score 1.29 runs. Every double results in a run. Every triple results in less than a run. Every home run produces more than a run. Every walk results in 0.24 runs; four walks result in nearly one run (0.96 runs). These interpretations are especially meaningful for students who have some knowledge of baseball.

SUMMARY AND CONCLUSION

This paper presents a pedagogy to give students exposure and experience in basic data analytics. The exercises provide access to several different data sets which focus on several different questions and hypotheses. Each exercise challenges students to summarize or analyze the data with simple, descriptive analytics and to use the results of the analysis to tell the story of "what wins baseball games?" The written interpretation of the analyses is, arguably, the most important part of the exercises because it is the translation of the analysis into a meaningful interpretation of the analysis. My experience with these exercises is that motivated students will find interest in using the data to answer questions. Most students find the context of sports interesting and have at least a working knowledge of the game of baseball.

The application of data analytics described here is clearly limited to hitting. The full model of how baseball works is illustrated. Clearly there are fielding and pitching components to winning games. Given the three components and the variety of variables offered in the MLB data, other experiential exercises and assignments can be created.



REFERENCES

Lewis, M. (2003). Moneyball: The art of winning an unfair game. New York: W.W. Norton.

http://www.mlb.com/, MLB.com, The Official Site of Major League Baseball, Accessed May 14, 2017

http://www.stevetheump.com/, Steve O's Baseball Umpire Resources, Accessed May 14, 2017