Creative Education, 2015, 6, 1477-1483
Published Online August 2015 in SciRes. http://www.scirp.org/journal/ce
http://dx.doi.org/10.4236/ce.2015.614148

Scientific Research Publishing

Using Baseball Data as a Gentle Introduction to Teaching Linear Regression

Patrick R. McMullen

School of Business, Wake Forest University, Winston-Salem, NC, USA
Email: mcmullpr@wfu.edu

Received 4 July 2015; accepted 4 August 2015; published 7 August 2015

Abstract

This effort describes a successful classroom exercise to introduce simple and multiple linear regression to working professional MBA students. The exercise starts by exploring the relationship between a baseball team's payroll with its winning percentage. The exercise then continues with the introduction of additional predictor variables so that the students are able to build a strong predictive model for winning percentage. Student feedback consistently praises the exercise as an effective way to learn about linear regression.

Keywords

Teaching, Linear Regression, Statistics, Survey, Baseball

1. Introduction

Teaching statistics to MBA students is a challenge. A part of the reason for this is that many different academic backgrounds comprise the class: liberal arts, engineering, social sciences, hard sciences, history, education and so on. With such a diverse group, learning statistics is harder for some students than others. Linear regression is no exception to this—while most students can easily understand the concept of a slope and intercept, the statistical significance of the slope and intercept, and predictive ability of the model, can be more challenging, particularly for those students who struggle with the basics of hypothesis testing in earlier class sessions.

This dilemma has been noted over the years, and because of this, a simple example has been developed for the Working Professional MBA Classes at Wake Forest University's School of Business that has eased the introduction of linear regression to the students. Instead of a typical example exploring the relationship between price and demand for some generic good, a baseball example has been used with consistent success.

The idea for this example came from an earlier edition of the classic text book by Albright, Winston and Zappe (2011), where a scatter plot was shown to illustrate the relationship between the annual payroll of a Major

How to cite this paper: McMullen, P. R. (2015). Using Baseball Data as a Gentle Introduction to Teaching Linear Regression. *Creative Education, 6,* 1477-1483. http://dx.doi.org/10.4236/ce.2015.614148

League Baseball team and the team's winning percentage for that same year. Other studies have investigated salary to individual performance (Watnik, 1988; Hoaglin & Velleman, 1995), but the author is not aware of any continuous exploration of the relationship between team success and team payroll. Over the years, this simple example has been developed to study the relationship between payroll and winning percentage in terms of the statistical significance of the slope and the model's predictive ability. The model has been further developed to include multiple predictor variables so that the model's predictive ability is maximized while simultaneously maintaining a parsimonious model.

2. Does a Baseball Team's Payroll Result in Success?

The short answer to the above question is "not really," but this question is the "teaser" for the exercise. A discussion is then encouraged in the classroom about the relationship between a baseball team's payroll and their winning percentage. The following comments from the students are typical:

- The New York Yankees typically have the highest payroll in the game, but sometimes they're good, sometimes they're not so good.
- Teams like the Pittsburgh Pirates, Tampa Bay Rays and Oakland Athletics usually perform well with small payrolls.
- Big city teams usually have a large payroll, regardless of whether or not they are successful. Teams in New York, Chicago and Los Angeles are what the students consider "big city" teams.

All of the above points are well taken. Data are then presented to them which show the team's payroll and winning percentage for the most recently completed regular season. For the 2013 regular season, the data are as follows:

From inspection of **Table 1**, the points made by the students above are essentially supported by the data. Converting the data into a scatter plot, along with the "best-fit" regression line, we have what is shown in **Figure 1**.

The points made by the students are further supported by the scatter plot. In essence, any team "above" the regression line outperforms their payroll, and any team "under" the regression line underachieves according to their payroll.

At this point, the statistical inference portion of the exercise is performed. The slope term has a t-statistic of 1.73, while the associated p-value is 0.0953. This does not indicate a strong relationship between payroll and winning percentage. Only at $\alpha = 0.10$, the most liberal level of significance commonly used, can one reject the null hypothesis of a meaningless slope. In essence, then, the relationship between payroll and winning percentage is dubious at best. This finding is quite surprising to the students. In fact, there are always students at this point in the exercise who question the New York Yankees high-payroll strategy.

To further dampen this analysis, the R^2 term is discussed, showing that only 9.62% of the variation in a team's winning percentage can be explained by their payroll. The lesson learned here is that there is a weak relationship between payroll and winning percentage, and our ability to use this model as a predictive tool is non-existent.
The next order of business is to inform the students that there may be other ways to "explain" winning percentage other than the team's payroll.

3. Multiple Linear Regression as a Tool to Explain Winning Percentage

To improve the ability to "explain" the variation in the team's winning percentage, the concept of multiple linear regression is introduced, where it is mentioned that adding new predictor variables can help to increase the R^2 term. That is, it is mentioned that one can do a better job of explaining the variation in the team's winning percentage. It is also emphasized that we must be frugal in adding predictor variables, as many predictor variables in a single model can make the model difficult to interpret. In the context of our baseball example, Payroll is removed and five new predictor variables are introduced. **Table 2** shows the new data set.

The five new predictor variables are as follows:

- ERA: Earned Run Average. The number of earned runs forfeited per nine innings pitched during the 2013 season.
- BB: Bases on balls. The number of times a player was "walked" during the 2013 season.
- RBI: the number of runs batted in during the 2013 season.
- Slug: The slugging percentage of a team during the 2013 season. Total bases divided by plate appearances.

Table 1. Payroll and winning % by team for the 2013 season (*USA Today*).

Team	Abb.	Payroll	Win %
New York Yankees	NYY	$228,995,945	52.47
Los Angeles Dodgers	LAD	$216,302,909	56.25
Philadelphia	PHI	$159,578,214	45.06
Boston	BOS	$158,967,286	59.88
Detroit	DET	$149,046,844	57.41
San Francisco	SF	$142,180,333	46.91
Los Angeles Angels	LAA	$142,165,250	48.15
Texas	TEX	$127,197,575	55.83
Chicago White Sox	CHW	$124,065,277	38.89
Toronto	TOR	$118,244,039	45.68
St. Louis	STL	$116,702,085	59.88
Washington	WAS	$112,431,770	53.09
Cincinnati	CIN	$110,565,728	55.56
Chicago Cubs	CHC	$104,150,726	40.74
Baltimore	BAL	$91,793,333	52.47
Milwaukee	MIL	$91,003,366	45.68
Arizona	ARI	$90,158,500	50.00
Atlanta	ATL	$89,288,193	59.26
New York Mets	NYM	$88,877,033	45.68
Seattle	SEA	$84,295,952	43.83
Cleveland	CLE	$82,517,300	56.79
Kansas City	KC	$80,491,725	53.09
Minnesota	MIN	$75,562,500	40.74
Colorado	COL	$75,449,071	45.68
San Diego	SD	$71,689,900	46.91
Oakland	OAK	$68,577,000	59.26
Pittsburgh	PIT	$66,289,524	58.02
Tampa Bay	TB	$57,030,272	56.44
Miami	MIA	$39,621,900	38.27
Houston	HOU	$24,328,538	31.48

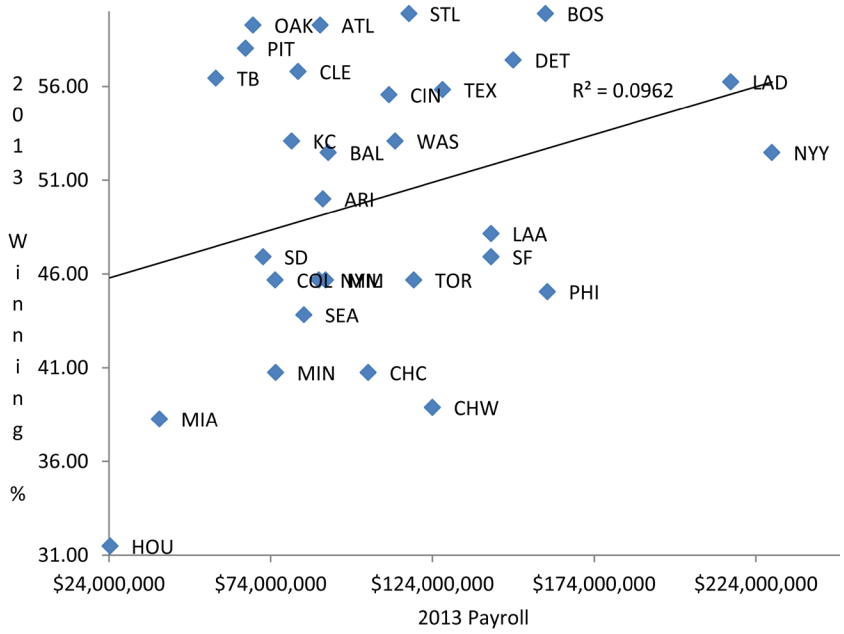

Figure 1. Payroll and winning % for 2013 season.

Table 2. Revised data set for 2013 season (*USA Today*).

Team	Win %	ERA	BB	RBI	Slug	WS Manager
NYY	52.47	3.94	466	614	0.376	1
LAD	56.25	3.25	476	618	0.396	0
PHI	45.06	4.32	417	578	0.384	1
BOS	59.88	3.79	581	819	0.446	0
DET	57.41	3.61	531	767	0.434	1
SF	46.91	4.00	469	596	0.381	2
LAA	48.15	4.23	523	696	0.414	1
TEX	55.83	3.62	462	691	0.412	0
CHW	38.89	3.98	411	574	0.378	0
TOR	45.68	4.25	510	669	0.411	0
STL	59.88	3.42	481	745	0.401	0
WAS	53.09	3.59	464	621	0.398	1
CIN	55.56	3.38	585	664	0.391	0
CHC	40.74	4.00	439	576	0.392	0
BAL	52.47	4.20	416	719	0.431	0
MIL	45.68	3.84	407	610	0.398	0
ARI	50.00	3.92	519	647	0.391	0
ATL	59.26	3.18	542	656	0.402	0
NYM	45.68	3.77	512	593	0.366	0
SEA	43.83	4.31	529	597	0.390	0
CLE	56.79	3.82	562	711	0.410	2
KC	53.09	3.45	422	620	0.379	0
MIN	40.74	4.55	533	590	0.380	0
COL	45.68	4.44	427	673	0.418	0
SD	46.91	3.98	467	578	0.378	0
OAK	59.26	3.56	573	725	0.419	0
PIT	58.02	3.26	469	603	0.396	0
TB	56.44	3.74	589	670	0.408	0
MIA	38.27	3.71	432	485	0.335	0
HOU	31.48	4.79	426	566	0.375	0

- WS Manager: the number of World Series titles won by the team's manager at the beginning of the 2013 season.

The five new predictor variables are used in a multiple linear regression model with winning percentage as the response variable. Of course, the R^2 term increases dramatically from the simple linear regression. However, the model is cumbersome due to the inclusion of non-significant predictor variables and multicollinearity—correlation between predictor variables, hindering our ability to properly interpret the significance of individual predictor variables.

After removing the predictor variables due to insignificance, and removing others to be explained by remaining variables due to multicollinearity, the following model is obtained:

$$\text{Winning \%} = f(\text{ERA}, \text{RBI})$$

Including payroll, six total predictor variables were candidates to explain winning percentage, and only two are needed to explain winning percentage: ERA and RBI. The slope terms show the following statistical properties (**Table 3**):

This model results in an R^2 term of 0.8830% - 88.30% of the variation in winning percentage is explained by ERA and RBI.

The astute students will usually note that the two remaining predictor variables make total sense. One is an offensive statistic—RBI is a measure of run-production, while the other is a defensive statistic—ERA is a pitcher's ability to prevent the opposition from scoring runs.

The cynical student will also note that the outcome from the exercise was obvious from the start—teams that score lots of runs and prevent the other team from scoring runs will be successful. This point is very well taken, and cannot be refuted. However, it is pointed out to the cynical student that several other, presumably important, variables were similarly explored and eliminated because of their lack of ability to explain winning percentage. These cynical students then agree that exploring other variables (such as payroll) was in fact worth doing.

At the conclusion of this exercise, it is also mentioned that having only two remaining predictor variables (ERA and RBI) is a good thing, as a parsimonious model is much easier to articulate as opposed to a more complicated model.

4. Conclusion and Effectiveness of Exercise

The example described has been used for about six years now. For AACSB re-accreditation purposes, learning objectives are monitored for each topic covered in quantitative methods. For each exam administered, several are randomly sampled, and for each, exam problems reflecting basic course concepts are graded for how well the student performs on the selected problems. While the administration is uncomfortable in sharing specific statistics, linear regression has always shown that learning objectives have been attained, and since the introduction of the baseball example, the attainment of these learning objectives has further improved.

Additionally, a survey was given to former students who were exposed to the baseball example. Three hundred students were invited to take the survey, and (133) of them responded, resulting in a response rate of 44%. The survey and resulting responses are included in the **Appendix 1**. Of particular note are a few following items. 47% of respondents feared quantitative methods, and 36% feared linear regression. 100% of the respondents found the baseball example a helpful introduction to linear regression, while 91% of the respondents found the baseball example preferable to a more traditional example of linear regression, which often involves supply and demand and the like.

Over the years, the example has evolved into a two-part exercise. The first week, the relationship between payroll and winning percentage is explored. Once it is established that this relationship is weak, other, more traditional linear regression examples are pursued. The second week, multiple linear regression is introduced, so that the student becomes aware of using multiple predictor variables to improve the R^2 term, along with the pitfalls

Table 3. Statistics for multiple linear regression model.

Term	t-statistic	*p*-value
ERA	−9.08	<0.0001
RBI	8.69	<0.0001

of multicollinearity. At that point, the baseball example is revisited (with the additional predictor variables), where the intent is to maximize the R^2 term with as few predictor variables as possible. Students are typically surprised that approximately 90% of the variation in a team's winning percentage can be explained by just two predictor variables. 98% of the students surveyed found this particular transition from simple to multiple linear regression helpful.

The survey also gives the students the opportunity to discuss their thoughts on the baseball introduction to regression. A summary of some frequent comments are as follows:

- The baseball example was a more understandable introduction to understanding the relationship between variables, as compared with more traditional examples shown in business school quantitative methods books. Students mentioned that they can better relate to the relationship between payroll and winning percentage as compared with something more esoteric.
- The baseball example was a good way to better understand the difference between independent variables (payroll) vs. dependent variables (winning percentage).
- Adding more independent variables, such as ERA and RBI is a good way to increase the predictive ability of the model.

The conversation is then brought to a close by mentioning sabermetrics, data-mining and the newer "analytics" tools currently used by sports teams to find patterns in data that can result in a competitive advantage (Lewis, 2003).

Acknowledgements

The author would like to sincerely thank the Working Professional MBA students at Wake Forest University for their feedback on the use of this teaching approach. This paper would not have been possible without their survey participation.

References

Albright, S. C., Winston, W. L., & Zappe, C. J. (2011) *Data Analysis and Decision Making* (4th ed.). Mason, Ohio: Southwestern/ Cengage Learning.

Hoaglin, D., & Velleman, P (1995). A Critical Look at Some Analyses of Major League and Baseball Salaries. *The American Statistician, 49*, 277-285.

USA Today. http://www.usatoday.com/sports/mlb/salaries/2013/team/all/

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game.* New York: W. W. Norton & Company.

Watnik, M. R. (1988). Pay for Play: Are Baseball Salaries Based on Performance? *Journal of Statistics Education, 6.*

Appendix

Appendix 1. Survey instrument used.

Q#	Question	Response
1	Did you fear Quantitative Methods?	No: 53%, Yes: 47%
2	Did you fear Linear Regression?	No: 64%, Yes: 36%
3	Did you find the simple linear regression example, payroll vs. winning percentage, a helpful introduction to linear regression?	No: 0%, Yes: 100%
4	Did the payroll vs. winning percentage problem help you better understand the difference between independent and dependent variables?	No: 1%, Yes: 99%
5	Were you surprised that the relationship between payroll and winning percentage was weak?	No: 27%, Yes: 73%
6	Were you surprised that the predictive ability (R-squared term) of the payroll vs. winning percentage model was basically non-exist?	No: 18%, Yes: 82%
7	Did you find the payroll vs winning percentage example an easier way to learn simple linear regression as opposed to other traditional type of business school examples (such as supply vs demand, etc.)	No: 2%, No Opinion: 7%, Yes: 91%
8	Did you find it helpful that we continued our baseball-related simple linear regression example to include multiple independent variables so as to increase our predictive ability for winning percentage?	No: 2%, Yes: 98%
9	Did you find the linear regression part of the class beneficial?	No: 0%, Yes: 100%
10	Please provide any comments you would like to make.	