A STUDENT FRIENDLY ILLUSTRATION AND PROJECT: EMPIRICAL TESTING OF THE COBB-DOUGLAS PRODUCTION FUNCTION USING MAJOR LEAGUE BASEBALL

James B. Larriviere, Spring Hill College
Ralph Sandler, Spring Hill College

ABSTRACT

There has been a plethora of thinking and research about better methods of teaching important economic concepts to undergraduate and graduate level students. One concept students have a particularly difficult time grasping is the Cobb-Douglas (C-D) production function. Not only does the concept create a hurdle for students, but the actual estimation and interpretation of the C-D production function has perplexed even the most dedicated students.

This paper offers a pedagogical tool to make the instruction and estimation of (C-D) production functions much more palatable to students in the intermediate level undergraduate economic courses and graduate students in managerial economics. To that end, the paper has two parts. In the first part, we introduce the production function and make the connection between team sports and the productivity of inputs. Next, we relate a specific type of production function, the Cobb-Douglas, to major league baseball teams' performance. A model is introduced and explained, with a description of the various inputs used to produce team wins. The first part concludes with the estimation of a log-linear C-D production function and an evaluation of the results.

The paper's pedagogical contribution is explained in the second part of the paper. We provide a student project that professors can use in their courses to facilitate the teaching of C-D production functions. The project comes complete with data from the 2006 MLB season, which students use to estimate a log-linear C-D production function. The steps involved in the data transformation process are provided along with the empirical results. Questions and answers about the regression results are included and could be used to reinforce students' learning and understanding of the model and output.

INTRODUCTION

Cobb-Douglas Production Function

Considering that most students have played a team sport at some point, this familiarity will make it easier for the student to grasp the connection between what the manager/coach is trying to produce (team wins) and productive inputs (which are measured by players' offensive and defensive skills). As with most team sports, baseball being no exception, team composition

is critical for maximizing wins, so the manager's task is to obtain players with the right mix of skills and abilities, bring them together, and win ballgames. For the manager, understanding the impact of inputs such as hitting, pitching, errors and stolen bases on a team's winning percentage is important for at least two reasons. First, if empirical results indicate that pitching has a greater positive impact on team wins than hitting, then the manager can trade or hire accordingly. For example, a manager may trade a player with a strong batting average in order to get a pitcher with a really low earned run average. Second, those players whose skills have the greatest positive impact on team wins (a high marginal product) may negotiate for higher salaries, which would certainly affect the team payroll. Managers (or owners) do not want to overpay for a player if his skills will not have much impact on team wins.

Let us begin with a brief explanation of the production function and how it can be used to examine major league baseball team success. The production function shows the resulting quantity of output that can be produced with a given set of inputs. It also provides a conceptual framework for decisions involving the allocation of the firm's (team's) resources. One particular form of the production function is the Cobb-Douglas (C-D). The C-D is used extensively in economic studies, and its properties have kept it popular for over eighty years.

Given the importance of production theory to microeconomic analysis and the obvious practical appeal of using the Cobb-Douglas, it is interesting to note the range of coverage found in a sample of economic textbooks. The Pindyck and Rubinfeld text (2005) is typical of the very brief discussion found in undergraduate intermediate microeconomic books. MBA level managerial economic textbooks tend to provide more coverage, sometimes including the results of empirical research using the C-D (see for example, Thomas and Maurice (2005); McGuigan, Moger and Harris (2006); Hirschey (2003); Baye (2009); Keat and Young (2006)). The textbooks listed above do not have an extensive case-like project that requires students to take a set of actual data, which is provided to them, and empirically estimate a C-D production function.

Economists have used the C-D to estimate the impact various inputs have on major league baseball teams' success. Two noteworthy articles by Charles Zech (1981) and Mark Woolway (1997) use cross-sectional data from the 1977 and 1993 baseball seasons, respectively, to empirically estimate a C-D production function. Both articles sought to establish a relationship between team victories (output) and various inputs, like hitting, pitching, and defensive skills. The variables used in our model will follow Zech (1981) and Woolway (1997) model specifications.

The C-D is typically illustrated by the following equation:

$$Q = AL^bK^c$$

where Q is output, A is the total factor productivity (the change in output not caused by the inputs e.g. by technology change or weather), and L and K are inputs, typically labor (L) and capital (K). The exponents, b and c, are to be estimated. The C-D assumes some degree of

substitutability between the inputs, albeit not perfect substitutability one finds in a strictly linear model specification. Since the C-D is a multiplicative model, not a linear model, taking the logarithms of the data is necessary in order to estimate the function using OLS linear regression. The standard log-linear model is:

$$\text{Log } Q = \log A + b \log L + c \log K.$$

MODEL

To estimate major league baseball's production function, data was obtained from *The ESPN Baseball Encyclopedia (2007).* The output measure (Q) is team victories in the regular season. The input measures consist of four primary categories: offense, pitching, defense and managerial effectiveness.

Offense requires hitting for average and hitting for power; consequently, we use team batting average (BA) and home runs (HR). We also use stolen bases (SB), an offensive sub-skill that one would expect to have an influence on winning games. We expect all of these to have a positive relationship to team victories.

Earned run average (ERA) is considered to be the most important measure of pitching effectiveness (Woolway 1997). ERA is measured by dividing the number of earned runs the pitcher allows by the number of innings pitched and is then multiplied by nine (the regulation length of a Major League Baseball game). The lower the team ERA the better; therefore, the variable should have an inverse relationship to winning percentage. Defense, the third major input category, is measured by the total number of errors committed by the team (E) and should similarly have an inverse relationship to winning percentage.

Zech (1981) did not find managerial effectiveness to be a significant factor in team wins. Nevertheless, we think some specification for this input should be considered and have included the manager's lifetime winning percentage (MW) in the equation. The variable should have a positive relationship to team victories.

In order to control for the fact that variables are measured in different units, we have indexed them by dividing all team variables (dependent and independent) by either their National or American League averages.

The Cobb-Douglas production function for Major League Baseball will be defined as follows:

$$V = A * BA^{a} * HR^{b} * SB^{c} * ERA^{d} * E^{e} * MW^{f} \qquad (1)$$

In order to estimate the parameters of this nonlinear equation, it will be transformed into a linear form. This is accomplished by taking the natural logarithms (log) of both sides of equation (1). The transformed equation is:

$$\text{Log } V = \log A + a \log BA + b \log HR + c \log SB + d \log ERA + e \log E + f \log MW \quad (2)$$

where

V = team victories/league average
BA = batting average/league average
HR = home runs/league average
SB = stolen bases/league average
ERA = earned run average/league average
E = errors/league average
MW = manager's lifetime winning percent/league average

EMPIRICAL RESULTS

An OLS regression was performed on equation (2), and the estimated coefficients for a, b, c, d, e, and f are provided in Table 1. Three key variables, batting average (BA), home runs (HR) and pitching (ERA) were found to be statistically significant (at the .05 level). The variables had the correct sign, except for stolen bases.

The relative size of the estimated parameters is also revealing. Consider the estimate for batting average (BA). A 10% increase in team batting average is associated with a 14.27% increase in team winning percentage. As for the impact of home runs on team winning percentage, a 10% increase in home runs is associated with a 2.07% increase in team winning percentage. The impact of hitting for average (BA) is nearly seven times greater than hitting for power (HR). The long lasting debate among baseball fans as to the relative value of hitting and pitching is also addressed. The two hitting variables combined (BA and HR) have a greater impact than pitching, a result consistent with the findings of both Zech (1981) and Woolway (1997). It is also interesting to note that three input variables (SB, E, and MW%) were not found to have a significant effect on team winning percentage, which is consistent with Zech's (1981) earlier findings.

Our results so far have established a relationship between team victories (output) and various inputs for just one year (2005). However, further evidence is necessary before coming to any strong conclusions about the keys to success for major league baseball teams. Our statistical evidence should be reaffirmed over several years. Therefore, the same cross-sectional regression model is applied using data for the 2000, 1996, 1990 and 1985 seasons. [1] Results are provided in Table II, a summary of which follows:

1. The results of our regression analysis using the same variables over four additional years are rather consistent. The R^2 for all five years ranges from .82 to .91. Batting average

(BA), home runs (HR), and earned run average (ERA) all have the correct sign and are statistically significant at the .05 level in each of the five equations.

2. The independent variable stolen bases (SB) was only significant in one year (1996), while the number of errors committed by a team (E) was significant in two years (1996 and 1985). In each case, the variable reflected the correct sign.
3. The independent variable MW, which represents the manager's lifetime winning percent, was only significant in the 1996 season.

In theory, the role of the manager should be indispensable to a team's success. Managers have the responsibility of training and motivating players, maintaining player morale, as well as making tactical decisions during a game to enhance the team' success. Yet, we do not find strong statistical evidence in our results that supports the theory. It may well be managers serve only a marginal role in a team's success. However, it seems unreasonable to believe that rational, profit-maximizing owners would pay relatively large salaries to managers despite their questionable value to the organization. Another possible explanation is that our results may just reflect the variable used as a proxy for managerial efficiency (MW). The absence of solid quantitative data makes it very difficult to come up with alternative measures. Finding an effective measure of managerial efficiency would be an excellent topic for further research.

STUDENT PROJECT

Estimating log-linear Cobb Douglas Production Function

This second part of the paper describes a project that requires students to estimate a Cobb-Douglas production function using 2006 MLB season data. By having them estimate and evaluate a regression equation using real economic data, students are much more likely to understand the economic and statistical concepts.

It is suggested that instructors provide students with copies of **Part I** of this paper. The instructor may assign the Instroduction to the students, discuss the paper in class, and then assign the student project described here in the Student Project**.** The raw data for the 2006 baseball season is provided in Appendix A. *Excel,* as well as other spreadsheet statistical packages, can be used to transform the data and run the statistical analysis.

Data Transformation

The data in Appendix A are measured in different units; therefore, it is customary to use a type of index in order to utilize the data in the model. The first step toward indexing the variable values is to find the league average for each variable. Next, each team's variable values are divided by the league average. For example, if Boston's wins 86 games, and the American

League's average number of victories is 83, then Boston's indexed number of wins is 86/83 = 1.036. This process is applied to all seven variables.

The second step in the data transformation process is determining the natural log of the data. The natural log is determined for each indexed variable. In the *Excel* package, the LN function is applied. Once all data are transformed into their log form, regression analysis can be performed.[2]

Regression Results and Sample Questions

Regression results from *Excel* are provided in Appendix B for instructors. The instructor may use the following questions to evaluate students' understanding of the model and interpretation of the estimated coefficients (reasonable answers are given in italics after each question.)

Question 1: Calculate the direction and statistical significance of the independent variables.

In evaluating the significance of the independent variables a useful "rule of thumb" is that if the absolute value of the t-statistic is greater or equal to 2, then the parameter estimate is statistically different from zero at the .05 level of significance (Baye, p. 100). In addition, a low P-value (lower than .05) suggests only a small chance that the true coefficients are actually zero. By these standards, the independent variables HR and ERA are statistically significant, where as the variables BA, SB, MW are not. The variable E is significant at the .10 level of significance. As hypothesized above, the coefficients of all variables have the correct sign, indicating that the model is consistent with our theory about the Cobb-Douglas production function as applied to Major League Baseball.

Question 2: Evaluate the overall performance of the model.

The R-square (coefficient of determination) and F-statistic tell us about the overall performance of the model. The R-square, which tells us the fraction of the total variation in the dependent variable explained by the regression, is .80 in Appendix B. In addition, the F-statistic, which allows one to objectively determine the statistical significance of any regression, suggests there is an infinitely small chance (3.97E-07) that the estimated regression model fits the data purely by accident.

Question 3: Define output elasticity. Given the regression results for 2006 (Appendix B), what impact would a 10 percent increase in HR have on output?

In a Cobb-Douglas production function, where the data are transformed by taking the natural log of all variables, the coefficients are all output elasticities. An output elasticity measures the percentage change in output (team victories) divided by the percentage change in some input variable. For example, the output elasticity for the coefficient "b" (HR) is

$$b = \frac{\text{percentage change in output (victories)}}{\text{percentage change in home runs}}$$

So that a 10 percent increase in a team's HR, ceteris paribus, should lead to a 2.29% increase in output where

Percentage change in output = (percentage change in HR) x b

= (+) 10% x .229
= 2.29 percent

Question 4: Reflect on your results within the broader context of empirical research. Specifically, compare and contrast your results with the five regression equations discussed in this paper.

As in the other five equations, the Cobb-Douglas production function using the 2006 Major League Baseball season provides a reasonably good fit (R^2=.80, low significance value of the F-statistic). In addition, HR and ERA are statistically significant (at the .05 level of confidence) and have the correct sign. However, unlike the other years, batting average (BA) is not significant in this equation.

CONCLUSION

This paper has two purposes. The first is to introduce the Cobb-Douglas production function to students in a team sport application: major league baseball. By using the popular national pastime as an illustration, we think students will become more interested in the broader topic of production analysis. In estimating the Cobb-Douglas production function, cross-sectional regressions were run for five different years (1985, 1990, 1996, 2000, and 2005) thus strengthening the confidence we have in our results that are reasonably consistent over time. Our empirical results demonstrate that:

1) Batting average (BA), home runs (HR) and earned run average (ERA) are statistically significant in each of the five years examined.
2) Consistent with previous research, the relative size of the coefficients suggests hitting (both BA and HR) contributes more to team victories than pitching.

3) In four of the five years examined, our measure of managerial effectiveness was not significant. We attribute this to the proxy used (MW). Finding an alternative measure of managerial effectiveness would be an excellent topic for further research.

In the second part of the paper we propose a student project that utilizes the Cobb-Douglas specification with data from the 2006 Major League Baseball regular season. Based on the model previously specified, students are provided the raw data and then required to transform the data, estimate the model and ultimately explain the economic and statistical concepts. We believe such a project will enhance their understanding of the material beyond that achieved by a lecture alone. The hands-on exercise centered around a popular sport is more likely to grab the attention of students than more traditional economics queries. Additionally, being able to compare the findings of past research with regression coefficients that the students themselves estimate is an effective teaching tool.

ENDNOTES

[1] Although our initial strategy was to select baseball seasons at five-year intervals, the 1995 season was shortened by the players' strike that may have given us abnormal results. Consequently, we used the 1996 season instead.

[2] Within *Excel*, have students click **tools**, and then **data analysis** (some computers may require operators to add-in the analysis tool pak) followed by **regression**. The dependent variable column should be added to the input Y-range and all independent variable columns should be placed in the input X-range. It is suggested that they request **labels** and a **95% confidence interval**.

REFERENCES

Baye, M. 2009. *Managerial Economics and Business Strategy.* 6th edition. Boston, MA: McGraw Hill Irwin.

Cobb, Charles W. and Douglas, Paul H. 1928. Theory of Production. *American Economic Review* 8:139-165.

Hirschey, Mark. 2003. *Managerial Economics*. Mason, OH. Thomson-Southwestern.

Keat, Paul G. and Young, Philip. 2006. *Managerial Economics, Economic Tools for Today's Decision Makers.* 5th edition. Upper Saddle River, NJ. Pearson- Prentice Hall.

McGuigan, J., Moyer, R., Harris, F. 2005. *Managerial Economics, Applications, Strategy And Tactics.* 10th edition. Mason, OH. Thomson-Southwestern.

Pindyck, R. and Rubinfeld, D. 2005. *Microeconomics.* 6th edition. Upper Saddle River, NJ. Pearson-Prentice Hall.

Thomas, C. and Maurice, S. 2005. *Managerial Economics.* 8th edition. Boston, MA. McGraw Hill-Irwin.

The ESPN Baseball Encyclopedia. 2007. 4th edition, edited by Gary Gillett and Pete Palmer. New York, NY. Sterling Publishing Co.

Woolway, Mark D. 1997. Using an Empirically Estimated Production Function for Major League Baseball to Examine Worker Disincentives Associated with Multi-Year Contracts. *The American Economist.* 41:77-83.

Zech, Charles E. 1981. An Empirical Estimation of a Production Function. *The American Economist.* 25:19-23.

Table 1: Regression Results **2005**			
Parameter	**Estimate**	**t-ratio**	**P-value**
A (Intercept)	-0.09	-0.686	0.500
a (BA)	1.427	4.316*	0.001
b (HR)	0.207	3.686*	0.001
c (SB)	-0.008	0.365	0.72
d (ERA)	-0.671	-6.385*	2E-06
e (E)	-0.224	-1.778	0.089
f (MW)	0.237	1.647	0.113
F-value 24.60 R square .87 *Significant at .05 level			

Table II: Regression Results **Multiple Years**					
	2005	**2000**	**1996**	**1990**	**1985**
Parameters					
A (Intercept)	-0.009	-0.008	-0.005	-0.008	-0.012
a (BA)	1.427*	1.550*	1.500*	1.291*	1.741*
b (HR)	0.207*	0.183*	0.224*	0.163*	0.306*
c (SB)	-0.008	0.001	0.094*	-0.0001	0.036
d (ERA)	-0.671*	-1.227*	-0.690*	-0.891*	-0.916*
e (E)	-0.223	-0.008	-0.200*	-0.110	-0.491*
f (MW)	0.237	-0.166	0.347*	-0.013	-0.219
F-value	24.60	37.09	39.54	14.20	25.29
R-square	0.87	0.91	0.91	0.82	0.89

Appendix A: Raw Data 2006							
Team AL	**V**	**BA**	**HR**	**SB**	**ERA**	**E**	**MAN W/L%**
NY	97	0.285	210	139	4.41	104	0.536
BOS	86	0.269	192	51	4.83	66	0.497
TOR	87	0.284	199	65	4.37	99	0.514
BAL	70	0.277	164	121	5.35	102	0.429
TB	61	0.255	190	134	4.96	116	0.419
CHI	90	0.28	236	93	4.61	90	0.56
CLE	78	0.28	196	55	4.41	118	0.492
DET	95	0.274	203	60	3.84	106	0.493
KC	62	0.271	124	65	5.65	98	0.416
MIN	96	0.287	143	101	3.95	84	0.562
OAK	93	0.26	175	61	4.21	84	0.568
SEA	78	0.272	172	106	4.6	88	0.501
ANA	89	0.274	159	148	4.04	124	0.537
TEX	80	0.278	183	53	4.6	98	0.514
Team NL							
ATL	79	0.27	222	52	4.6	99	0.563
NY	97	0.264	200	146	4.14	104	0.556
FLA	78	0.264	182	110	4.37	126	0.481
WAS	71	0.262	164	123	5.03	131	0.475
PHIL	85	0.267	216	92	4.6	104	0.535
STL	83	0.269	184	59	4.54	98	0.536
CIN	80	0.257	217	124	4.51	128	0.473
MIL	75	0.258	180	71	4.82	117	0.45
HOU	82	0.255	174	79	4.08	80	0.483
PIT	67	0.263	141	68	4.52	104	0.508
CHI	66	0.268	166	121	4.74	106	0.527
SF	76	0.259	163	58	4.63	91	0.503
LA	88	0.276	153	128	4.23	115	0.568
ARI	76	0.267	160	76	4.48	104	0.477
COL	76	0.27	157	85	4.66	91	0.447
SD	88	0.263	161	123	3.87	92	0.494

Appendix B: Regression Results 2006					
Regression Statistics					
Multiple R	0.897012198				
R Square	0.804630883				
Adjusted R Square	0.753665026				
Standard Error	0.062587744				
Observations	30				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance of F*
Regression	6	0.371062625	0.061844	15.7876456	3.97278E-07
Residual	23	0.090096191	0.003917		
Total	29	0.461158816			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
A (Intercept)	-0.007277222	0.011697174	-0.62214	0.53997215	
BA	0.799905128	0.494106737	1.618891	0.11910328	
HR	0.229302108	0.085878985	2.670061	0.01367548	
SB	0.040020952	0.036791781	1.087769	0.28796517	
ERA	-0.799607547	0.165322329	-4.83666	6.9971E-05	
E	-0.183639004	0.09186517	-1.99901	0.05755922	
MAN W/L%	0.280892048	0.19198527	1.463092	0.15697493	