

Maximization of Profit:
An Application of the Bootstrap and Regression Analysis

A Thesis Presented to the University Honors Committee
Southern Connecticut State University

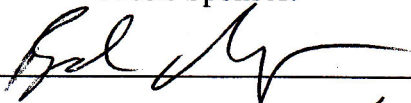
In Partial Fulfillment of the Requirements for Departmental Honors in
Mathematics and for Graduation from the Honors College

By: Melissa Harrigan

9 November 2007

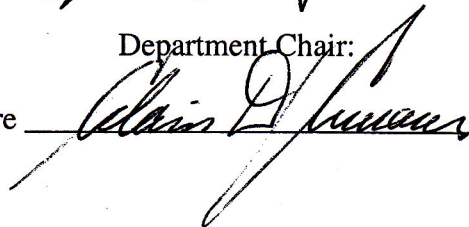
Thesis Sponsor:

Signature



Department Chair:

Signature



Abstract

The concepts of the bootstrap and regression analysis have been around for centuries; however, it hasn't been until recently that the uses and applications of the bootstrap have appeared in statistical research. One of the main reasons for this is that the bootstrap and regression analysis are both applications that require the use of computers and computer programming. Since technology has improved recently, it has become easier to make the complicated calculations that are associated with creating models using the bootstrap and regression analysis. This study will demonstrate that the use of the bootstrap and regression analysis are an efficient and accurate ways to maximize the profit of an at-home EBay[®] business. Regression analysis was used in this study to create a model that would estimate the profit of the business, and from this model it was possible to find a way to maximize profit. Bootstrapping was then used to determine the accuracy of the results from the regression model. The results of this thesis show how to maximize profit for each type of rock sold and through bootstrapping; the model that was used was found to be extremely accurate.

Table of Contents

Section	Page No.
Introduction	1
Literature Review	1
Mathematical Background	2
Methods	7
The Final Model	11
Interpretation and Analysis of the Model	12
The Bootstrap	15
Discussion	17
Conclusion	19
Appendix A	21
Appendix B	24
Appendix C	26
Appendix D	30
Appendix E	33
Appendix F	36
Bibliography	39

Introduction

The concepts of regression analysis and bootstrapping have been around for centuries, but it hasn't been until recently that these two concepts have been used together and have started to make an impact in the mathematical community (Chernick 1999, Lange 1999). Regression analysis is a statistical method that takes a set of data points and creates a model that can then be used to better understand the relationship of variables or make predictions about future results. Bootstrapping is a resampling method that uses the original data to estimate a parameter or estimate the standard error of an estimator (Chernick 1999). As technology has improved, statisticians have been able to better understand the use of bootstrapping and the potential impact it may have on the mathematical community in the future. This thesis will apply the two methodologies of regression analysis and bootstrapping to a real-world application by using data from a particular at-home business to find a way to maximize profit.

Literature Review

The concepts of bootstrapping and regression analysis have been well-known topics in the statistical community for many years. Some research involving the bootstrap method includes, but is not limited to, a book written by J.S. Hjorth (1994), which discusses bootstrap methods and the use of computers. Also, R. Cheng (2005, 2001) has given many speeches over the past six years at the annual Winter Simulation Conference about his research experiments involving bootstrap resampling. According to Chernick (1999), "regression analysis is one of the most widely used statistical techniques." Additional research on regression analysis includes an article by Hoyt,

Leierer, and Millington (2006) that analyzes different multiple regression and correlation methods. Also, articles by Jianxin and MacKenzie (2006); and Vermunt (2005) are examples of more current research that has currently been done on regression analysis. While these two techniques have been used for many years, they are not often used together. These two types of analyses when used together have just recently begun to make an impact in the statistical community. An article by Wang, Linton and Hardie (2004) discusses how bootstrapping is used to approximate the standard error of certain estimators that were then used to create a particular linear regression model. Additional research involving the bootstrap and regression analysis include recent articles by Hui, Modarres, and Zheng (2005); Li (2005); and comprehensive and informative books by Good (1999); and Efron and Tibshirani (1993).

Mathematical Background

In mathematics, *models* are used to display and explain results and conclusions in a simplified and understandable manner. More specifically, with regression analysis, models are used to display the results and analyze the data (Devore 2004). In this study, we created a model using regression analysis to optimize the profit of an at-home business. Through the model, we were able to show the relationship between multiple variables that affect profit; but are related in a nondeterministic fashion. Two variables that have a *nondeterministic* characteristic are related to one another, but knowing the value of one variable does not mean that it is always possible to determine the exact value of the second variable (Devore 2004). For example, a regression model that shows the relationship between a student's high school GPA and their college GPA would be an

example of a nondeterministic relationship. This is because a student's academic performance in high school is a good predictor of how they will perform in college. However, two students with the same high school GPA may not have the same college GPA. Since these two variables are related but there is no way of finding an exact high school GPA by knowing a student's college GPA (or vice versa), this relationship is considered to be nondeterministic. For the model in this thesis, we were able to determine which variables would maximize the profit.

The model used in this thesis uses several variables; therefore multiple regression is used to find a relationship between the dependent variable, profit, and the independent variables. When creating a model using regression we want to find the estimate of the following equation,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon.$$

In this equation, y represents the dependent variable, x_1, x_2, \dots, x_k are the independent variables, and $\beta_1, \beta_2, \dots, \beta_k$ determine the contributions of the independent variables x_1, x_2, \dots, x_k to the model. Also, β_0 is the y -intercept of the model and ϵ is the random error of the model (McClave and Sincich 2003). Since there is a random error term in the model, there is no definite relationship between the independent variables; and therefore they are related in a nondeterministic fashion. Thus for the model created in this study, y represents the profit, x_1, x_2, \dots, x_k are the variables that affect the profit, and $\beta_1, \beta_2, \dots, \beta_k$ represent how much each variable affects the profit. We want to find an estimate of $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ because they are unknown population parameters. Through regression analysis we are able to create a model that will estimate the profit of the business that we studied. Since we use multiple independent variables in our model, we decided to create a

second-order regression model to estimate the profit. In addition to the linear terms, the second-order model also includes interaction terms and quadratic terms. An *interaction term* is the cross-product of two independent variables (Mendenhall, Sincich 2003). For example, x_1x_2 would be the interaction term between the two independent variables x_1 and x_2 . A *quadratic term* is the square of an independent variable (Mendenhall, Sincich 2003). Thus, x_1^2 would be the quadratic term associated with the independent variable x_1 . A complete second-order model with two independent variables would have the following equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \hat{\beta}_3x_1x_2 + \hat{\beta}_4x_1^2 + \hat{\beta}_5x_2^2$$

where \hat{y} and $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_5$ are the estimates of the variables from the original equation (McClave and Sincich 2003). Also, x_1 and x_2 are the linear terms. The model that was used in this thesis has more than two independent variables, so the equation of the regression model will be very similar to this example, with the addition of more linear, interaction and quadratic terms. From regression analysis it is possible to determine which variables significantly affect the change in profit, and therefore we can use the model to find a way to maximize the profit.

In addition to creating the regression model, in this study we also had to interpret and analyze the model. The first thing we do when interpreting a model, is to determine if it is an adequate model. The *adjusted multiple coefficient of determination*, or adjusted R^2 , is a very good indicator of the adequacy of a regression model because it represents the proportion of variability in profit explained by the model. Thus, the higher the adjusted R^2 the less inconsistency there is with the data being used in the model. The reason we want to look at the adjusted R^2 value instead of the R^2 (multiple coefficient of

determination) is because the adjusted R^2 takes into account both the sample size and the number of β parameters in the model (Mendenhall, Sincich 2003).

Another way to check the reliability of the model is to use the F -test. For the global F -test, the null hypothesis is $\beta_1=\beta_2=\dots=\beta_k=0$ and the alternative hypothesis is that at least one of the coefficients (β_i) is nonzero (Mendenhall, Sincich 2003). Thus, if the null hypothesis is rejected then there is sufficient evidence to support that the model is more useful than no model at all for predicting y (Mendenhall, Sincich 2003). When interpreting the F statistic, we want a larger number because this indicates a larger proportion of the total variability accounted for by the model (Mendenhall, Sincich 2003). Thus, a larger F statistic indicates a more reliable model. When interpreting our regression model, we want a large F statistic and a p -value less than 0.001. Additionally, when we interpret the mean square of the regression model versus the mean square of the residual error, we want the mean square of the regression model to be larger. If both, the F statistic and the mean square of the regression model are large numbers, then we can conclude that the model is adequate.

Residual analysis is also used to determine the adequacy of a model. A *residual* is an estimate for the random error term of the equation that we want to model and can be represented with the following equation:

$$\hat{\varepsilon} = y - \hat{y}.$$

In the equation $\hat{\varepsilon}$ represents the residual, y is the observed value of the dependent variable and \hat{y} is the predicted value. In this study, y represents the actual profit and \hat{y} is the estimated profit found through the model. By analyzing the residuals, we can determine if there are any outliers in the model. An observation is considered to be an

outlier if its residual is greater than three in absolute value. By identifying the outliers, it is possible to determine which observations may have more influence in the model (Mendenhall, Sincich 2003).

In creating the regression model, one important assumption that is made is that the random errors, ϵ , are independent. It is possible to see that the random errors are independent by analyzing a graph of the standardized residuals, $\hat{\epsilon}$, and the actual profit values. If there is no apparent pattern in the graph then this means that the random errors are independent. This becomes important when determining if the regression model being used is adequate (McClave, Sincich 2003).

The other statistical method that is used for this thesis is the bootstrap. Bootstrapping uses the original data as the population and then, by sampling with replacement, bootstrap samples are created. When the bootstrap samples are created using the *sampling with replacement* method, it is possible for a data point to be chosen, returned back to the population sample and then chosen again as another member of the bootstrap sample. These bootstrap samples are then used to estimate the values of the population parameters (Devore 2004).

A *parameter* is a value that defines a certain characteristic of the population, such as the mean, median, standard deviation or the population proportion (Devore 2004). In many instances, the population is too large or the population parameters are unknown; thus in order to analyze the population an estimator is used to estimate the unknown parameters. An *estimator* is a function used to estimate an unknown population parameter because the value of this function changes from sample to sample (Devore 2004). Since the value of the function changes, the function is considered a *random*

variable. For example, in this study the function that determines how much an item sells for is not the same in every case, and therefore it is considered a random variable. Since there are many data points in this study, and the population parameters ($\beta_1, \beta_2, \dots, \beta_k$) are unknown, it is possible to use the bootstrap method to estimate the population parameters. For example, to estimate the population mean by using this method, a number of random bootstrap samples ($\theta_1, \theta_2, \dots, \theta_n$) would be created. Then the mean of each bootstrap sample ($\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$) would be calculated. Finally, the bootstrap estimate of the population mean ($\hat{\theta}$) would then be found using the following equation, where n is the number of bootstrap samples (Efron, Tibshirani 1986):

$$\hat{\theta} = \frac{\sum_{i=1}^n \bar{x}_i}{n}$$

In this example, the estimator is the mean of each bootstrap sample, and the unknown parameter would be the actual population mean. Bootstrapping can also be used to assess the accuracy of an estimator, by estimating the standard error of an estimator. This is important when creating models because the more accurate the estimator, the more reliable the model. For this thesis, the bootstrap method was used to determine the accuracy of the parameters $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ found in the regression model.

Methods:

Before the regression model could be created, the raw data that was going to be used to generate the model had to be collected. For this study, the data comes from the at-home business VAM, which is a business conducted through EBay[®]. The items sold

are various types of rocks. This particular company was chosen because the business was willing to supply their data, which was also readily available and very well organized. VAM sells between thirty to forty different types of rocks and there are two separate auctions each week, where about twenty items are sold in each auction. The sales data from 2006 was used because this is a fairly new company, and the sales data is from their second complete year of operation. Once the data was collected, the next step was to determine what parts of the data were going to be used in the model. The independent variables that were used in the process of creating the final regression model are the size of the rock, the type of rock, the purchase price of the rock, and the total cost of selling the rock. Once the data was extracted, the next step was to create different second-order models in order to determine which model would produce the most adequate results.

Once the data was imported from the Excel[®] spreadsheet that was provided by VAM to Minitab[®], it was possible to create numerous regression models in an efficient amount of time. However, before any models could be created, the qualitative data that was being used had to be coded so that when used in the model it would be treated as quantitative data. The only qualitative data that was used in the models was the type of rock that was sold. There were four different types: iron, stone, stony-iron, and impactite. Since these don't have numerical values, these variables had to be coded in Minitab[®] so that they could be used in the model. In order to code them, three 'dummy' variables were used and given the following titles in the spreadsheet: impactite, stone and stony-iron. After coding, only 1's and 0's appeared in these three columns, and a 1 indicated that type of rock was sold and a 0 indicated that it wasn't sold. For example, if a 1 appeared in the stone column, then that would mean that a stone rock was sold and

0's would appear in the impactite and stony-iron columns for that observation. However, if all three columns had 0's in them this would indicate that an iron rock was sold.

Additionally, before any models could be created all the interaction and quadratic terms had to be calculated using the raw data because the final model was going to be a second-order model. Minitab[®] has a tool that can multiply any columns together and will store the results in a new column. So to create the interaction term, size*sale amount, Minitab[®] would multiple the value in the size column by the value in the sale amount column and store the result in a new column, which was entitled size*sale amount. Once the qualitative data was coded and all the interaction and quadratic terms were calculated, it was possible to start the regression analysis process.

The first model that was produced was the complete second-order model, which meant that all the first order terms, interaction terms and quadratic terms were included in the model. It became apparent that this model would not be useful because the coefficients of the sale amount and total cost terms were 1 and -1, respectfully; and the coefficients of the other terms were all zero. Although this model would not be used, it did provide valuable information that would be used in creating the final model. From this first model we discovered that sale amount and total cost are linearly dependent and therefore our final model would either exclude all sale amount terms or all total cost terms. Linear dependence causes a problem with creating regression models because regression analysis uses invertible matrices as part of the process of creating the models. Thus, if two variables are linearly dependent, they form a matrix that has a determinant of zero and thus is not invertible.