Price of Diamonds

Mitchell Pudil

January 20, 2019

1 Introduction

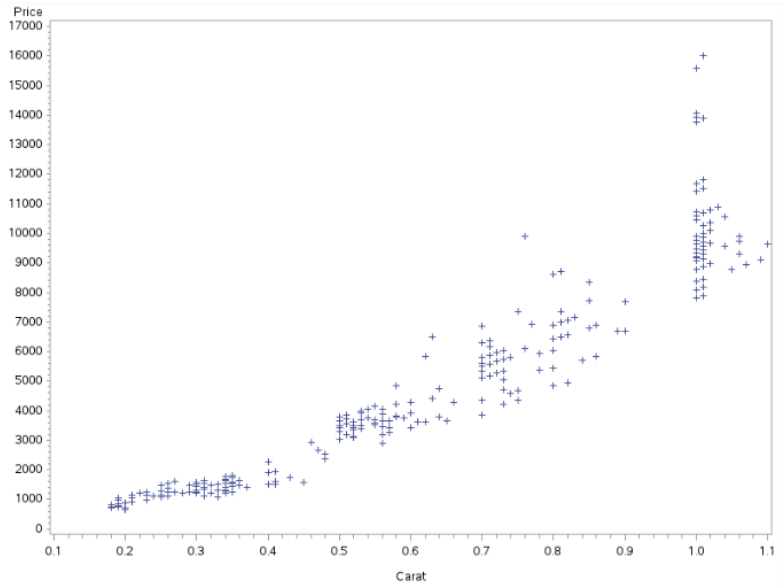
In March 2000 Dr. Chu asked his MBA students to predict the price of a diamond from the four C's. This problem focuses on predicting Price (in Singapore$) using Carat (the weight of the stone, where one carat is equivalent to 0.2 grams).

2 Data

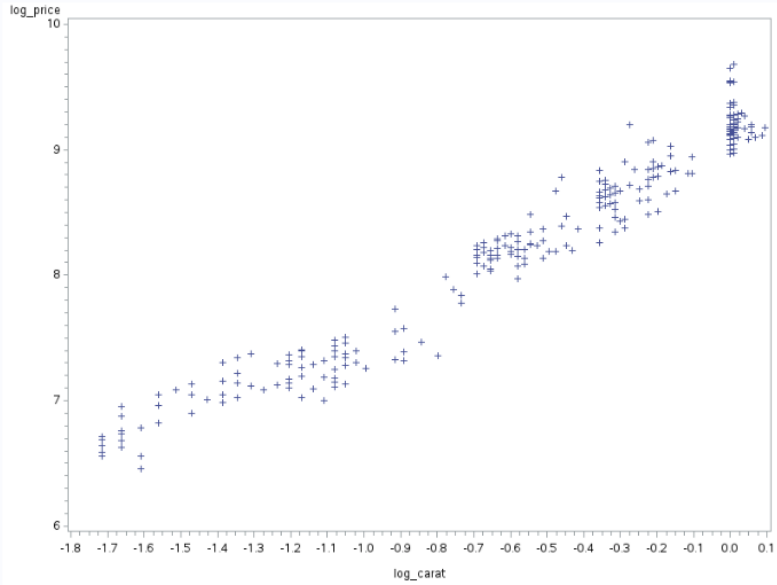
Chu's 2001 Journal of Statistics Education paper is at http://www.amstat.org/publications/jse/v9n2/datasets.chu.html and it describes how the MBA students gathered the data from an advertisement in Singapore's Business Times edition of February 18, 2000 of 308 round diamond stones (the most popular shape). The data is available at http://www.amstat.org/publications/jse/v9n2/4Cdata.txt

3 EDA

In order to find what model we should use, let's compare the scatterplot of Carat and Price to the scatterplot of ln(Carat) and ln(Price). Our goal here is to find a linear relationship between these variables so we can use them in a linear regression. Here is the scatterplot of Carat and Price:



And here is the scatterplot of log(Carat) and log(Price):



Since the logged version has more of a linear relationship, we will use that model, which will be more formally shown in the Analysis section

4 Analysis

The response variable for this experiment is the log of price, and the explanatory variable is the log of the number of carats. Since we are using the log of price and carat, The regression model we will be using is:

$$log(Price) = \beta_0 + \beta_1 log(Carat) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

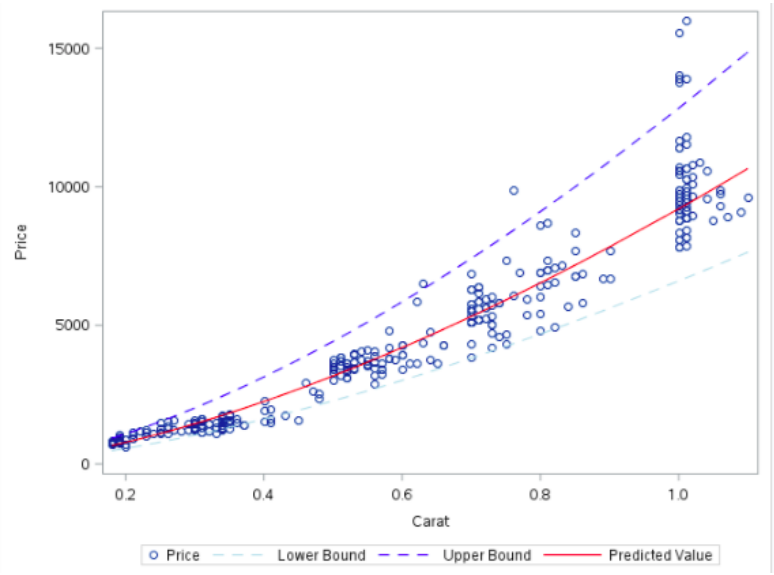
After running the regression, it was determined that the model parameter for $\beta_0 = 9.128$ with standard error (s.e.) of 0.014, $\beta_1 = 1.537$ with s.e. of 0.019. This means that the equation for the model is:

$$log(Price) = 9.128 + 1.537 * log(Carat) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

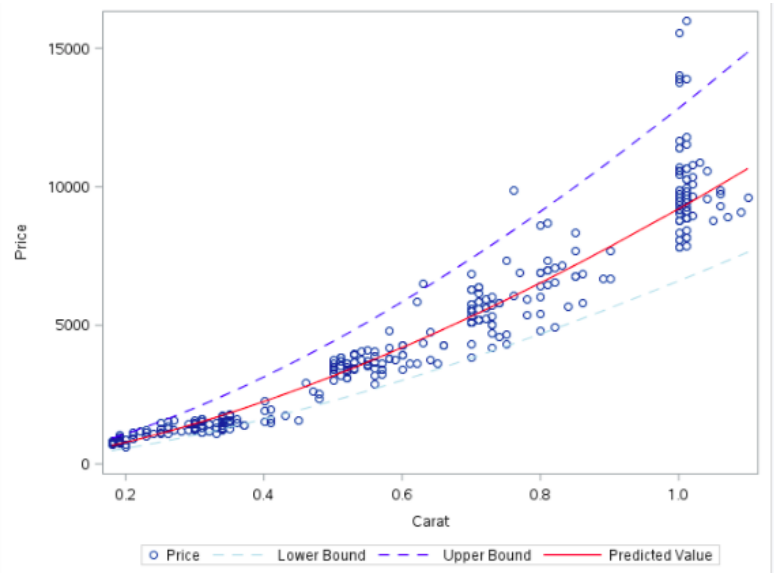
The value of β_1 means that for a 1% increase in Carat, the price increases by 1.537%. The 95% confidence interval for this is (1.500, 1.574), so there is a significant effect of carat size on price. On the other hand, here is the estimated model for $\widehat{Price}$:

$$\widehat{Price} = -2298.36 + 11599 * log(Carat) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

For the jewlry store owner, I have created a publication quality graph that shows the uncertainty associated with price, which will be used for the presentation. The graphic is shown below.



As a newly-engaged couple going ring shopping, I have calculated for you a 95% confidence interval for the most popular one carat diamond. This confidence interval is ($6602.48,$12839.74). Below is a graph the uncertainty associated with price in the estimated model.



The R^2 for price is 0.905, which means that carat size explains 90.5% of the

variation in the price. This means that the simple bivariate model chosen earlier is a great predictor of price. Below are the summary statistics of the absolute prediction error:

Analysis Variable : Absolute_Prediction_Error				
N	Mean	Std Dev	Minimum	Maximum
308	602.5675202	855.8674905	2.7522880	6658.77

The model predicts well when the carat size is smaller. As the carat size gets larger, we have a higher variance. A way to fix this prediction is to get more data on larger diamonds so that we can more accurately predict these higher carat values.

5 Conclusion

A research task that would match the strengths of this type of model is to determine how acreage affects housing prices (data at https://catalog.data.gov/dataset?tags=real-estate). One of the weaknesses of the data was that we were only using one variable to predict, which in generally can lead to problems with endogeneity.