Looking at the Price of Diamonds: A Comparison between a Linear Model (LM), Generalized Linear Model (GLM), Bayesian Generalized Linear Model (Bayes GLM), and Markov Chain Monte Carlo (MCMC) for the Hierarchical Gaussian Linear Regression Model

Aaron A. Gauthier

Bayesian Statistics

CCAS – Data Science Department

The George Washington University

December 4, 2019

Looking at the Price of Diamonds: A Comparison between a Linear Model (LM), Generalized Linear Model (GLM), Bayesian Generalized Linear Model (Bayes GLM), and Markov Chain Monte Carlo (MCMC) for the Hierarchical Gaussian Linear Regression Model

Introduction

Problem. Diamonds are amongst the most valuable natural resources in the world. They are also amongst the hardest for the average person / consumer to quantitively figure out its value. This project will allow me to understand what attributes of a diamond affect its prices and which combinations of characteristics lead to the best retail price. How do they know how much a diamond retails for? Is the price accurate? Normally, one needs to identify the characteristics of a diamond and determine if the diamond is desired and a good price. This is normally done through some sort of comparative analysis across multiple retail sites. After much painstaking research and qualitative analysis, a consumer decides whether to buy a diamond based on multiple criteria. This task is very difficult and time consuming for non-experts, so I thought that maybe utilizing a Bayesian GLM and doing a comparison of various models may help to shed some light on diamond data. This way, novices can avoid wasting time trying to identify a diamond that is too costly or worse not the proper value for the price paid.

The data for this project has been web scraped from various diamond retail websites and has been formatted in a comma delimited format (.csv) or (.txt) file. There are approximately 200,000 lines of data. I used a small subset of data ~ 31,594 lines for this analysis. The data will include diamond characteristics such as cut, color, clarity, price (in various denominations), ID number, certification, and depth measurements. I hope to build some models that can help us understand if the pricing of diamonds is accurate comparatively across lots of data.

Motivation. Bayesian techniques can now be applied to many complex modeling problems due to the explosion of cheap, yet powerful computers and processors. Utilizing the Bayesian approach will continue to challenge and possibly uproot traditional frequentist statistical methods with have dominated so many areas of science and research for so long. The primary objective of this project is to validate the price of diamonds based on their attributes namely cut, color, clarity and carat. Additional attributes that may be explored (time permitting) could include depth and brilliance (shine) characteristics amongst others. This project will also serve to identify diamonds that are priced accordingly and in other instances see if some diamonds are priced such that they are a

good buy – price arbitrage. What are the characteristics of a diamond that contribute most to its price / value? How accurately can we predict a diamonds price? Hopefully some of the work completed here can be further used to come up with a good predictive model of diamond price. This is the first step in the process of utilizing a Bayesian GLM to get to that accurate predictive capability and helps people make better decisions about whether to buy a diamond.

Domain Knowledge. While reading about diamonds, academic research and personal research I have accumulated a rather deep understanding of the aspects that make a diamond valuable and hence desired. I believe that I have quite a bit of domain knowledge especially after doing a previous project on diamond price predictions utilizing a Linear Regression Model. In this project I will do a comparison of four different models to see if the price is accurate or rather within tolerance or a credible range.

Proposed Methods

I am going to compare multiple models utilizing different packages within R. I also utilized Label Encoding for the features I am utilizing. The standardization step was omitted because nearly all the data is categorical, so we will mostly have data with label encoded variables. One hot encoding would introduce up to 450 addition columns of data which would make it very hard to get an accurate result which is why I decided to utilize label encoded variables. Additionally, standardizing is not necessary for this problem. A simple model will be specified. It will be modeled utilizing traditional techniques then compared to that with a Bayesian approach.

Linear Model (model1.1). I utilized a simple linear regression model utilizing the *lm* package in R. A simple linear regression model describes the relationship between two variables x and y. The numbers α and β are called parameters and ϵ is the error term. Here is the equation: $y = \alpha + \beta x + \epsilon$

Generalized Linear Model (GLM – model1.2). The *glm* is specified by giving a symbolic description of the linear predictor and a description of the error distribution. There are multiple categories of the *glm* such as binomial, gaussian, Gamma, inverse.gaussian, poisson, quasi, quasibinomial and quasipoisson. I picked a *gaussian glm.*

Bayesian Methods. Bayesian methods focus on five essential elements. The first is the incorporation of the prior information. Prior information is generally specified quantitatively in the form of a distribution whether normal / Gaussian, Poisson or binomial amongst others. It represents the probability distribution for a coefficient – meaning the distribution of the probably values for a coefficient we are trying to model. The prior is combined with a likelihood function. The likelihood function represents the data. The combination of the prior with the likelihood function results in the creation of a posterior distribution in order to create a distribution of likely values for the price parameter.

Simulates are drawn from the posterior distribution to create an empirical distribution of likely values for the price parameter. I have used basic statistics to summarize the empirical distribution of the simulates from the posterior. The mode, median or mean of this distribution represents the maximum likelihood estimate of the true coefficient's price value and credible intervals which can capture the true price value with the probability attached. Priors should be rationally and honestly derived. Since the diamond data comes with recognized industry certifications, we can guarantee that each diamond has been reviewed by an expert in their field and certified accordingly to "grade" a diamond based on multiple characteristics.

Bayesian GLM (model1.3). A symbolic description of the model to be fit. I picked the gaussian family.

MCMC Hierarchical Gaussian Linear Regression Model(model1.4). This function generates a sample ('simulates') from the posterior distribution of a linear regression model with Gaussian errors using Gibbs sampling. It has a multivariate Gaussian prior on the beta vector, and an inverse Gamma prior on the conditional error variance.

Empirical Results and Analysis

EDA. The web scraping tools used had some issues with it – mainly depending on the format of the website it had a hard time grabbing data. I collected data from six websites but visited over 50 in order to get six good data files that I could work with. I decided to use file number five which after doing some manual cleansing of duplicates, I resaved the file. I then read in the file, plotted a few of the characteristics then I took some shape and unique value counts. I used the unique value counts to create function to convert the feature values into label

encoded variables in a separate Python3 Jupyter notebook script(s). I also used python to remove the dollar signs ($) in the price column. I then took the data frame post label encoded variable conversion and outputted it to a comma delimited flat file (.txt) which I used with R. I also did a quick correlation plot. There was very little multicollinearity. I saw that polish did not matter and shape and color were weakly correlated. Carat, Symmetry and Cut seemed to matter most. I did not use the certification data since the 99.99% of it was of one type of certification – which leads me to believe this is the most popular (possibly trusted) certification in the diamond industry.

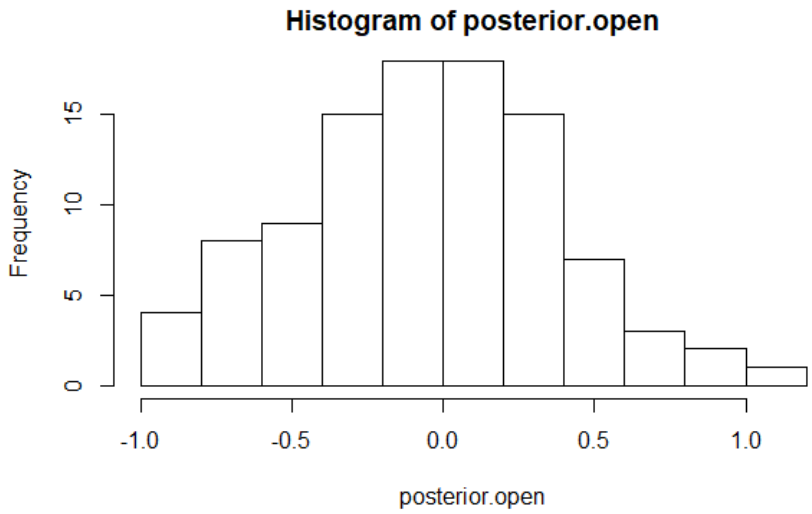
Selected Features and Encoding. I decided to focus on Shape, Carat, Cut, Color, Clarity, Polish and Symmetry for the data models to see how they influenced the distribution of a diamonds overall price. Price was the dependent variable (y). I label encoded the following features – Shape, Cut, Color, Clarity, Polish and Symmetry.

Conclusions

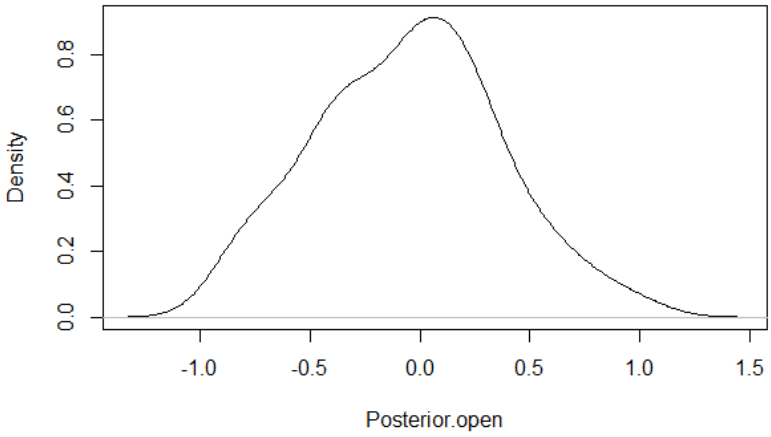
Model1.1. This was the traditional *lm* – Ordinary Least Squares (OLS) regression. This was my first baseline from which I will use to compare my other results.

Model 1.2. The same results from model1.1 are reflected in 1.2. This model specified a default Gaussian (normal family distribution). The primary benefit of the *glm* function is the ability to specify error distributions other than normal. This basically validates the model results from both models.

Model 1.3. The Bayesian GLM utilized the 'arm' package in R which contains the bayesglm function (Gelman, et al., 2010). I conducted the Bayesian GLM with the family = Gaussian and got a summary of the output. You will notice the specification of the prior mean, scale and degrees of freedom. Each 'family' of distributions requires specific prior specifications (reference package documentation in RStudio console). The output here matches up with model1.1 and 1.2. One of the benefits of the Bayesian perspective is that it allows us to make credible interval statements. The *bayesglm* function represents a short cut of the Bayesian approach to inference. Utilizing *bayesglm* I got a distribution of the 'simulates' which are used in place of an actual empirical distribution. Looking at the histogram of the it is almost a perfectly distributed Gaussian / normal distribution as anticipated.



The density plot below can be interpreted as there is a 95% probability that the true price value of the coefficients is between -0.808765 and 0.778626. In order to truly make a Bayesian inference about our coefficients you must go one step further. Meaning, I must do an additional step of re-creating the empirical distributions above in order to validate our results. I utilized the MCMCregress function as part of the 'MCMCpack' (Martin, Quinn, & Park, 2010). This provided me a Markov Chain Monte Carlo simulation of the distributions.



Model1.4. Based on the MCMC results I can say that there is a 95% probability that the true price value of the shape coefficient is between -0.8959 and 0.8308. This is a slightly wider distribution than the model1.3 distribution. The model1.3 distribution lies in between the model1.4 simulated distribution which tells me that

model1.3 was quite accurate. Hence if models 1.1, 1.2 and 1.3 all matched closely, and model 1.4 essentially validated the results of model1.3 then that tells me that in this instance all the models are accurate and reflect each other accurately. This ultimately means that the price for the diamonds5a.txt reflects the data – which means its consistent and in line with the experts rating. This means that this approach is validated as accurate.

References

Douc, R., Moulines, E., Priouret, P., & Soulier, P. (2018). *Markov Chains.* France : Springer.

Gelman, A., Pittau, M.G., & Su, Y., Yajima, M., Hill, J., & Zheng, T. (2010) Package 'arm'. Available at http://cran.r-project.org/web/packages/arm

Kruschke, J.K., (2015). *Doing Bayesian Data Analysis.* New York: Elsevier.

Martin, A.D., Quinn, K.M., & Park, J.H. (2010). Package 'MCMCpack'. Available at http://cran.r-project.org/web/packages/MCMCpack/MCMCpack.pdf

Nybert, S.O., (2019). *The Bayesian Way.* New Jersey: Wiley.