Predicting Diameter of Asteroids

Minho Choi Morgan Taylor

December 17, 2019

1 Introduction

One of the least explored areas in physics is astronomy. Relative to other fields, space is a new field that we still do not know and understand fully. However, with the development of technology, we are discovering new facts about astronomy and space recently. More information about space help us not only to improve our knowledge in chemistry, physics, and math but also to recognize potential dangers that may impact our lives. One of the space materials that contain both of these aspects is an asteroid. The asteroid impact can have serious damage to the earth, but at the same time it can deliver us valuable information about the outer space.

To understand the potential dangers of the asteroid impacts first, we can look at two past impacts. One of the most famous impacts is the impact that killed all the dinosaurs around 65 million years ago. The size of the asteroid is conjectured to have longest diameter of 10km, and the velocity of the asteroid is predicted to be around 25km per second [1]. More recent asteroid impact is meteor impact at Chelyabinsk, Russia [4]. It is verified that a near-Earth asteroid with diameter of 20m went through the atmosphere near Chelyabinsk with a speed of 19.16km per second. The meteor quickly burnt in the atmosphere creating an air burst, which yield around 500 kilotons of TNT energy, equaling to 26-33 times as much energy as the Hiroshima atomic bomb. Thus, the asteroid impacts have potential to influence our lives in a detrimental way.

In contrast, there are several reasons why the asteroid impacts can be important in science. First, meteorites, asteroids that make it to the surface of the Earth, are valuable tools in research of the cosmos since they provide us with materials to study from the outer space. Space research is often expensive and requires complex machinery and engineering for it to be possible. Meteorites are a simple, but useful, way to study compositions not commonly found on Earth. Asteroids can come from various places in the universe, and tracking an asteroid that is not imminently dangerous aids in its study when it does impact. If an asteroid does pose imminent danger, it is imperative that we know why, to create a safety plan. Therefore, it is important to examine if the asteroid pose imminent danger or not, and this issue is highly dependent on the diameter and velocity of the asteroid.

This leads us to our research question. Can we predict the diameter of an asteroid? If we can what is the most significant variable in predicting the diameter of an asteroid? Based on the articles on asteroid diameter [3, 5], we found that there are relationships between natural log of diameter and natural log of albedo, a ratio of the light received by a body to the light reflected by that body, and between natural log of diameter and absolute magnitude, a visual magnitude an observer would record if the asteroid were placed 1 Astronomical Unit (au) away, where 1 au is roughly the distance from Earth to the Sun. Hence, our research question is: **Are albedo (in natural log) and absolute magnitude significant variables in predicting the diameter (in natural log) of an asteroid?** In addition, another research question is: **Is there any other variable that has significance in predicting the diameter of an asteroid?** Specific conjectures and hypotheses that we tested are discussed in Section 2 with more details.

Our paper is organized as follows: Section 2 contains the conjectures and hypotheses that are tested throughout the paper, Section 3 contains an exploration of the data-set including pre-processing and visualization of the data-set, Section 4 contains a summary of the statistical methods that we used and the results that we found, and Section 5 contains a conclusion of the paper and a discussion of future research.

2 Conjectures and Hypotheses

2.1 Conjectures

We base our conjectures on two articles on asteroid diameter. First, Morbidelli et al. insist that the absolute magnitude can be calculated using natural log of albedo and natural log of diameter [3]. The formula that they derived is the following:

$$H = 2.5(6.244 - log(p) - 2log(D)) \tag{1}$$

where H is absolute magnitude, p is albedo, and D is diameter. In comparison, Shevchenko and Tedesco propose similar relationship [5]:

$$log(p) = 6.2472 - 2log(D) - 0.4H \tag{2}$$

where again p is albedo, D is diameter, and H is absolute magnitude. If we change both equations (1) and (2) to represent $log(D)$ in terms of H and $log(p)$, we get:

$$(1) : log(D) = 3.122 - 0.5log(p) - 0.2H$$

$$(2) : log(D) = 3.1236 - 0.5log(p) - 0.2H$$

Therefore, we examine if our data-set also follows this relationship. So, when selecting variables and selecting models we also add the natural log of each variable.

2.2 Hypotheses

First, we select variables and built a model using forward selection. We use the model built by forward selection to test significance of the regressors absolute magnitude and natural log of albedo (fortunately, you will see in Section 4 that the model contains both regressors). Our null hypothesis and alternative hypothesis to test the significance of the regressors absolute magnitude and natural log of albedo are following:

$$H_0 : \beta_H = \beta_p = 0$$

$$H_1 : \beta_H \neq 0 \text{ or } \beta_p \neq 0$$

where β_H is the slope parameter for absolute magnitude and β_p is the slope parameter for natural log of albedo. Furthermore, since we have conjectures on the slope parameters, we construct 95% confidence intervals for β_H, β_p, and intercept to see if the conjectured slope values are in between the intervals.

We employ the full model that includes other various regressors with the absolute magnitude and natural log of albedo. The following null hypothesis and alternative hypothesis test the significance of all regressors except the absolute magnitude and natural log of albedo:

$$H_0 : \beta_1 = \cdots = \beta_n = 0$$

$$H_1 : \text{At least one of } \beta_1, \cdots, \beta_n \text{ is not zero}$$

where $\beta_1, \cdots, \beta_n$ are the slope parameters for all regressors except the absolute magnitude and natural log of albedo. For all tests, we apply 5% significance level.

3 Data

3.1 Data-set Description

The data-set that we used is provided by Victor Basu through the Kaggle website [2]. He has clearly stated that the database that contains the asteroids data-set is officially maintained by Jet Propulsion Laboratory of California Institute of Technology which is an organization under NASA. Hence, all values in the data-set is collected and calculated by NASA's various tools. The original data-set has 786226 rows and 27 columns. The columns consist of orbital details of the asteroids, such details as semi-major axis, eccentricity, inclination, perihelion distance, aphelion distance, absolute magnitude, geo-metric albedo values, and lastly the response variable diameter of an asteroid.

3.2 Data-set Pre-processing

We follow some of the pre-processing procedures that Basu performed to remove unnecessary data [2]. First, 646785 rows out of 786226 rows do not contain the diameter values, which is the response variable, so we remove those 646785 rows. We then remove the rows with no albedo and absolute magnitude values, since they are two most important variables in predicting diameter of an asteroid according to our conjectures. Next, we want to focus on the near Earth asteroids only to examine our research questions more specifically, so we only extract the rows that are classified as near Earth asteroids. Then, we remove all the columns that contained any empty or missing value. Moreover, to apply statistical methods universally and more simply, we only choose the quantitative variables and remove categorical variables. Lastly, we remove variables that are not characteristics of an asteroid, such as number of observations used to obtain values. At the end, we get a data-set with 800 rows and 13 columns.

3.3 Data Visualization

For data visualization, we first examine the scatter plots of diameter against each variable. However, none of the scatter plots shows linear relationship. Then, we create the scatter plots of natural log of diameter against each variable. The scatter plot of natural log of diameter against absolute magnitude (H) clearly shows linear relationship:

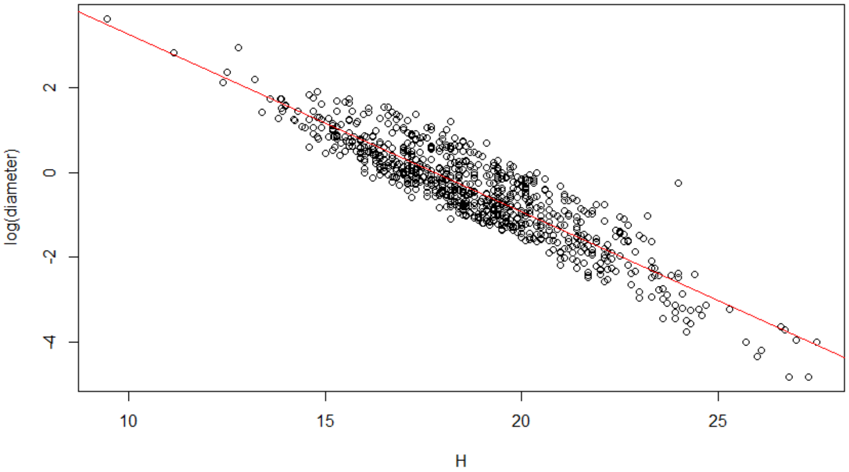

Figure 1: Scatter plot of log(diameter) against absolute magnitude (H)

Lastly, we also investigate the scatter plots of natural log of diameter against natural log of each variable. Excluding the absolute magnitude, the natural log of data-arc, a time period between the earliest and the latest observation to trace the asteroid's path, show a linear relationship with the natural log of diameter:

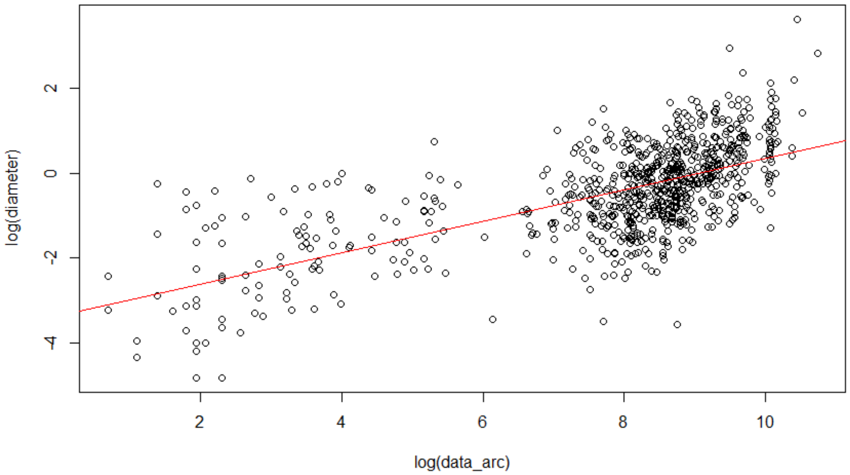

Figure 2: Scatter plot of log(diameter) against log(data-arc)

4 Statistical Methods and Results

As mentioned in Section 2, we first select variables and built a model using forward selection. We do not investigate the model built by backward elimination, because there are too large number of variables to test; there are 12 variables and if we include natural log of each variable, there are 24 variables to test for backward elimination. Hence, we mainly focus on the model built by forward selection. When running forward selection, we include natural log of each variable. To reduce multicollinearity issue, when one variable is selected during the procedure, we remove natural log of that variable from the procedure. Similarly, when natural log of one variable is selected during the procedure, we remove that variable from the procedure. We use F-statistics to make a decision on the selection and p-value of 0.1 as a cutoff.

The model built by forward selection is following:

$$log(D) = \beta_0 + \beta_H \times H + \beta_p \times log(p)$$

where D is diameter, H is absolute magnitude, and p is albedo. First, we simply verify that the model does not run into multicollinearity problem by examining the variance inflation factors. The variance inflation factor values are less than 5 for both regressors, so the model does not suffer multicollinearity issue. We now test the significance of the regressors H and $log(p)$. Using the analysis of variance (ANOVA) table, we get F-statistic of 30670 with (2, 797) degrees of freedom. This corresponds to p-value of 0. Since the p-value is less than 0.05, we have a strong evidence to reject the null hypothesis, so either absolute magnitude or natural log of albedo has significant relationship with natural log of diameter. Next, we construct confidence intervals for β_H, β_p, and the intercept. The least-squares regression equation for the model is following:

$$log(\hat{diameter}) = 7.152741 - 0.4572 \times H - 0.4875 \times log(albedo)$$

Hence, the 95% confidence intervals for:

- β_H is [-0.4609606, -0.4535054]
- β_p is [-0.4967544, -0.4782596]
- β_0 is [7.083021, 7.222461]

We can see that none of the conjectured slope parameters are in between the confidence intervals. However, the lower boundary of 95% confidence interval for β_p is close to -0.5, meaning that at least our β_p is close to the conjectured slope parameter of natural log of albedo.

Although the slope parameters are not close to conjectured values, we see that the model satisfies most of the linear model assumptions. First, the 3d scatter plot and adjusted R^2 value of the model show that the model has linear fit:

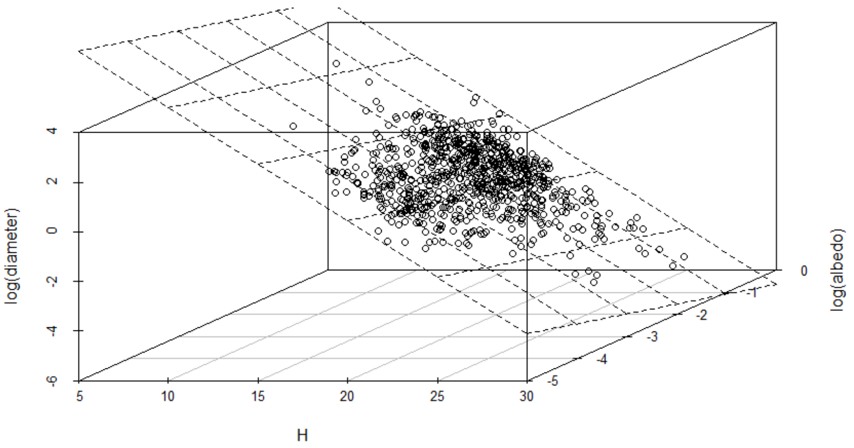

Figure 3: 3d scatter plot of log(diameter) against H and log(albedo)

The adjusted R^2 value of the model is 0.9871, which means that 98.71% of the variation in natural log of diameter is explained by the least-squares regression line.

Next, we look at the normal quantile plot of the residuals of the model:

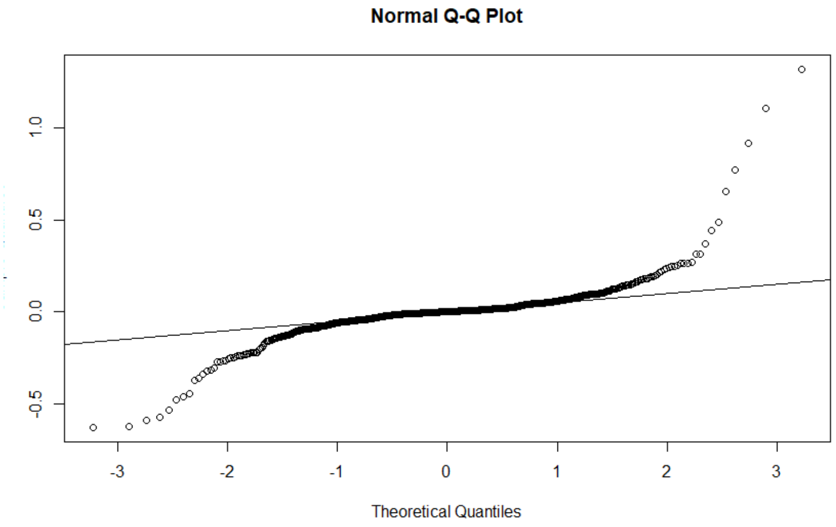

Figure 4: Normal quantile plot of the model

We see most of the points lie on the normal line, but there are clearly several outliers in the data points. Hence, in general, the model satisfies the assumption of Gaussian distribution although there are some outliers.

Lastly, we construct the residual plot of the studentized residuals:

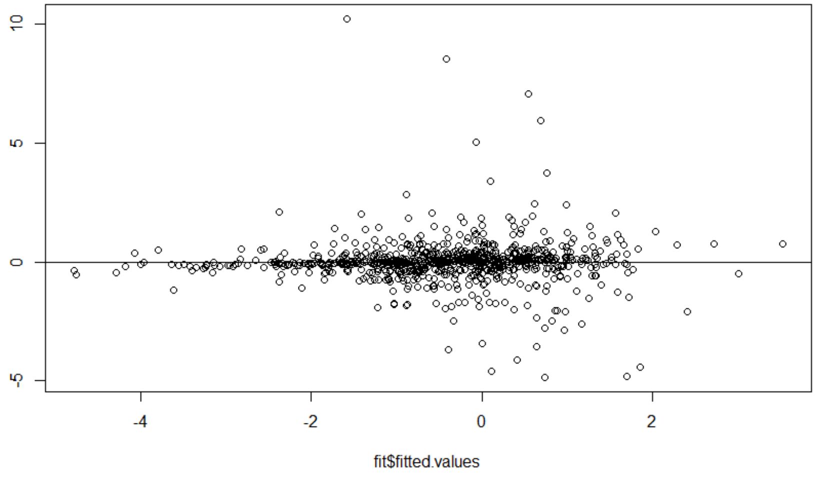

Figure 5: Residual plot of the studentized residuals of the model

We see most of the points positioned near the horizontal line of $y = 0$. However, due to some outliers in the plot, it is hard to say that the error terms have a common variance. Therefore, the model meets all linear model assumptions except homoscedasticity.

Secondly, we now answer our second research question: Is there any other variable that has significance in predicting the diameter of an asteroid? We consider a full model:

$$\begin{aligned} log(D) = {} & \beta_0 + \beta_H \times H + \beta_p \times log(p) + \beta_1 \times a + \beta_2 \times e + \beta_3 \times i + \beta_4 \times om \\ & + \beta_5 \times w + \beta_6 \times q + \beta_7 \times ad + \beta_8 \times per\text{-}y + \beta_9 \times data\text{-}arc + \beta_{10} \times moid \end{aligned}$$

We test the significance of all regressors except H and $log(p)$. Using the analysis of variance (ANOVA) table, we get F-statistic of 0.7222 with (9, 788) degrees of freedom. Note that the degrees of freedom is (9, 788), not (10, 788). This is because one of the regressors is simply a linearly combination of other regressors. The F-statistic corresponds to p-value of 0.6888. Since the p-value is greater than 0.05, we fail to reject the null hypothesis, so none of the regressors: $a, e, i, om, w, q, ad, per\text{-}y, data\text{-}arc$, and $moid$, has a significant relationship with natural log of diameter given that H and $log(p)$ are in the model.

In addition, we conduct multicollinearity diagnostics using variance inflation factors to the full model. Many regressors in the full model have variance inflation factor values over 5, and this indicates that the full model runs into multicollinearity problem. Hence, a subset model is better as long as there is no multicollinearity issue, and the model built by forward selection is one of the subset models of the full model that does not run into multicollinearity issue. Hence, we now compare the predictive ability of the model built by forward selection to that of the full model by calculating the Mallow's C_p value. For the model built by forward selection, we get Mallow's C_p value of 0.53445 which is less than 3 (the number of regressors plus 1). Therefore, we confirm that the model built by forward selection is a good model.

5 Conclusion

Our findings and results provide clear answers to our research questions. First, forward selection and the test on the significance of absolute magnitude and natural log of albedo show that there is a significant relationship either between natural log of diameter and absolute magnitude or between natural log of diameter and natural log of albedo. In addition, we recognize that although the model built by forward selection do not follow the conjectured slope parameter values, the model satisfies most of the linear model assumptions. Next, testing significance of other regressors in the full model show that there is no other variable that has significant relationship with natural log of diameter when absolute magnitude and natural log of albedo are in the model. Thus, we suggest complete answers to both of our research questions.

Our results imply that it is possible to predict diameter of an asteroid using absolute magnitude and albedo of an asteroid. Hence, we can predict the diameter of an asteroid, which is one of the two most important factors in determining the severeness of an asteroid impact. It is difficult to claim that we are confident in reproducing similar results using different data-sets. However, based on the high value of adjusted R^2 and linearity of the model built by forward selection, we are confident that absolute magnitude and albedo will have significance on predicting the diameter of asteroids.

For future research, we can investigate more on other ways to calculate the diameter of an asteroid such as using Kepler's Law. With different calculation methods, we believe that different variables may have more significance on predicting the diameter. Moreover, we can try to examine the values of slope parameters more in detail to figure out why our least-squares estimators differ to the conjectured slope parameter values. At last, we can explore the relationship between the diameter of an asteroid and Palermo scale, which assesses the risk of asteroids in the event that they impact the Earth.

References

1. Alvarez, Luis W. "Experimental evidence that an asteroid impact led to the extinction of many species 65 million years ago." *Proceedings of the National Academy of Sciences of the United States of America* 80, no. 2 (1983): 627.

2. Basu, Victor. "Prediction of Asteroid Diameter With the Help of Multi-Layer Perceptron Regressor." *Proceedings of ieeeforum International Conference* (2019): 1-5.

3. Morbidelli, Alesandro, R. Jedicke, W. F. Bottke, P. Michel, and E. F. Tedesco. "From magnitudes to diameters: The albedo distribution of near Earth objects and the Earth collision hazard." *Icarus* 158, no. 2 (2002): 329-342.

4. Reddy, Vishnu, Juan A. Sanchez, William F. Bottke, Edward A. Cloutis, Matthew RM Izawa, David P. O'Brien, Paul Mann et al. "Chelyabinsk meteorite explains unusual spectral properties of Baptistina Asteroid Family." *Icarus* 237 (2014): 116-130.

5. Shevchenko, Vasilij G., and Edward F. Tedesco. "Asteroid albedos deduced from stellar occultations." *Icarus* 184, no. 1 (2006): 211-220.