MATH 377: Model Fitting Assignment

Connor Gephart

22 October 2019

1 Introduction

This project is an introduction to model fitting for several sets of example data. The goal was to find a model that best fit the given data, and to draw up some graphs to defend that it is the best. There were three data sets given: Wire Stress, Pines, and Planets. Each set is detailed below.

2 Wire Stress

2.1 Scatter Plot

The Wire Stress data is the elongation, in 10^{-5} inches per inch, of a certain wire based upon how much stress, in pounds per inch squared, is applied to it. The data is shown in the following scatter plot, with the linear trend line applied. It can be seen that this function is quite appropriate for the data, with an R^2 value of 0.9989. A polynomial trend line could also be applied, but this runs the risk of overfitting the model for this specific set of experimental data.

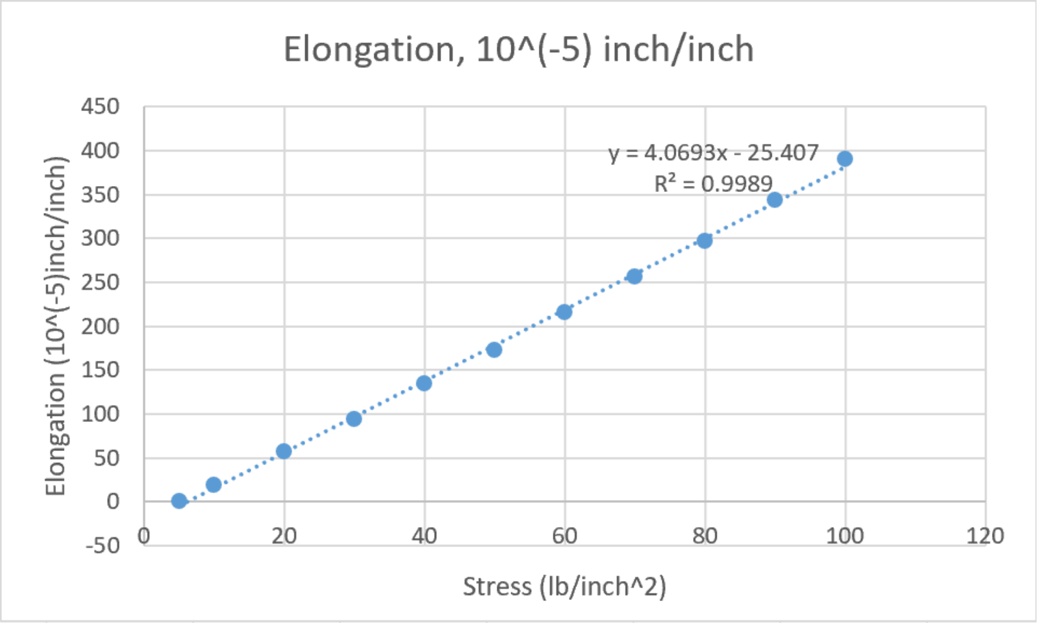

Figure 1: Wire stress v. elongation scatter plot

2.2 Residuals

However, it is often best to try to simplify the models to make them easier to deal with. In this case, the equation can be rounded to the following: $y = 4.07x - 25.4$. This can then be compared by plotting the residuals between this model and the actual data, shown below. Additionally, the relative error plot is shown. The values in this graph were found by taking the residuals and dividing by the actual measurement. The relative error for the first data point is omitted, because it is known experimentally to be 0, but the model output was $5.5 * 10^{-5}$ inches per inch. The largest relative error was approximately 19% for the second data point. However, no other data point approaches this high, as the second largest relative error is -2.9%.

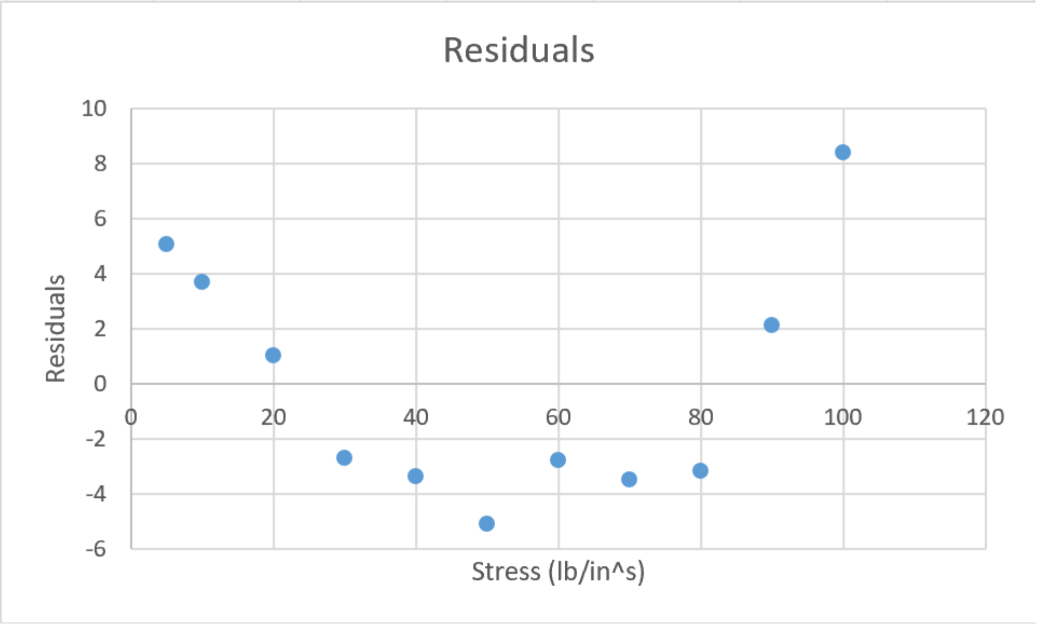

Figure 2: wire stress residuals plot

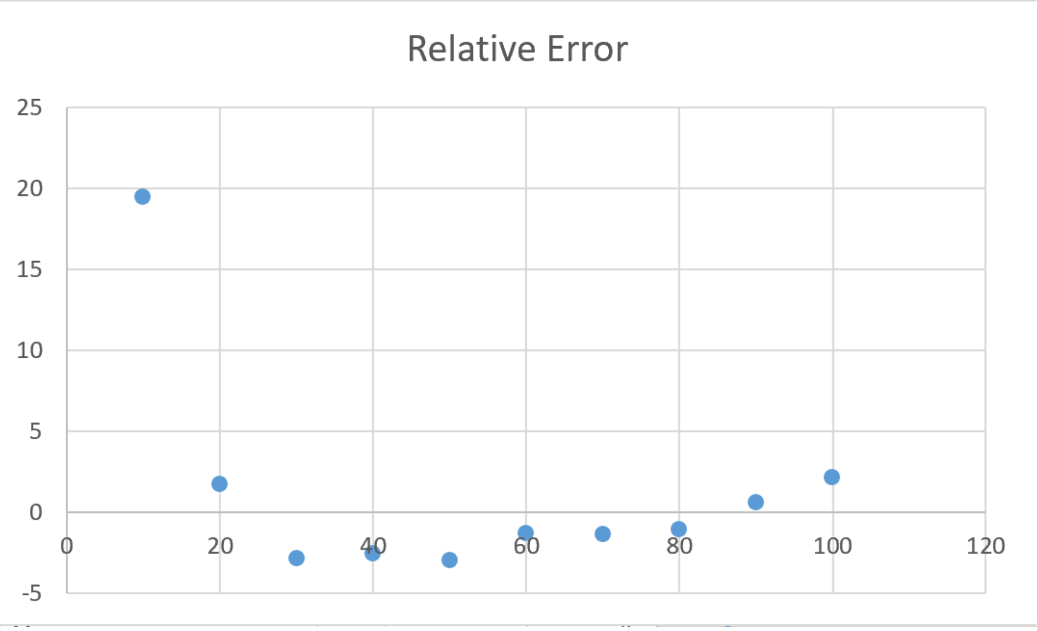

Figure 3: Wire stress relative error plot

3 Pines

3.1 Scatter Plot

The next data set compares the diameter, in inches, of a pine tree to the volume, in board feet per 10, of wood. The data has a power trend line applied below, with an R^2 value of 0.9858.

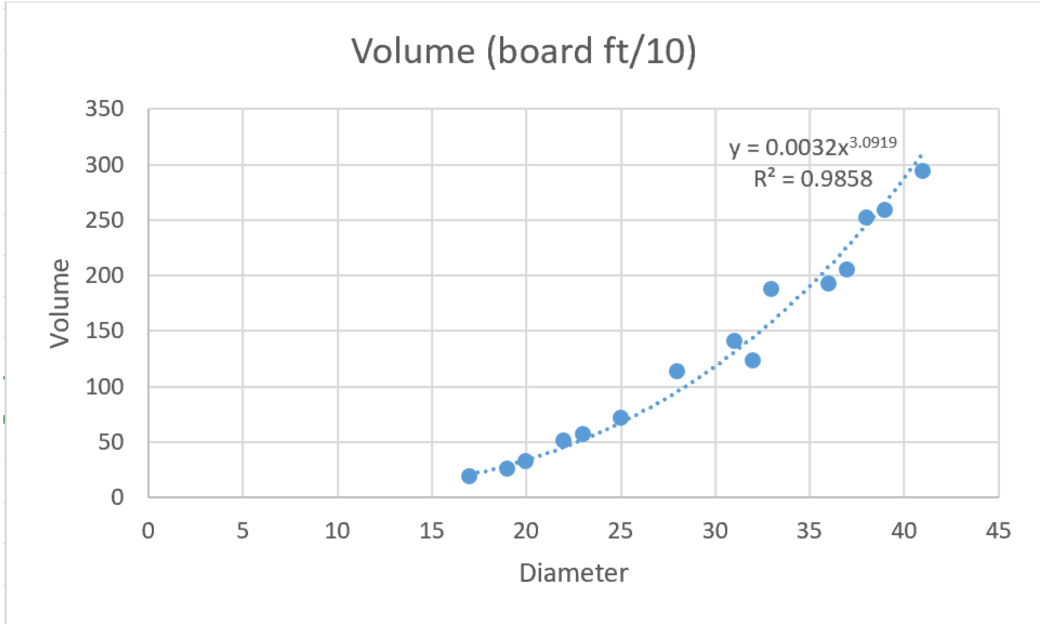

Figure 4: Pine diameter v. volume scatter plot

3.2 Residuals

The practical application of this model would be for lumber workers who wish to know if a tree should be cut down based on its diameter. The equation given in the above graph is $y = 0.0032x^{3.0919}$, which is not the easiest formula to remember or use. The moderately simplified equation $y = 0.0032x^{3.1}$ is used to generate the residual plot below. The coefficient of 0.0032 can be understood as applying the formula to large quantities of wood. This model is slightly worse for fit than the one above, as it has a largest relative error of approximately -20%, and the rest of the values fluxuate more than in the previous example. This can be seen by the relative error chart below.

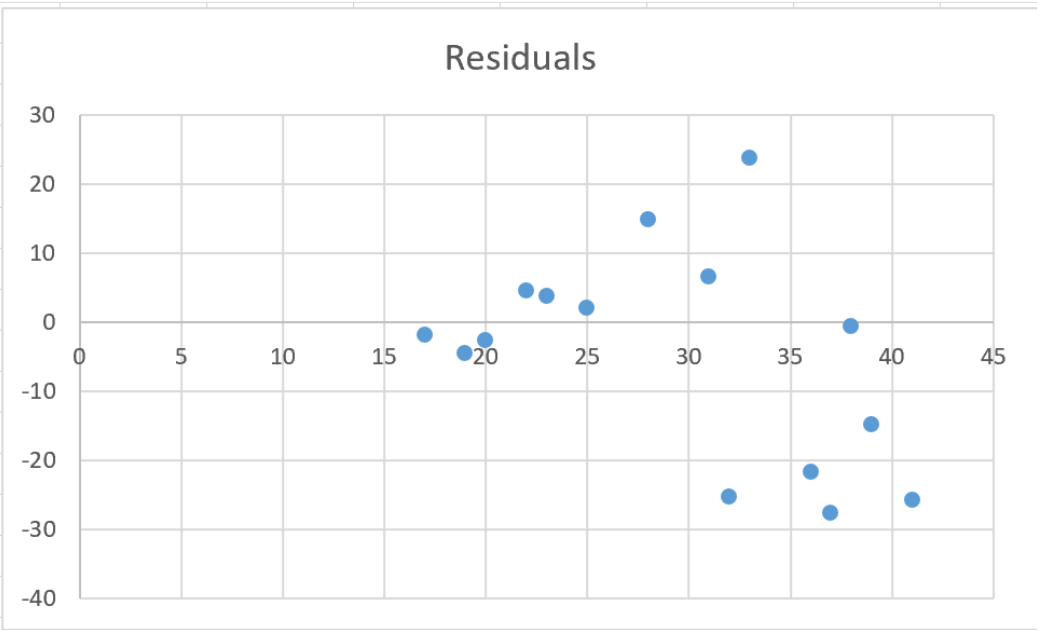

Figure 5: Pine residuals plot

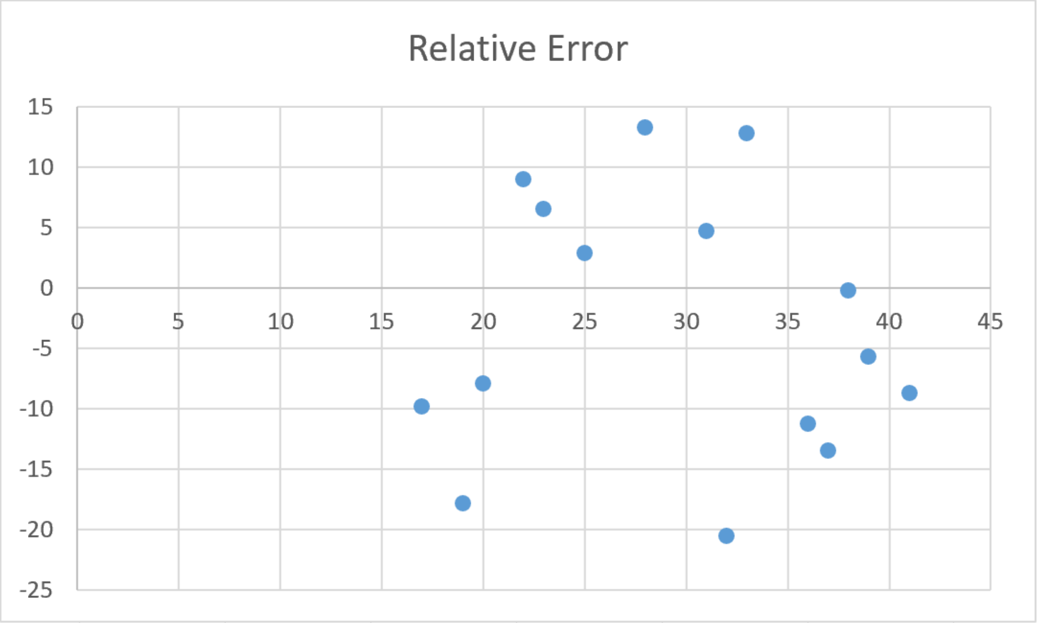

Figure 6: Pine relative error plot

3.3 Question

In the project description, the following question was posed: If your model is applied to 1000 mature thick trees, what is a possible error of the estimate of total lumber volume? This model was not nearly as accurate as the other two in this project, and this error would become more pronounced with a larger sample size. Additionally, most of the error is an overestimate, meaning the actual volume will be less than predicted. This will result in losses for the lumber company.

4 Planets

The goal of this set of data is to derive one of Kepler's Laws about the length of a planet's period, in days, and the minimum distance from the sun, in millions of kilometers. Similar to the pines data, a power trend line is applied to the data, but this time with $R^2 = 1$. However, comparing the equation of this line, $y = 2.9403x^{0.6659}$ to the actual equation, $y = 3x^{\frac{2}{3}}$ shows this to be an example of overfitting. The original scatter plot of the data, and then the scatter plot of the residuals using the actual equation are shown below. The relative error chart is also shown below. This model is a pretty good fit as the largest is about 3.5%, and all the values are less than that.

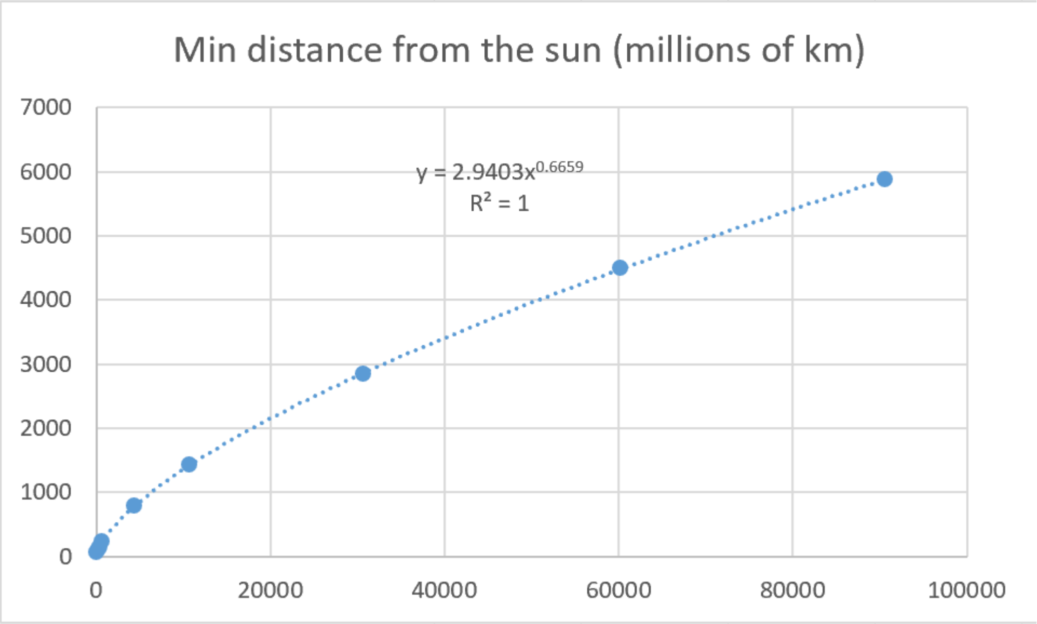

Figure 7: Planets period v. minimum distance from sun scatter plot

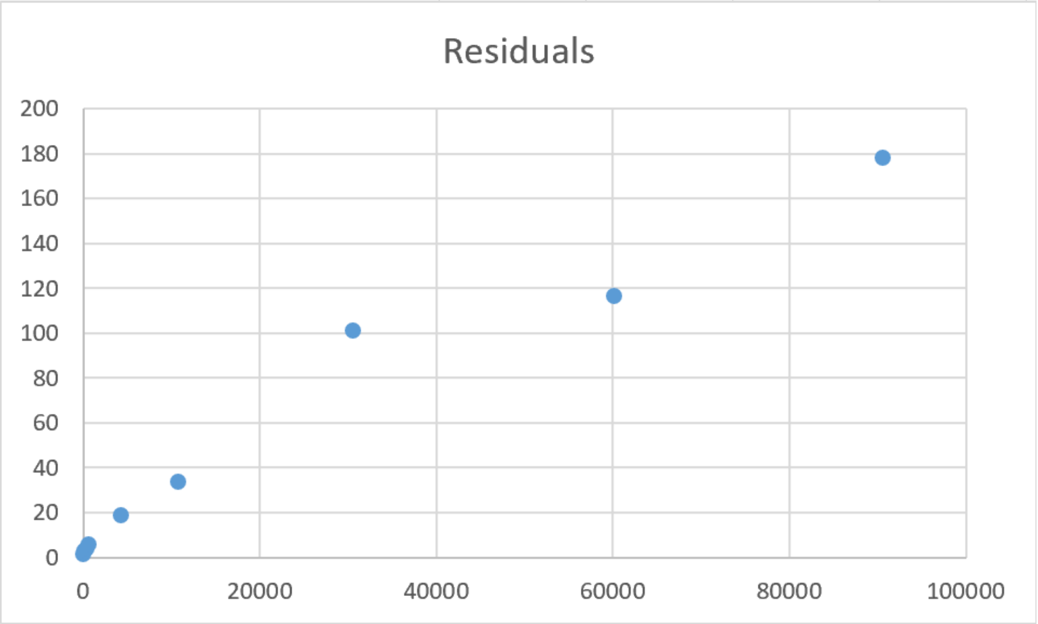

Figure 8: Planets residuals plot

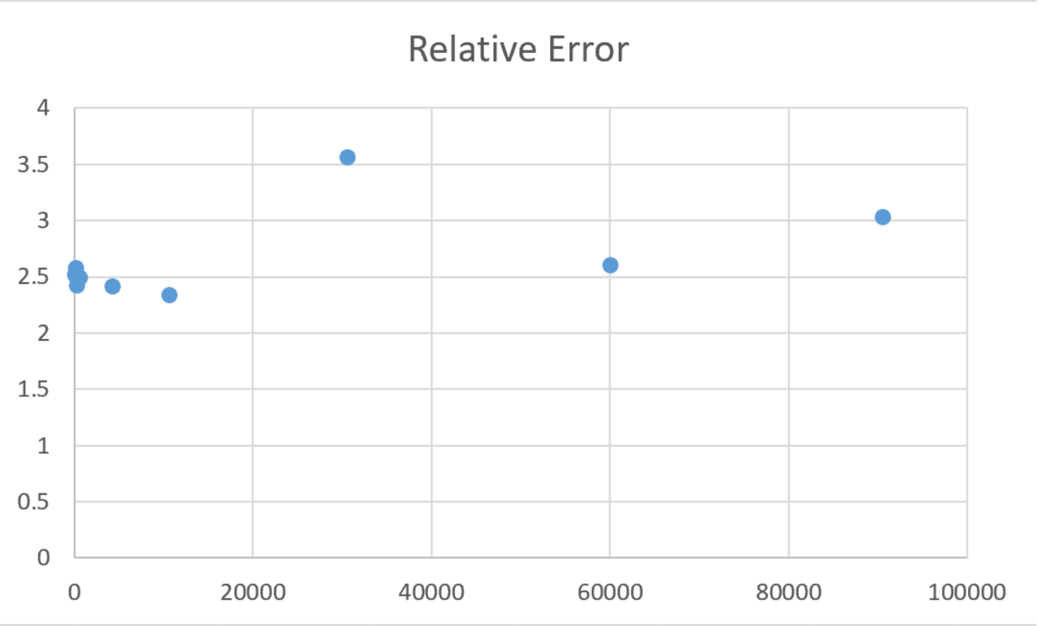

Figure 9: Planets relative error plot