Holcombe Department of Electrical and Computer Engineering
Clemson University

Lab 1: Model Fitting

August 31, 2021

Tim Driscoll

Introduction:

For this lab, code was developed in MATLAB to fit models to a series of different raw data sets. This lab utilized the normal equation to determine various unknowns from a function consisting of a linear combination of terms.

Part One:

A 2-dimensional linear line was used for part one to model five data points represented as (x,y) pairs. This model chose was a straight line, resulting in the need to determine two unknowns, both the slope and the y-intercept. For this data set and model, the value of M would be two. The matrices needed to solve the normal equation would be as follows:

$$(1)\quad A = \begin{bmatrix} 5 & 1 \\ 6 & 1 \\ 7 & 1 \\ 8 & 1 \\ 9 & 1 \end{bmatrix}$$

$$(2)\quad x = \begin{bmatrix} Slope\ (m) \\ y - intercept\ (b) \end{bmatrix}$$

$$(3)\quad \mathrm{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 5 \end{bmatrix}$$

Using matrices one, and three equation four could be used to solve for the two unknown values represented by matrix three, the x matrix. Equation three is represented by

$$(4)\quad x = (A^T A)^{-1} A^T b$$

The values of x were solved for in MATLAB and produced the results, $\begin{bmatrix} 1.000 \\ -4.600 \end{bmatrix}$.

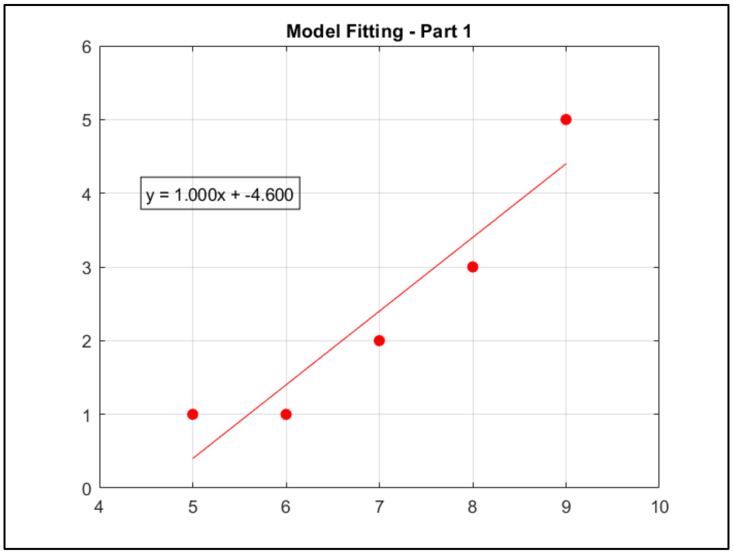

Image 1. Graph Displaying Part 1 Model

The values from x could then be plugged into the equation of a straight line to produce the desired model. Image one above is a graph displaying the raw data plotted as red circles and the model plotted as a red line. The model equation is also shown in the black box on the graph.

Part Two:

Part two of the lab involved the inclusion of an additional point to the data set that was used in part one of the lab. The point that was added to the data set was (8,14). This point was a relatively large outlier in the y-direction of the graph. The other point at x equals 8 was (8,3), creating a large disparity of 11 in the y direction. For part two the following matrices were used to solve the normal equation:

$$\text{(5)} \quad A = \begin{bmatrix} 5 & 1 \\ 6 & 1 \\ 7 & 1 \\ 8 & 1 \\ 8 & 1 \\ 9 & 1 \end{bmatrix}$$

$$\text{(6)} \quad x = \begin{bmatrix} Slope\ (m) \\ y - intercept\ (b) \end{bmatrix}$$

$$\text{(7)} \quad \mathrm{b} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 14 \\ 5 \end{bmatrix}$$

Using matrices five, and seven equation four could be used again to solve for the two unknown values represented by matrix six, the x matrix. When the values of x were solved for in MATLAB it produced the results, $\begin{bmatrix} 1.815 \\ -8.677 \end{bmatrix}$. The plotted model and data can be viewed in image two below.

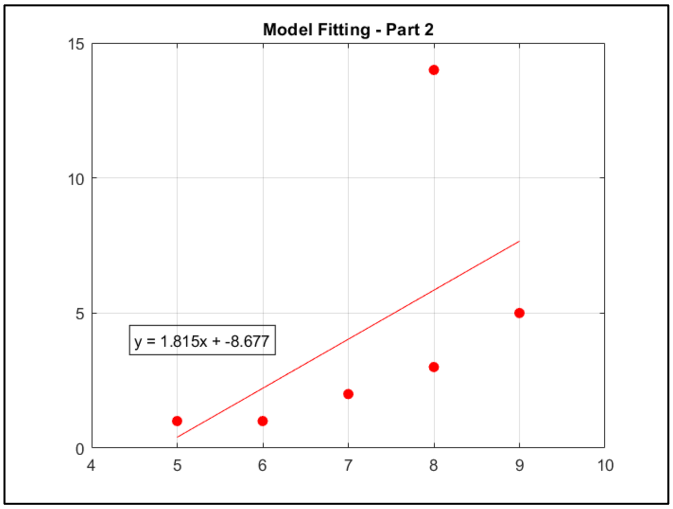

Image 2. Graph Displaying Part 2 Model

As observed when comparing images one and two the model is negatively impacted by the insertion of the point (8,14) into the data set. The slope of the line was forced to increase, and the y-intercept was shifted down by almost a factor of 2x in the second model. The additional point caused the model to rotate to the left and largely increase the residual for each point in the data set.

Part Three:

For the third part of the lab data was observed for 3,398 meals eaten by 83 different people. For this data the relationship between bites taken and kilocalories per bite was observed. The relationship was plotted and observed before a series of models were tested to determine the best fitting. The raw data is plotted in image 3 below and was analyzed to determine which model would fit the best.

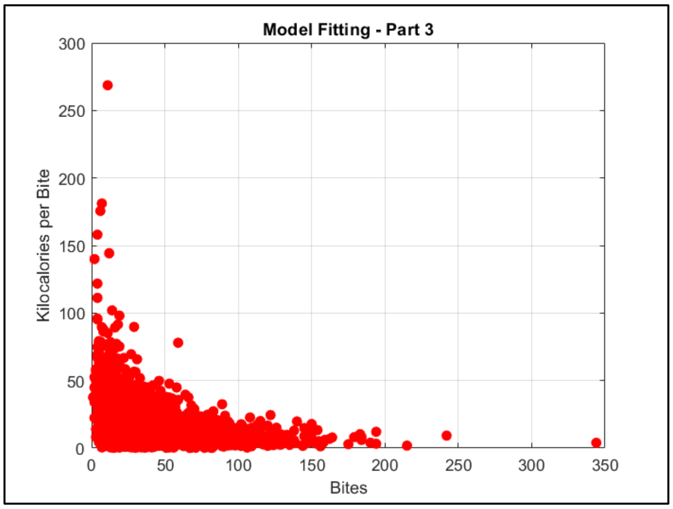

Image 3. Graph Displaying Part 3 Data

The first thing that I noticed when viewing the data was that there was an inverse relationship between the number of bites and the kilocalories per bite. As the number of bites increased the kilocalories would decrease. This observed relationship did not appear to be directly proportional, so variations of the inverse power model and exponential model were tested and observed for how well they fit the data. Some models that were tested included $y = a\frac{1}{x} + b$ and $y = ae^{-x} + b$, but I found that $y = a\frac{1}{\sqrt{x}} + b$ was the best model for the data. For this model the following matrices were constructed to utilize equation 4:

$$(8) \quad A = \begin{bmatrix} \frac{1}{\sqrt{x_1}} & 1 \\ \frac{1}{\sqrt{x_2}} & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \frac{1}{\sqrt{x_N}} & 1 \end{bmatrix}$$

$$(9) \quad x = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$(10) \quad \mathrm{b} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix}$$

In matrices eight and ten respectively x_1 through x_N and y_1 through y_N represent all the provided x and y points in the given data set. For the data set the value of N was 3398, thus matrix A was 3398x2 and matrix b was 3398x1. When the values of x were solved for in MATLAB using the normal equation it produced the results, $\begin{bmatrix} 82.6404 \\ 0.4003 \end{bmatrix}$. The plotted model and raw data can be viewed in image four below.

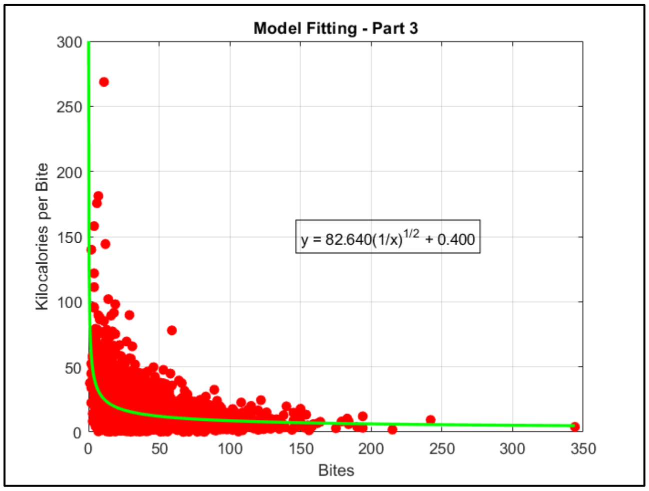

Image 4. Graph Displaying Part 3 Model

In image 4 above the raw data is again displayed by the red circles, the fitted model is displayed by the green line to add contrast, and the black box displays the equation of the model. The model does a good job fitting to the entire span of the data and showing the previously described inverse relationship. Although this is true the model could more accurately represent the cluster of points from 0 to 150 bites with the removal of outliers in the data, but those outliers represent the nature of an inverse power model. These outliers could include the group of points where the kilocalories per bite ranges over 100. Similar to the differences seen between part 1 and part 2 outliers can drastically alter the fitting of the model and increase the value of the residual at each point. Viewing the data in its entirety the chosen model most accurately represents the data and reduces the value of the residual at each point when compared to the other models that were tested.