Tracking Systems
Nonlinear Regression

Mayukh Sattiraju
msattir@clemson.edu

ECE 854
Lab 3 Report

September 28, 2017

1 Introduction

This assignment attempts to fit models to a given set of data. The models proposed increase in complexity in an attempt to fit the model better. The fit is measured by defining an error, chi-squared in this case. The model coefficients are determined by solving a closed form of the normal equation constructed by the given data.

Iterative methods such as Gradient Decent etc, can also be used to solve this problem. But in cases where it is not mathematically intensive to compute, the closed form solution can provide results almost instantaneously.

The data used here is the data collected for 3398 meals eaten by 83 people. The data includes the number of bites taken in a meal, and the number of kcal consumed in that meal. The objective here is to find a model that best describes the number of bites taken and the number of kcal consumed in that meal. The paper attempts to first fit the model using linear models and then explores more complex models to better capture the data. For all models proposed here, we'll use the normal equation to derive the model coefficients.

2 Closed form Normal Equation

All model coefficients determined here are generated by solving the closed form of a Normal Equation. To set up the matrices, we would need the data points, and the equation of the model we intend to fit.

The expression for the unknowns (model coefficients here) is derived by this equation,

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

where the matrices A and b are constructed from the data points and the 'model' we intend to employ.

2.1 Linear Model Fitting

Here we attempt to find the best line that can fit the model. The 'best' line is determined by the line that has the least chi-squared error with the given data.

The model we intend to fit is given by,

$$y = ax + b \tag{2}$$

where a and b are the (best-fit) unknowns we intend to find.

For part A of the assignment we consider 5 data points, and we construct A and b matrices for this example as below.

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \\ x_5 & 1 \end{bmatrix}, b = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \tag{3}$$

Plugging the data points in Eq. 3 and using matrices A and b from Eq. 3 in Eq. 1 we can solve for the unknowns of a and b and plug it into Eq. 2 to get the best-fit linear model for the data.

The model obtained for the given set of data is,

$$y = 1.0x - 4.6 \tag{4}$$

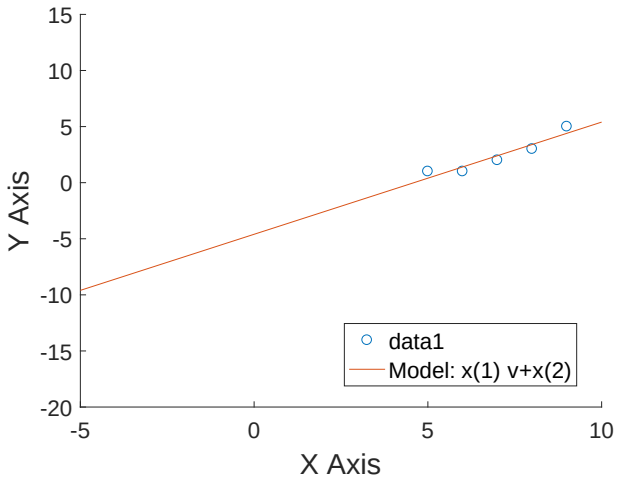

Figure 1: Plot for Part A

The plot of the data and the above line fitted to the data is shown in Figure 1

2.1.1 Effect of Outliers

Here we again try to find the best fit line but this time the dataset includes a new data point.

Here also we employ Eq. 1 to determine the unknowns.

The new model, now including the new data point, is given as

$$y = 1.8x - 8.67 \tag{5}$$

The plot for this updated model with its data points is show in Figure 2

The new added point (outlier) pulls the line towards the new point and away from the group. This is because the new average lies slight;y offset from the previous average.

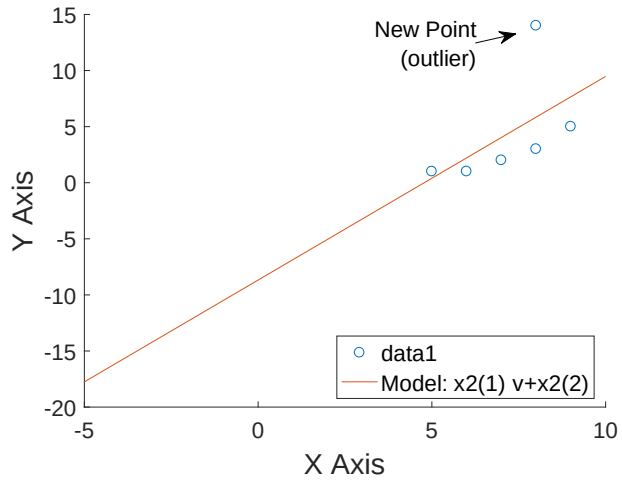

Figure 2: Plot for Part B

2.2 Non Linear Models

For this part we load the provided dataset of 3398 data points. Here we intend to find the model that best describes the relationship between the number of bites taken by a test subject and the number of kcal of food consumed in that meal.

Here is a distribution of the provided data,

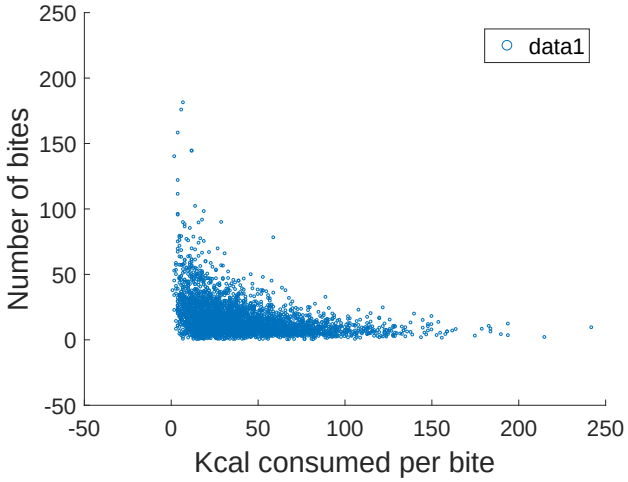

Figure 3: Distribution of Number of Bites vs Kcal per bite

To find a best-fit model we start with a model of complexity 1 (degree 1) and then increase the degree. To ensure that each successive model is better than the previous one, we compute the loss and print it. Theoretically, this loss depicts how well the model fits the data (smaller the number a better fit the model is) and should decrease as we increase the model complexity. This probably is because now have more dimensions to represent the data.

2.2.1 Model of Degree 1

The weights obtained for solving a closed form normal equation built using the data points as shown above and matrices constructed to fit a model of degree 1 is given in Eq 6

$$y = -0.17x + 23.4 \tag{6}$$

The loss obtained for this model is: $2.077e + 03$

The plot of the data and the model is shown in Fig 4

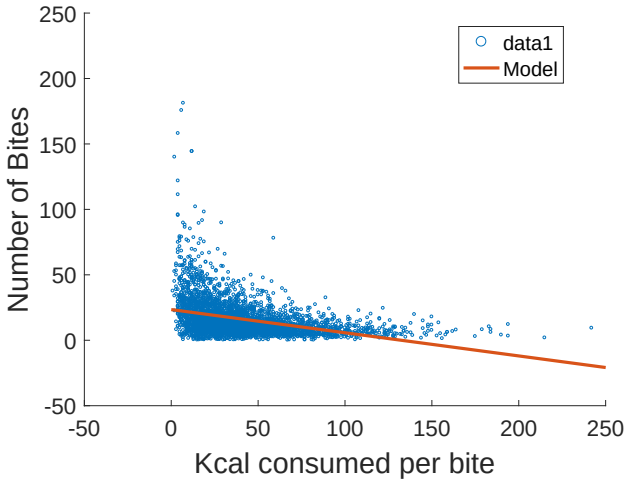

Figure 4: Model of Degree 1

The line attempts to divide the data into two equal halves, but we can do better by increasing the complexity.

2.2.2 Model of Degree 2

Solving for the weights using a closed form normal equation yields Eq 7

$$y = 0.001x^2 - 0.36x + 27.51 \tag{7}$$

The loss obtained for this model is: $1.99e + 03$

The plot of the data and the model is shown in Fig 5

This model is better theoretically (as it has a lesser loss than the previous model) and also looks like the parabola that the data might represent.

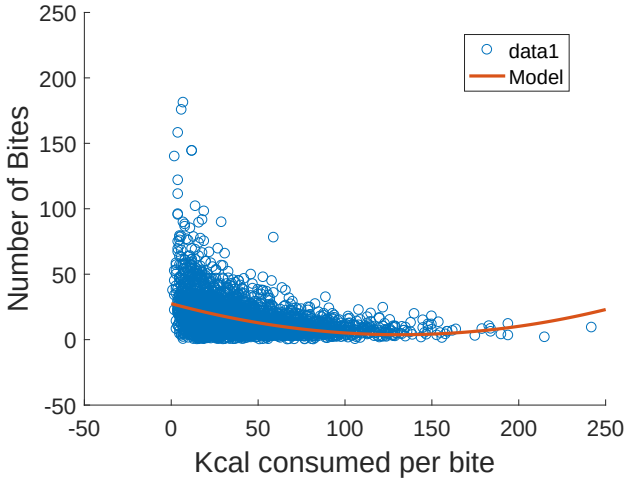

Figure 5: Model of Degree 2

2.2.3 Model of Degree 3

Solving for the weights using a closed form normal equation yields Eq 8

$$y = 0.0000009x^3 + 0.0045x^2 - 0.65x + 31.52 \tag{8}$$

The loss obtained for this model is: $1.931 + 03$
The plot of the data and the model is shown in Fig 6

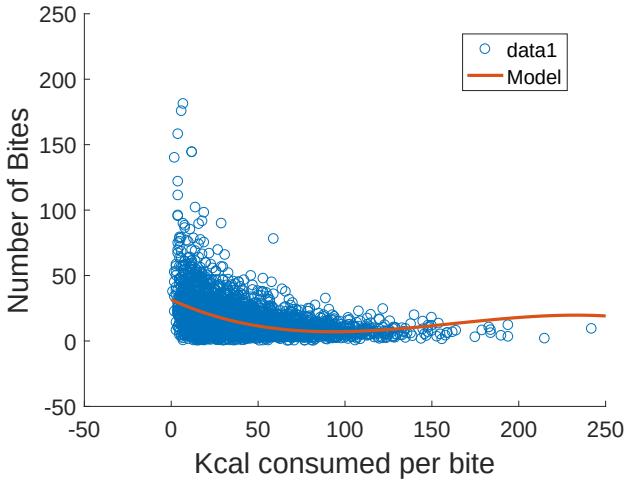

Figure 6: Model of Degree 3

This model also has a lesser loss than the previous model.

2.2.4 Model of Degree 4

Solving for the weights using a closed form normal equation yields Eq 9

$$y = 0.0000000x^4 - 0.000058x^3 + 0.011x^2 - 0.95x + 35.52 \tag{9}$$

The loss obtained for this model is: $1.888 + 03$

The plot of the data and the model is shown in Fig 7

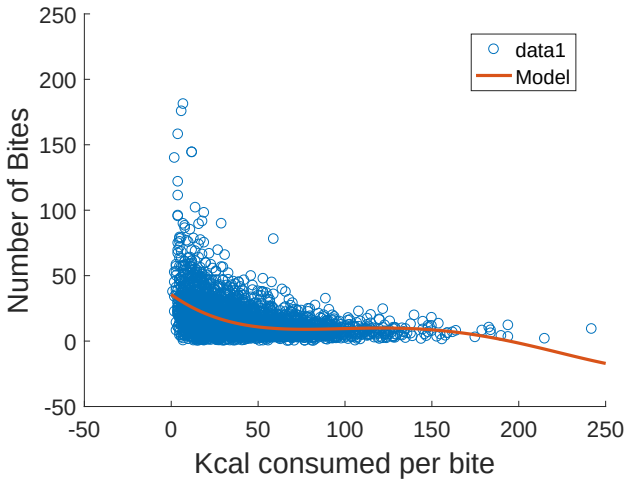

Figure 7: Model of Degree 4

Though this model also has a lesser loss than the previous model, the coefficient of the highest power is approaching zero. Hence, though the complexity is increasing the model has saturated at this loss. There would not be a significant reduction in the loss as we increase the complexity.

3 Results

Various models were generated for the given data. For part A and B of the assignment the models obtained are given in Table 1

Part	Model	Remarks
Part A	$y = 1.0x - 4.6$	5 data points
Part B	$y = 1.8x - 8.67$	Includes outlier

Table 1: Models for Part A and B

The models for the 3398 data values, each of increasing complexity is shown in Table 2. Here the last column shows the loss (chi-squared) for the models.

Complexity	Model	Loss
Degree 1	$y = -0.17x + 23.4$	$2.077e + 03$
Degree 2	$y = 0.001x^2 - 0.36x + 27.51$	$1.99e + 03$
Degree 3	$y = 0.0000009x^3 + 0.0045x^2 - 0.65x + 31.52$	$1.931e + 03$
Degree 4	$y = 0.0000000x^4 - 0.000058x^3 + 0.011x^2 - 0.95x + 35.52$	$1.888e + 03$

Table 2: Models for fitting data of Part C

4 Conclusion

For part A and B of the assignment we see that addition of a new point (outlier) has a direct impact on the model's weights. This method might be susceptible to random noises and outliers. Thus, it might be beneficial to filter the data (to remove extreme noises) before generating weights using this approach.

For part C of the assignment, Table 2 shows that increasing the complexity better fits the data and the loss also decreases. We can also visually see that the parabola seems to better describe the data than the line. But, this method also has limitations. We can see that as the degree increases the coefficient of the highest degree vanishes towards 0. This saturates the dimensions (degrees) the model uses to describe the data. Practically, we can not increase the complexity to infinity, the model would saturate before that. To overcome this, models outside polynomials can be explored, this can also extend to models which are linear in the unknowns.