Report for lab 6

Introduction

In this lab we learned how to use the given data to find an underlying relationship between them. The method of linear least square parameter fitting was used to solve linear parameter fitting problem. As to the non-linear problem, the nonlinear regression with the Gauss Newton method was used. The model which fit the data was plotted together with sample point as a comparison and the student t test was performed to test the model.

Methodology

Firstly, the linear least square approach was used to solve the linear problem. Detailed mathematical formula used in this method is in lab assignment file. A c++ program-"regression.cpp" was written for the calculation of a1 and a0. In order to solve

$$A^TAx=A^Tb$$

the lapack library was used. A library I wrought previously"matrix.h" was also used for the matrix calculation.

The statistical approach was used to perform a T test with null hypothesis to be $a1=a_t$ where at was specified as 0, which is not equal to the value I got. After the T value was calculated based on the given formula, the threshold was checked on the t table with degrees of freedom to be 98. Then the decision was made to reject the null hypothesis.

As to the second question, the Gauss Newton method was used to fit the model with data. For each step, the a0, a1, a2, a3 was updated by the given formula. The SSE value was used to denote the size of residual which is calculated by:

$$SSE^n= \sum(Y_i-f(X_i, b^n)^2)$$

If the difference of SSE between two consecutive steps is larger than tolerance(setting to be 0.0001), the Gauss Newton iteration stopped. Two different group of initial guess was used to justify the method.

The result of model which fits the data was then plotted together with the data points.

Result

For question 1, the a0 and a1 have been calculated with the value 16.404 and 2.009. The model y=a1x+a0 was plotted in Figure 1 together with sample data. We can see the sample dots are evenly distributed on both sides of the straight line. The T value calculated is 83.473, which is far above the threshold (1.98 if the probability level is 0.05, degrees of freedom is 98, got from online t table). So we can reject the null hypothesis that $a1= =a_t$.

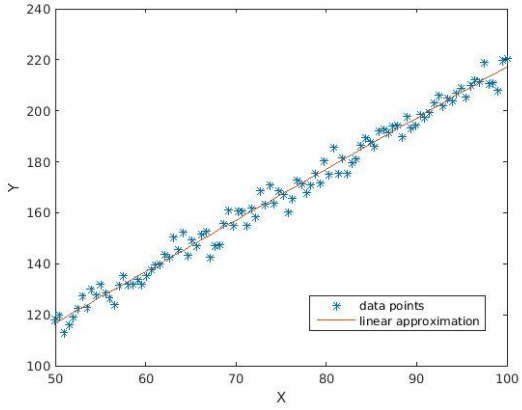

Figure 1: the sample dots and the fitted linear function: y=2.009x+16.404

For question 2, the value of a0, a1, a2, a3 were calculated with two initial guesses(all 1, or all 1000). Both of the guesses leads to the same output:
a0=10.7314599
a1=2.2448251
a2=-3.6905773
a3=0.2384493

in just one step of iteration. The result model was shown in Figure 2 with the data points. We can see the sample dots are evenly distributed on both sides of the curve.

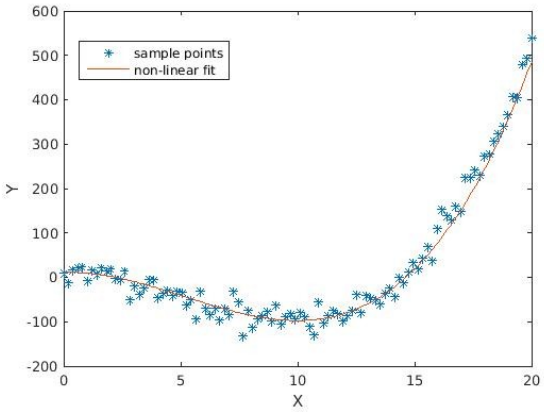

Figure 2: The sample dots and the fitted nonlinear function.

Discussion

For question 1, the T value I got was 83.473. According to the online t value table, if the p value threshold is 0.05 and the degrees of freedom is 90, the t value should be larger than 1.98 or smaller than -1.98 to reject the null hypothesis. Since the T value in this simulation is far larger than 1.98. The null hypothesis that has been rejected, which suggest that a1 is not equal to 0. It means that the value y is not independent from x.

For question 2, the two initial guesses all leads to the same result, together with the fact that the sample dots are evenly distributed on both sides of the curve, we can see that the Gauss Newton iteration is successfully implemented. Also it only takes one step for this method to reach the stopping criteria. This is because the parameters a0 to a3 is just the coefficients of each terms, in fact, this problem can be treated as a linear problem by treating x, x^2, x^3 as three separate variables.