**Computational Sciences and Informatics Course 690**

**Assignment 3**

**Jericho McLeod**

**Introduction:**

The following problems and solutions reflect studies designed to reinforce course lectures pertaining to the fitting and use of linear systems to solve problems, as well as linear straight line and polynomial regression. Emphasis has been placed on comparisons of methods related to the introduction of error based upon the direction and type of regression.

**Methods:**

The assigned problems are from Scientific Computing, by Michael T. Heath, are Chapter 3’s first and third exercises, as well as a selection from Numerical Methods for Engineers, by Steven Chapra and Raymond Canale, specifically; Chapter 17’s third and sixth problems. These will henceforth be referred to respectively as problems one through four.

Problem one contained several components; first, the set up an overdetermined system with which to solve the problem, then to determine if it was consistent, and if not consistent, to solve it in all possible solvable combinations and determine if a particular result was more desirable. The final step was to solve the system using a normal equation and compare results to the chose combination from the previous step.

The first stage of this required fitting a set of numeric data given in the form Y = x(sub 1) + x(sub 2)t to an overdetermined matrix system in the form Ax = b. This is trivially achieved by creating the matrix A from the t variable at exponents 0 and 1 and the b vector from the Y values, then concatenating the two systems. The next step was completed using row calculations to reduce the calculation to a form that could not be solved, at which point it was shown to be inconsistent, and then creating the possible pairs of a system with three rows. The third component of this problem, solving normal equations, used the formulas below.

Or, in more common notation:

The results of this equation were then verified with a brief python script relayed in Appendix 1.

The subsequent problem required the same formulas for solving a normal equation, followed by Cholesky Factorization, which states that ATA = LLT can be used to obtain ATAx = ATb by multiplying A by its Transpose, AT, to obtain LLT, then solving Lz=ATb for z, then LTx=z for x, which then allows the problem to be solved as a linear least squares problem. Since least squares have previously been solved, this was achieved using the python code given in Appendix 2, while an algorithm to solve the Cholesky decomposition is found in Appendix 3. The matrix resulting from the python code was then used to solve the least squares problem in the previously described fashion.

The third problem fit a least squares regression to data presented as two vectors, and in addition to the slope an intercept to find the standard error of the estimate as well as the correlation coefficient, then to repeat the previous steps with the variables reversed and interpret the results.

The least squares problem was solved using the previously provided methods. Standard error was obtained using the following formula:

where

The correlation coefficient was used using Pearson’s Correlation, given as follows:

(Kent State University, 2018)

Which, when substituting the formula for covariance and variance for a more complete formula, is:

The fourth problem primarily used the same methods but added a new component in regression the data to a polynomial rather than a straight line. The form for this is given as:

Given that the problem requested that the problem be fitted to a parabola only a second order equation was used, the familiar linear form:

Can be obtained using the matrix system:

From which point a solution introduces no additional methods, although the python script shown in Appendix 4 was used for the exponential summations and to solve the linear system

**Results:**

**Problem 1 (Heath 3.1)**

If a vertical beam has a downward force at its lower end, the amount by which it stretches will be proportional to the magnitude of the force. Thus, the total length of Y of the beam is given by the equation:

Y = x(sub 1) + x(sub 2)t,

Where x1 is its original length, t is the force applied, and x2 is the proportionality constant. Suppose the following measurements are taken:

T: 10, 15, 20

Y: 11.6, 11.85, 12.25

A: Set up an overdetermined 3x2 system of linear equations corresponding to the data collected.

A:

And

B: Is this system consistent? If not, compute each possible pair of values for x1 and x2 obtained by selecting any two of the equations in the system. Is there a reason to prefer any one of these results?

Subtracting row 1 from rows 2 and 3:

Subtracting 2x Row 2 from Row 3:

0x + 0xT = 0.15 is not mathematically possible: this system is inconsistent.

Possible pairs that could be solved:

(1,2), (1,3), (2,3)

Row (1,2) = → → →

Row(1,2) =

Row (1,3) = → → →

Row(1,3) =

Row (2,3) =→→→

Row(2,3) =

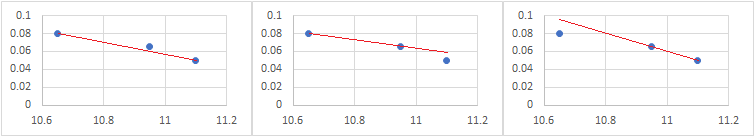


Figure 1: Plotted points for Row 1 & Row 3, Row 1 and Row 2, and Row 2 and Row 3, respectively

Choosing the first and last points, or Row 1 and Row 3, as a solution will result in the least error among the three possible combinations for the given data points. This is visualized in Figure 1

C: Set up the system of normal equations and solve it to obtain the least squares solution to the overdetermined system. Compare your results with those obtained in part B.

Applying a set of normal equations gives us:

And

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 10 | 11.6 | 116 | 100 |
| 15 | 11.85 | 177.75 | 225 |
| 20 | 12.25 | 245 | 400 |
| Σ = 45 | Σ = 35.7 | Σ = 538.75 | Σ = 725 |

Thus, = 0.065

Using the means for t and y, 15 and 11.9 respectively, along with the x1 value allows us to solve for the intercept using rearranged to :

Thus solving gives: 10.925 and 0.065

In conclusion, this is closely aligned with the error-minimizing guess of utilizing row 1 and row 3 to form an estimate with low residuals.

**Problem 2 (Heath 3.2)**

Suppose you are fitting a straight line to the three data points (0,1), (1,2), (3,3).

A: Set up the overdetermined linear system for the least squares problem.

Similar to the previously used overdetermined linear system

B: Set up the corresponding normal equations.

Again, using the equations previously used, albeit updated to more common notation:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 0 | 1 | 0 | 0 |
| 1 | 2 | 2 | 1 |
| 3 | 3 | 9 | 9 |
| Σ = 4 | Σ = 6 | Σ = 11 | Σ = 10 |

Therefore, = 0.642857

And

C: Compute the least squares solution by Cholesky factorization.

Therefore, ATAx = ATb =

The Cholesky factorization of ATA is:

Thus, solving for Ly=ATb gives y=[3.464, 1.388) and solving for LTx = y with back substitution gives x = [1.1427, 0,6429] as the least squares.

**Problem 3 (Chapra Canale 17.3)**

A: Use least squares regression to fit a straight line to:

X: 0, 2, 4, 6, 9, 11, 12, 15, 17, 19,

Y: 5, 6, 7, 6, 9 8, 7, 10, 12, 12,

Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the regression line.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 0 | 5 | 0 | 0 | 0.022042 |
| 2 | 6 | 12 | 4 | 0.196714 |
| 4 | 7 | 28 | 16 | 0.545507 |
| 6 | 6 | 36 | 36 | 0.933842 |
| 9 | 9 | 81 | 81 | 0.953035 |
| 11 | 8 | 88 | 121 | 0.531011 |
| 12 | 7 | 84 | 144 | 4.331289 |
| 15 | 10 | 150 | 225 | 0.019206 |
| 17 | 12 | 204 | 289 | 1.337435 |
| 19 | 12 | 228 | 361 | 0.203884 |
| Σ = 95 | Σ = 82 | Σ = 911 | Σ = 1277 | Σ = 9.073965287 |

Fitting the least squares using the same formulas as the prior problem:

Yields:

= 0.35246996

= 4.85153538

Thus,

The standard error can be defined as:

where

Since we have little information regarding the sample, we will use n-1 in the SSE formula.

Thus:

For the correlation coefficient we used Pearson’s formula:

Thus: = 0.91476728

Figure 2

Figure two shows the relationship of the predicted values to the observed values. The error observed is the vertical distance between the observations and regressed line of the predicted values.

B: Then repeat the problem, but regress x versus y: that is, switch the variables. Interpret your results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 5 | 0 | 0 | 25 | 3.620944 |
| 6 | 2 | 12 | 36 | 5.184631 |
| 7 | 4 | 28 | 49 | 7.028221 |
| 6 | 6 | 36 | 36 | 2.968803 |
| 9 | 9 | 81 | 81 | 5.756547 |
| 8 | 11 | 88 | 64 | 3.899915 |
| 7 | 12 | 84 | 49 | 28.61095 |
| 10 | 15 | 150 | 100 | 1.504593 |
| 12 | 17 | 204 | 144 | 2.315214 |
| 12 | 19 | 228 | 144 | 0.228883 |
| Σ = 82 | Σ = 95 | Σ = 911 | Σ = 728 | Σ = 61.11870504 |

Fitting the least squares using the same formulas as the prior problem:

Yields:

= 2.37410072

= -9.96762590

Thus,

The standard error can be defined as:

where

Since we have little information regarding the sample, we will use n-1 in the SSE formula.

Thus:

For the correlation coefficient we used Pearson’s formula:

However, one will note that due to the commutative property this is unchanged when X and Y are exchanged, thus the Pearson’s coefficient remains the same: 0.91476728.

Figure 3

Figure 3 showed the results of reversing the values. The error is now the horizontal distance to the regressed line. It should be noted that the observed error is distorted due to the lack of scale between X and Y: the increased distance between points along the X axis exaggerates the visually apparent error in this model, although it does present with relatively greater error.

Reviewing the models, the error present in each represents the residual values observed between the regressed line and the observed values based on the axis being interpreted. This causes them to have different amounts of error and to create slightly different regressed lines when plotted along the same values, as shown in the following chart, figure 4.

Figure 4

**Problem 4 (Chapra Canale 17.6)**

Use least-squares regression to fit a straight line to:

X: 1, 2, 3, 4, 5, 6, 7, 8, 9,

Y: 1, 1.5, 2, 3, 4, 5, 8, 10, 13,

A: Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the straight line. Assess the fit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 1 | 1 | 1 | 1 | 2.41975309 |
| 2 | 1.5 | 3 | 4 | 0.35667438 |
| 3 | 2 | 6 | 9 | 0.13040123 |
| 4 | 3 | 12 | 16 | 0.67148920 |
| 5 | 4 | 20 | 25 | 1.63271605 |
| 6 | 5 | 30 | 36 | 3.01408179 |
| 7 | 8 | 56 | 49 | 0.03780864 |
| 8 | 10 | 80 | 64 | 0.12056327 |
| 9 | 13 | 117 | 81 | 3.56790123 |
| Σ = 45 | Σ = 47.5 | Σ = 325 | Σ = 285 | Σ = 11.98138889 |

Fitting the least-squares regression to a straight line:

Yields:

Thus,

Standard Error:

where

Since we have little information regarding the sample, we will use n-1 in the SSE formula.

Thus:

For the correlation coefficient we used Pearson’s formula:

Thus: = 0.95622229

Figure 5

The fit shown in figure 5 is deficient, as the observed line is clearly curved whereas the linear regression line is straight. However, in this particular sample, the relevant range is small enough to create a relatively high correlation between X and Y.

B: Recompute A, but use polynomial regression to fit a parabola to the data. Compare the results to those of A. Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the straight line. Assess the fit.

A parabola will be fit by polynomial equation order

Using a brief python script to obtain the values used renders the coefficients and intercept such that:

Computing the standard error of this equation with

where

Again, using n-1 to account for working with a sample of unknown proportion, yields

This is relatively small compared to the least square linear regression standard error of 0.95622229. The improvement in the model is evident in a chart of the predicted dependent variable values as well, as shown in figure 6. The polynomial regression, shown in grey, nearly covers entirely the plot of the observed data, as a contrast to the orange plot of the straight line linear regression.

Figure 6

# **Bibliography**

Higham, N. J. (2009). *Cholesky Factorization.* Retrieved from Manchester School of Mathematics: http://www.maths.manchester.ac.uk/~higham/papers/high09c.pdf

Kent State University. (2018, 9 26). *Pearson Correlation*. Retrieved from Kent State University Libraries: https://libguides.library.kent.edu/SPSS/PearsonCorr

**Appendix 1:**

import numpy as np

a = np.array([[1,10],[1,15],[1,20]])

b = np.array([[11.6],[11.85],[12.25]])

x = np.linalg.lstsq(a,b, rcond=None)

print(x)

**Appendix 2:**

a = np.array([[1,0],[1,1],[1,3]])

aT = a.transpose()

b = np.array([[1],[2],[3]])

ata = np.dot(aT,a)

atb = np.dot(aT,b)

L = np.linalg.cholesky(ata)

strings = ("A","A Transpose","A Transpose \* A","A Transpose \* B","Cholesky Lower")

variables = (a,aT,ata,atb,L)

for i in range(0,len(strings)):

print(strings[i])

print(variables[i])

print()

**Appendix 3:**

Set pi =1, i=1:n

Do for k = 1:n

Find s such at ass = maxk<=i<=naij

Swap rows and columns of k and s of A

And swap pk and ps

akk = Sqrt(akk)

Do for j = k + 1:n

akj=akj/akk

End for

Do for j = k + 1:n

Do for i = k + 1:j

aij = aij - akiakj

End for

End for

End for

Set P to the matrix whose jth column is the pith column of I.

(Higham, 2009)

**Appendix 4:**

# Polynomial Regression

# Finding summations of x^n and yx^n:

x = [1,2,3,4,5,6,7,8,9]

y = [1, 1.5, 2, 3, 4, 5, 8, 10, 13]

x\_power = {'sumx0':0,'sumx':0,'sumx2':0,'sumx3':0,'sumx4':0}

xy\_power = {'sumx0y':0,'sumx1y':0,'sumx2y':0}

def sumpowerx(sum\_,n):

for i in x:

sum\_+=i\*\*n

# print(sum\_)

return(sum\_)

def sumpowerxy(sum\_,n):

for i in x:

# print(i)

sum\_ += i\*\*n \* y[i-1]

return(sum\_)

for k,v in x\_power.items():

n = list(x\_power.keys()).index(k)

x\_power[k] = sumpowerx(v,n)

for k,v in xy\_power.items():

n = list(xy\_power.keys()).index(k)

xy\_power[k] = sumpowerxy(v,n)

print(x\_power)

print(xy\_power)

#%% Solving the linear system:

a = np.matrix([[9,45,285],[45,285,2025],[285,2025,15333]])

b = np.matrix([[47.5],[325],[2438]])

x = np.linalg.solve(a,b)

print(x)