MATH 392, Seminar in Mathematics and Climate

Computer Project #2: Working with Climate Data

DUE DATE: Friday, April 13, 5:00 pm

The goal of this project is for you to gain experience working with actual climate data. You will fit CO_2 measurements from the Mauna Loa Observatory and the South Pole Observatory to both linear and quadratic curves. You will interpret and analyze these fits and then make some predictions. We will be using Matlab to perform the calculations and produce plots, but the analysis could be performed on any standard statistical package. Some of the ideas and material for the lab were borrowed from Chapters 9 and 10 of the course textbook.

For this project, it is **required** that you work in a group of two or three people. Any help you receive from a source other than your lab partner(s) should be acknowledged in your report. For example, a textbook, website, another student, etc. should all be appropriately referenced at the end of your report. The project should be **typed** although you do not have to typeset your mathematical notation. For example, you can leave space for a graph, computations, tables, etc. and then write it in by hand later. You can also include graphs or computations in an appendix at the end of your report. Your presentation is important and I should be able to clearly read and understand what you are saying. Spelling mistakes and sentence fragments, for example, should not occur. Only **one project per group** need be submitted.

Your report should provide coherent answers to each of the following exercises. Be sure to read carefully and answer all of the questions asked.

Note: Please include a printout of your Matlab script file(s) (or send via email).

1 Linear Regression

Suppose we have a set of data points $\{(x_i, y_i) : i = 1, \ldots, n\}$. A plot of the data seems to reveal a particular relationship between x and y (e.g., linear, quadratic, exponential, periodic). How do we find the best fit to the data? The goal in regression analysis is to find a functional relationship between x and y that best approximates the data. In other words, we seek some function $f(x)$ so that $y_i \approx f(x_i)$ for each data point.

The Method of Least Squares

Suppose we suspect the relationship between x and y is a simple linear function, $y = mx + b$. Ideally, we want to find m and b so that $y_i = f(x_i) = mx_i + b$, or, more practically speaking, we want the *error* $y_i - f(x_i)$ to be as small as possible.

Here is an approach using linear algebra. The equations we want to solve are

$$\begin{aligned} mx_1 + b &= y_1 \\ mx_2 + b &= y_2 \\ &\vdots \\ mx_n + b &= y_n. \end{aligned}$$

Since this is n equations in the 2 unknowns m and b, this system is *overdetermined* and there is typically no solution. The goal then, is to find the *best fit* to the data.

The above system of equations can be written more compactly as $Xv = Y$, where

$$X = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad v = \begin{bmatrix} m \\ b \end{bmatrix}, \quad \text{and} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{1}$$

The values of X and Y are determined by the data, but v is unknown. We would like the error vector $Y - Xv$ to be as small as possible. In other words, we want to minimize the length of the error vector, $||Y - Xv||$. This is equivalent to minimizing the square of the length,

$$||Y - Xv||^2 \;=\; (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \cdots + (y_n - (mx_n + b))^2.$$

Minimizing the sum of the squares of the errors is called the *method of least squares*. We will prove the following theorem in class. Recall that X^T denotes the transpose of the matrix X.

Theorem 1.1 *The least squares solution to $Xv = Y$ is the vector v satisfying the equation*

$$X^T X v \;=\; X^T Y. \tag{2}$$

It is important to note that $f(x)$ does not have to be a linear function for this technique to work. The important feature is that the unknown parameters enter the problem linearly. For example, if we had wanted to fit the data to a quadratic function $f(x) = \alpha_2 x^2 + \alpha_1 x + \alpha_0$, then the system can still be written as $Xv = Y$, but now X and v become

$$X = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} \alpha_2 \\ \alpha_1 \\ \alpha_0 \end{bmatrix}. \tag{3}$$

The beautiful aspect of the method of least squares is that equation (2) holds for *any* system of the form $Xv = Y$.

If X is an $n \times p$ matrix (n data points, p unknown coefficients in the formula for $f(x)$), then $X^T X$ is a $p \times p$ square symmetric matrix. If the columns of X are linearly independent, which is typical in applications, then $X^T X$ is invertible and equation (2) has a *unique* solution given by

$$\boxed{v \;=\; (X^T X)^{-1} \cdot X^T Y.} \tag{4}$$

One can check that the dimension of the vector v is $p \times 1$. This vector is easily computed in Matlab.

Residuals and the Coefficient of Determination

Once we have obtained the vector v using equation (4), we have determined the best fit $f(x)$. The *residuals*, given by $r_i = y_i - f(x_i)$, measure how far each data point is from the fit. If we define $\hat{Y}$ as $\hat{Y} = Xv$, then the *residual vector* r is

$$r \;=\; Y - \hat{Y} \;=\; Y - Xv.$$

The length of r gives some indication to the quality of the fit. A more commonly used measure in statistics is the *coefficient of determination* or R^2 value.

Let $\overline{y}$ represent the mean of the y_i data, that is,

$$\overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \;,$$

and let $\overline{Y}$ be the $n \times 1$ vector

$$\overline{Y} = \begin{bmatrix} \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{bmatrix} .$$

The coefficient of determination is given by the formula

$$\boxed{R^2 \;=\; \frac{||\hat{Y} - \overline{Y}||^2}{||Y - \overline{Y}||^2} \;.} \tag{5}$$

The value of R^2 measures how well the fit $\hat{Y}$ does in comparison to the mean $\overline{Y}$. The closer to 1, the better the fit.

Useful Matlab Commands

The table below provides some useful commands in Matlab for working with data and matrices.

Command	**Meaning**
inv(A)	inverse of the matrix A
A'	transpose of the matrix A
a = [1:0.1:7]	vector with equally spaced entries between 1 and 7
ones(35,1)	column vector with 35 ones
a(6)	6th entry in the vector a
a(6:12)	6th through 12th entries in the vector a
a.^2	the square of each entry in the vector a
A = [a b]	a matrix whose columns are the vectors a and b
norm(a)	length of the vector a
mean(a)	mean of the entries in the vector a
A*B	product of the matrices A and B
A*b	product of the matrix A and vector b
plot(x,y,'r')	plot of the points in vectors x and y in red
plot(x,y,'r', a,b,'b')	two plots on the same set of axes in different colors

Exercise 1: Warmup

Before trying to work with actual climate data, let's begin with a simple example. Consider the 10 data points (x_i, y_i) given in Table 1. A plot of this data (see Figure 1) indicates a linear relationship.

x	y
1.0	3.3
1.6	3.5
2.3	3.2
3.1	3.5
3.7	4.1
4.2	4.9
5.4	5.0
6.1	5.3
6.5	5.6
7.0	6.5

Table 1: Some sample data.

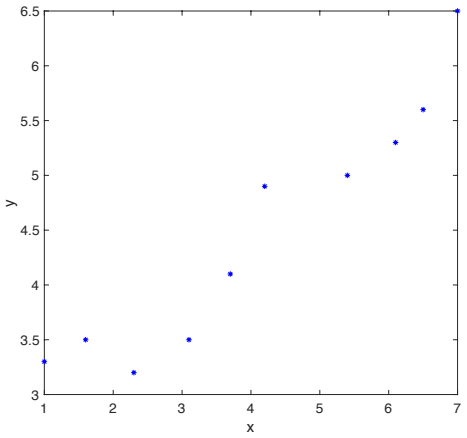

Figure 1: A plot of the ten data points from Table 1.

(a) Using Matlab, enter the data into two column vectors x and y. Form the special data matrix X shown in equation (1).

(b) Fit the data to a straight line using the method of least squares and equation (4). Give the values of m and b. Print out a plot of the original data along with the linear fit.

(c) Using equation (5), find the R^2 value and comment on the quality of your fit.

(d) Using your linear fit, estimate the value of y when $x = 10.0$ (one decimal place).

Exercise 2: The Keeling Curve and Mauna Loa CO_2 Data

Recall from class our discussion of the famous *Keeling curve*, one of the most important data sets in climate science. Measurements of carbon dioxide (CO_2) at the Mauna Loa Observatory at the Mauna Loa volcano in Hawaii were begun by Charles Keeling in March of 1958 and continue to this day under the guidance of his son Ralph Keeling. Air samples have been taken hourly, every day, using the same measuring technique for 60 years.

For this portion of the project you should import the data into Matlab from the spreadsheet titled *CO2Data-MLO.xlsx*. This file contains monthly averages of CO_2 dry air mole fractions (in parts per million by volume) measured in air samples collected in glass flasks. The measurements given range from March of 1958 to February of 2018. To import the data into Matlab, click on the `Home` tab, and then the `Import Data` button (green arrow). This will open up a new window and allow you to click on the correct spreadsheet.

Import the relevant data into Matlab as column vectors. The columns to import are column C, the decimal dates; Column E, the interpolated data; and Column F, the seasonal corrections. You should adjust the range of the import to exclude the first two rows (e.g., F3:F722 is better than F1:F722). Click on the `Import Selection` button and then choose `Import Data` from the drop-down menu. This should create a vector in Matlab that is visible in your workspace. You can change the name of the vector (do this) by clicking on the vector name in the workspace.

(a) Make a plot of the Keeling curve including both the interpolated data (in red) and seasonal corrections (in black). The year should be on the horizontal axis and the CO_2 concentrations on the vertical. What is the reason for the oscillation in the red curve?

(b) Explain the meaning behind the decimal dates. How were these obtained? For example, why is March of 1958 written as 1958.208 and July of 1958 written as 1958.542?

(c) Fit the interpolated data to a straight line using the method of least squares and equation (4). Print out a plot of the original data along with the linear fit. Find the R^2 value (four decimals) and comment on the quality of your fit. According to your fit, what is the average predicted increase in CO_2 per year?

(d) Fit the data to a quadratic function using the method of least squares and equation (4). You will need to compute the matrix X given in equation (3). Because the square of the years is quite large, you might consider shifting the date vector so that $t = 0$ corresponds to March of 1958 before trying the method of least squares. Print out a plot of the original data along with the quadratic fit. Find the R^2 value (four decimals) and comment on the quality of your fit. Which approximation captures the trends in the data better, linear or quadratic?

(e) What does the leading coefficient of the quadratic reveal about the longer term trend of the CO_2 data?

(f) Using your quadratic fit, estimate the amount of CO_2 at Mauna Loa in January of 2030, 2050, and 2100.

(g) Pick one calendar year where the data is complete. Plot the residuals for the quadratic fit over this year (from January to December). What do you notice about the graph?

Exercise 3: CO_2 Data from the South Pole Observatory

The South Pole Observatory is situated at the geographic South Pole and has been collecting data since 1957. It is one of the four major observatories operated by the NOAA Earth System Research Laboratory, Global Monitoring Division (GMD). Ozone data collection began in 1963 and CO_2 concentrations started being measured in 1975. Two GMD scientists work yearlong shifts collecting data at the observatory, including nine months in isolation and six months in darkness.

For this portion of the project you should import the data into Matlab from the spreadsheet titled *CO2Data-SPO.xlsx*. This file contains monthly averages of CO_2 dry air mole fractions (in parts per million by volume) measured in air samples collected in glass flasks. The measurements given range from July of 1975 to December of 2016.

(a) Import the data into Matlab as column vectors and then plot the monthly averages of CO_2 against time. You will first need to combine the year and month columns into a single column with decimal dates (e.g., July 1975 should be converted to 1975.5416).

(b) What peculiarity do you notice about your plot around the year 1980? Why does this occur?

(c) Fit the data to a straight line using the method of least squares. Print out a plot of the original data along with the linear fit. Find the R^2 value (four decimals) and comment on the quality of your fit. According to your fit, what is the average predicted increase in CO_2 per year?

(d) Fit the data to a quadratic function using the method of least squares. Print out a plot of the original data along with the quadratic fit. Find the R^2 value (four decimals) and comment on the quality of your fit. Which approximation captures the trends in the data better, linear or quadratic?

(e) Pick one calendar year where the data is complete. Plot the residuals for the quadratic fit over this year (from January to December).

(f) Using your quadratic fit, estimate the amount of CO_2 at the South Pole in January of 2030, 2050, and 2100.

(g) Write a paragraph comparing and contrasting your findings from the Mauna Loa and South Pole Observatories.