

Matlab Assignment

Due date: March 29, 2005

1 Linear Regression

1.1 Introduction

Recall that linear regression tries to find the best fit line, or coefficients a and b that satisfy $y = a + bx$ from data (x_i, y_i) . One of the ways to do this is to form the normal equations $A^T Ax = A^T b$, where the vector x consists of the coefficients a and b . This is called *ordinary least squares regression*, as in effect you can show that you are minimizing the vertical residual to the hypothetical best fit line. You may want to look on page 226 of your text book to refresh yourself of this example. In any case, the normal equations produce the following matrix equation:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Implicit in this assumption is that there is no error in the independent variable x_i , and that all error comes from the dependent variable y_i . For many applications, this assumption is not correct. Alternative linear regression techniques exist when *both* x_i and y_i have errors to them.

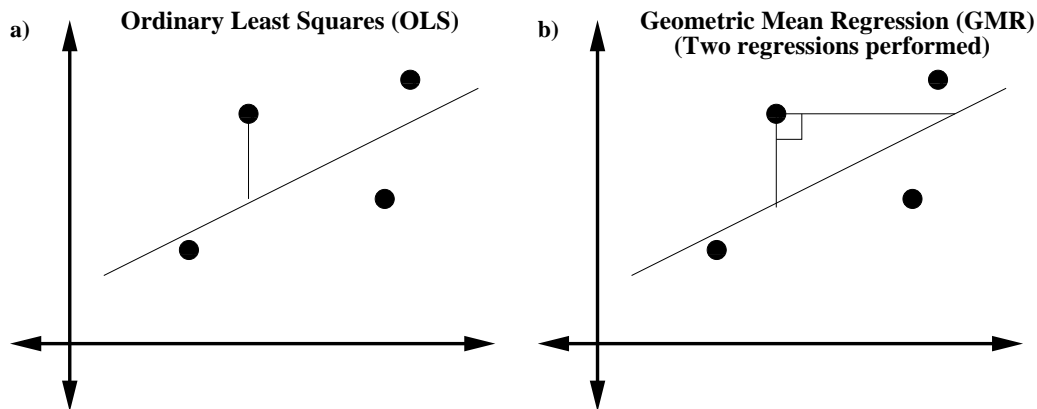


Figure 1: a) Ordinary least squares regression. The vertical residuals among data and a hypothetical best fit line are minimized. b) Geometric mean regression. The vertical *and* horizontal residuals among data and a best fit line are minimized.

One of these techniques is called *geometric mean regression* or *reduced major axis regression*. What this does is minimize the horizontal residuals as well. Conceptually two regressions are performed: an ordinary least squares regression of \bar{y} versus \bar{x} , and then the data are “switched” so that an ordinary least squares regression of \bar{x} versus \bar{y} is done. The two slopes are inversely related. In practice, it can be shown that the slope and intercept of a geometric mean regression are:

$$b_{GMR} = \frac{b_{OLS}}{|r|}, \quad a_{GMR} = \bar{y} - b_{GMR}\bar{x},$$

where GMR refers to geometric mean regression, OLS, ordinary least squares, \bar{y} refers the average (mean) of data y_i (similarly for \bar{x}), and r is the correlation coefficient:

$$r = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}.$$

So to calculate the regression coefficients for a geometric mean regression, you calculate the slope of an ordinary least squares regression, the correlation coefficient, and the respective means of the data. Figure 1 shows the relationship between ordinary least squares and geometric mean regression.

1.2 Assignment

1. Consider the relationship $y = a + bx$ and the relationship $x = \tilde{a} + \tilde{b}y$. How are b and \tilde{b} related?
2. Write a program that calculates the slope and intercept of an ordinary least squares regression and a geometric mean regression. You will find many of the needed commands in the program “simple_program.m” found on the course website. At the beginning of your program you should define the data you are fitting:

```
function regression

% A program to calculate the slope and intercept of a ordinary least squares and
% geometric mean regression.

x=[0:10];
y=[0:10];
```

Note that this data has a slope of 1 and an intercept of 0 no matter what regression you choose, so use this for debugging. Once you are satisfied that your program works well, download the data file “regression_data.mat” from the course website. This is data that I generated and know the slope and intercept for both regression types. Once you have saved the file to your working directory, you can include it by adding the following lines to your program:

```
function regression

% A program to calculate the slope and intercept of a ordinary least squares and
% geometric mean regression.

% x=[0:10];
% y=[0:10];

load regression_data.mat
```

The data file already has variables x and y , but are much longer than 11 elements.

Email me (zobitz@math.utah.edu) your program when you are completed and I will compare it to my results. (In class you will email me a test message.)

3. This question is a little more open-ended and should cause you to think. If you have written your program correctly, you should have obtained strikingly different regression coefficients for the ordinary least squares and geometric mean regressions. What would the benefits of using geometric mean regression over ordinary least squares regression? One thing to think about that may stimulate your thinking is thinking about the purposes of linear regression, and the idea of a “independent” and a “dependent” variable.

2 Powers of a Matrix

2.1 Introduction

For this part you are going to calculate powers of a two by two matrix and multiply each of these powers by a vector and “track” how the equation $A^t \vec{x}$ “evolves” through time, where t is a particular matrix power. This will give you exposure to generating plots in Matlab. A lot of the program code is written in the program “simple_program.m” available through the course website. On March 29 the reason for this part will become evident. You should print out the plots you generate and bring them to class that day.

2.2 Assignment

1. Write a program that will iterate through 10 powers of the equation $A^t \vec{x}$, where:

$$A = \begin{bmatrix} 0.86 & 0.08 \\ -0.12 & 1.14 \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} 100 \\ 300 \end{bmatrix}$$

Generate a plot of each iteration $A^t \vec{x}$, labeling where $t = 0$ and $t = 10$.

2. Do the same as above, but with

$$\vec{x} = \begin{bmatrix} 200 \\ 100 \end{bmatrix}$$

3. The same as above, but with

$$\vec{x} = \begin{bmatrix} 1000 \\ 1000 \end{bmatrix}$$

4. Either plot all three of the above simulations on the same plot, or carefully hand draw what you see on a piece of paper. Do you observe anything intriguing?