

## LAB 2

Math 441

Thursday, October 16

You do not need to turn in solutions to this lab. However, you are responsible to understand all of the material presented here.

1. **Simple Linear Regression (one explanatory variable, one response):** On October 21, 2005, The Bozeman Daily Chronicle ran the AP article “Congress approves gun-lawsuit shield.” According to the article, incidences of murder, robberies and aggravated assault are down from 1990’s highs. The data file **crime.txt** on the MATH441 web site shows the incidences of these crimes (in the millions) in the 2nd column, for the odd years from 1991 to 2003 (in the first column). Download this file then load it into MATLAB by **load crime.txt** to perform the following analysis.

- (a) Plot the data points, **plot(t,y,'\*')**, **hold on** where  $t$  contains the time variable and  $y$  is millions of crimes.
- (b) Construct the Vandermonde matrix  $A$ , set up the system  $Ac = y$  where  $c = \begin{pmatrix} a \\ b \end{pmatrix}$  and now find the least squares line  $y = a + bx$ . Which method (“classic” or QR factorization) did you use?
- (c) Plot this line: **trange=1:1:13; plot(trange,a + b\*trange,'r')**
- (d) Compare five different ways to find the least squares solution by implementing the following:

```
tic, x=inv(A'*A)*(A'*y), toc % classic method using inversion
tic, x=(A'*A)\(A'*y), toc % classic method using Gauss Elimination
tic, R=chol(A'*A); x=R\(R\'(A'*y)), toc % classic method using Cholesky
tic, [Q R]=qr(A,0); x=R\'(Q'*y), toc % QR method
tic, [U S V]=svd(A,0);x=V*diag(1./diag(S))*U'*y, toc %SVD method
```

Record the times for each. Which method is fastest? Does this agree with the assertion we made in class? A useful way to quantify the differences is by looking at a relative difference in time:  $(\text{time} - \text{fastesttime})/\text{fastesttime}$ . Rank the 4 methods from fastest to slowest.

2. **Least Squares with Categorical Data: Yes/Nos:** Last month I went to my first baby shower. Since my wife was going for an ultrasound that week, and we wanted the ultrasound technician to tell us the gender of our child, I was wondering how often these techs correctly classify the sex of the fetus. I polled all of the parents there (and there were a lot of them) as well as a few others. In the data file at the MATH441 web site called **ultrasound.txt**, I used 1’s to encode “Correct classification”, and 0’s encode “incorrect/no classification.”

- (a) Let  $y$  be the vector of 0 and 1 responses. Set up the linear system  $Ac = y$  to solve.
- (b) Find the least squares solution to the model  $y = a$ .
- (c) Note that geometrically,  $y = a$  is a horizontal line. Look at the least squares solution to  $a$  that you get. What statistic does the least squares solution correspond to?
- (d) Starting at  $Ac = y$ , you can use the analytical (“equation form”) of the least squares solution,  $c = (A^T A)^{-1} A^T y$  to corroborate that you always get the statistic you mention in #??.

**3. Multiple Linear Regression (multiple explanatory variables, one response):**

Consider data used to support global warming from the years 1890-1980. In the data file **globalwarming.txt**, available at the MATH441 web site, the first column is year; the second column is mean global temperature, given as a change in degrees Celsius from the 1951-1980 average (published by NASA's Goddard Institute for Space Studies); the third column is atmospheric carbon in parts per million (from the well-known "Keeling Curve"); the fourth column is solar magnetic cycle length in years (from a 1991 *Science* paper).

- (a) Let  $y = \text{temperature}$  (the second column of the data matrix). Set up the matrix system to find the least squares solution to

$$y = a + bx_{\text{year}} + cx_{\text{carbon}} + dx_{\text{solar}}.$$

Note that this corresponds to fitting a *hyperplane* to the data in  $(x_{\text{year}}, x_{\text{carbon}}, x_{\text{solar}}, y)$  space.

- (b) Find the least squares solution.
- (c) In the year 2020, if carbon levels increase to 350 ppm, and if solar magnetic cycle length is at 10.5 years, by how much is the mean global temperature predicted to warm up?