Assignment 3 : Basic Statistics, Least Squares, Correlation, and Data Fits

C. Alex Simpkins

October 31, 2006

1 Description

1.1 The Basics

Read this entire document before attempting the homework...

This assignment is worth 100 points, and a a 10 point extra credit will be possible!

In this assignment you will apply concepts of data approximation and fitting to some real data generated from your class surveys. Each modeling tool gives you another way to represent, simplify, quantify and make decisions about the real system you are dealing with. Understanding the measure of goodness of fit is important, so you will also not only fit data, but judge the goodness of fit and make decisions based on those fits.

You will compute basic statistics of the data, and make some statements about basic relationships between variables (ie could height be related to weight?).

You will be required as usual to turn in a well-commented listing of either your matlab scripts or the commands you used to perform your calculations.

1.2 The formatting details (fulfill all of these, and get your 10 bonus points!!! A midterm present)

To avoid unnecessary complexity, we will be turning in the assignment in the same way as before: Printed out. But to avoid a Karate chop to the wallet (this joke was from a student in your class) you are NOT required to turn in any color prints, just black and white (Yay!).

Formatting Requirements:

- Cover page with your name, the date, class, quarter, your section, and the homework number/title
- Pages must be numbered
- No plots should be JPEGs. Use some vector-based graphics format (or one of more appropriate compression strategies) such as PDF, EPS, GIF, etc when exporting figures from matlab to your document file (if you have any issues with this email or otherwise let us know). The plots must be clear and not blurry.
- You may recycle printer paper which has been printed on only one side IF what is on the back does not bleed through to the front or in any way interfere with your assignment. Also the material on the reverse side should be clearly unrelated to this assignment.
- You may NOT use lined paper (such as from a binder or college ruled paper)
- All figures must have axis labels and a figure title. If there is more than one data line or there is a scatterplot of *'s and lines, you must use a legend to clarify which line corresponds to what information source.
- If you have plots with more than one line, and since you are printing in black and white, you should use more than one linetype to set the lines apart.
- Turn in a well commented listing of your matlab code in an Appendix at the end of your assignment paper

2 Instructions

2.1 Download the data

Download the data files for this assignment on the handouts section of the page, or the assignments section. Both are the same data. The file is called **hw3.zip**. Unzip the file, and you should have a single data file called **classdata.mat**. You will also have a brief readme which describes the data file.

2.2 Load the data

The data file is a binary mat file. Load the file into matlab using the load command, or the import wizard:

```
load classdata.mat
```

You should now have a single variable with several rows and columns. The rows correspond (to keep in standard matlab format) to observations, the columns to variables. The order of the columns are as follows:

$Bday(day)$	$Bday(year)$	$Shoesize$	$height$	$pairsofshoes$	$gender$
0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	0	0	0

The name of the variable is `hw3data`

2.3 Compute the basic statistics (20 points)

2.3.1 The equations

Write down the equation for the mean and standard deviation (Hint: see matlab help for std).

2.3.2 The Computations

For each column (i.e. each variable) compute the following:

- **mean - use the matlab command** `mean()`
- **median - use the matlab command** `median()`
- **standard deviation - use the matlab command** `std()`

And tabulate the results (ie put the results in a table which has along one axis the variable names, and along another axis the statistic. The table may look something like the following:

	Mean	Median	StDev
num shoes	0	0	0
m/f	0	0	0
weight	0	0	0
height	0	0	0

Table 1: Example of a results table (this does not include all the variables

2.3.3 How to compute the statistics

Recall that the matlab functions will each operate on a matrix with rows being the observations and columns the variables. So for example if you have a variable called COOLDATA which, when at the command prompt in matlab you type

```
size(COOLDATA),
```

you would get matlab's response as (keep in mind that your variables may NOT be this size, this is a demonstration only)

```
ans =

81 6
```

This means you have COOLDATA as a variable with 81 rows (which correspond to observations) and 6 columns (which correspond to different variables such as height, weight, etc). To compute the mean of a matrix of variables and observations, then, just make sure the variable is in the form of the rows being observations, columns being individual variables, and type something like

```
MatrixMean = mean(COOLDATA);
```

and matlab will compute the mean of each column of data and return it as an array of size 1x6 (ie one row, 6 columns of variables, or one mean per variable).

2.4 Plot the distributions (20 points)

2.4.1 Plot the distributions of each variable using the function `hist()` in matlab. (10 points)

You should get a result that looks something like the following (not necessarily in shape, just overall appearance of the plot).

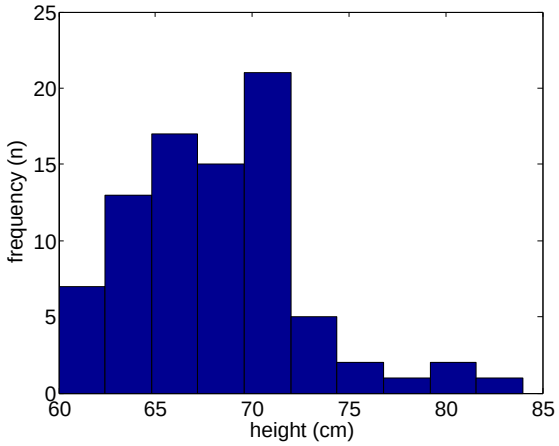

Figure 1: Distribution of heights of students in CogSci 109, the coolest class around!

If you have any problems with matlab crashing remember to type our old friend at the command prompt:

```
opengl neverselect
```

If you still have problems let Alex, Nick, Dan or Nathaniel know with an email or other communication, and we will help you resolve the issues.

2.4.2 Question (10 points)

- Is there a standard distribution you can see the variables behaving according to? Each variable might have a different distribution. Don't try to find one

for all of them, just one for each variable. For example, it may be unimodal, bimodal, normal, positive skew or negative skew, etc.

- Briefly, pick one distribution and make an observation about the characteristic of the distribution - ie, if the variable for sleep time were 2am, what would that suggest (brief answer, maybe one or two sentences - not intended to be complex)?

2.5 Scatterplots and linear fits (20 pts.)

2.5.1 First create the scatterplot (10 points)

Choose one pair of variables, and create a scatterplot. Do this by simply plotting one variable as the x-axis, and the other as the y-axis, and format the plot to use '*' instead of lines. For example:

```
plot(height, weight, 'r*')
```

We will later comment on the relationship after fitting a simple curve, so you can be creative with the pairing. Your scatterplot will look something like the following:

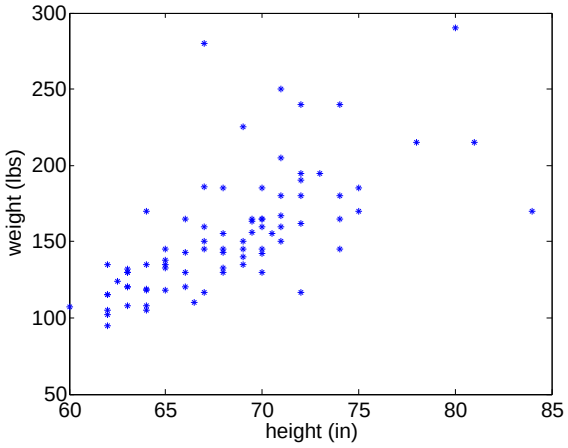

Figure 2: Height vs. Weight for the class

2.5.2 Linear and nonlinear least squares fits (10 points)

Now before saving this let us create a few data fits to model the trend. Then we will see what kind of correlation coefficient we get. You can save one image file which includes the scatterplot and the two fits we're going to perform.

Set up a linear least squares problem just like the handout given on least squares. Follow that example to fit a two parameter system ($y = a_0 + a_1 x$, with unknown parameters a_0 and a_1) and then plot the fit in the same plot as the scatterplot.

Do the same thing, but create a quadratic fit (fit the equation $y = a_0 + a_1 x + a_2 x^2$ by finding the best fit for a_0, a_1, and a_2)

Plot lines for each fit in the same figure window as the scatterplot figure. (hint: use the hold on command in matlab

Use the legend command in matlab to create a legend for the plot, and label each line or the data. See the examples for plotting on the course web page for examples. Remember that you have to call the legend command AFTER you have plotted all your data and fits.

2.6 Computing the correlation coefficient (20 points)

2.6.1 Compute the coefficient and tabulate it

Use the matlab command `corrcoef()` to compute the correlation coefficient for the pair of variables you chose. List the correlation coefficient matrix in a table such as the following:

0	height	weight
height	1	.2
weight	.2	1

2.6.2 Question

- Is the relationship between the pair of variables you chose a positive correlation, or a negative correlation? Explain briefly.
- Is the correlation strong or weak? Explain briefly.
- Briefly explain why the coefficients along the diagonal are always one.
- What might you postulate from these results (ie what can you suggest from your data as a question? We did not perform significance tests or hypothesis tests, but what questions might this data lead you to ask? Briefly state your response.)

2.7 Very Basic Model Error Analysis (20 points)

2.7.1 Norm-based error (10 points)

When modeling it is very important to create a quantitative description of modeling errors so one can evaluate and possibly improve the models generated if need be. One simple way to do this is the following.

Compute the 2-norm of the squared error for each of your fits. Let us define

$$\epsilon^2 = \sum_{i=1}^{n} \Big\{ y_i - P(x_i) \Big\}^2 \tag{1}$$

or you can write

$$\epsilon = \sqrt{\sum_{i=1}^{n} \Big\{ y_i - P(x_i) \Big\}^2} \tag{2}$$

This provides a basic piece of information in a single number about how good your fit to the data is. Use the matlab command `norm()` to compute the norm of the difference between your actual y-variable and the predicted y-variable from your fits. Do this by taking the linear and nonlinear fits performed above (you have an

equation from when you created the plots) and using them to compute predicted y-values given the x-values of the data. Then make a new variable such as NBE by subtracting actual y minus predicted y, then squaring those values. Finally compute the norm of those quantities with the norm command in matlab.

2.7.2 Question (10 points)

- Is the linear or nonlinear fit better?
- Why might one fit be better than another?