# COMS21202: Symbols, Patterns and Signals

## Lab 3: Least Square Method

NOTE: You will need to refer to the Matlab help pages to complete most of these examples.

1. Generate a sequence of 10 numbers $y_i$, $1 \leq i \leq 10$, using the following equation

$$y_i = 0.2 + 0.5x_i + e_i \tag{1}$$

where $x_i$ is a uniformly distributed random number between 0 and 1 and $e_i$ is a random number from $N(0, 0.1)$ indicating a normally-distributed error in the measurement.

Generate a scatter plot.

2. Change the above code to a function that receives the parameters $m$, $c$ and $n$ for the equation $y_i = c + mx_i + e_i$ and returns $n$ samples as shown in the previous exercise. Use the code to generate a 100 samples from the equation above.

Plot the new scatter plot for the 100 samples.

3. Using the sequences $x_i$ and $y_i$ from Q2, implement the matrix form of the method of least squares to determine estimates of the parameters $c$ and $m$ assuming that $y_i = c + mx_i + e_i$.

*Discuss: what is the effect of changing the mean and the standard deviation of $e_i$ on the estimated values of $m$ and $c$*

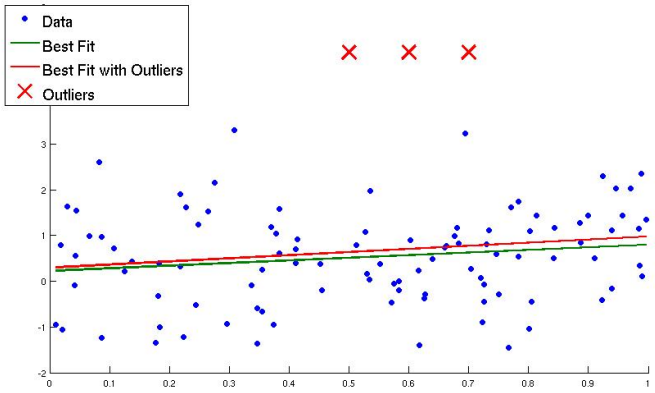
4. Draw the best fitting line defined by the estimated values of $c$ and $m$ on top of the scatter plot generated in Q2. Note: do not use any pre-defined functions. Write your own code to plot the line.

You should use the phrase (*hold on*) for multiple plots on the same figure.

5. Add a few outliers to your data. For example, add the points {(0.5, 5), (0.6, 5), (0.7, 5)} to the sampled sequences from Q2.

Re-estimate the best line that fits the data after the addition of the outlier.

Plot the best fit line before and after the outlier data, in a way similar to the figure below (including the colouring and the legend)




*Discuss: what is the effect of outliers on the best line estimate?*

6. Load the data in the file DMD.csv where column one is $x_i$ and column two is $y_i$.

Generate a scatter plot of $y_i$ vs $x_i$.

7. Assuming a polynomial relationship between $x_i$ and $y_i$, what would you suggest as a suitable complexity for the model? Write your model's equation - similar to Equation 1

   *Discuss: for the model you've chosen, what are the parameters of the model?*

8. Adjust the code in Q3 **if needed** to calculate the parameters of your chosen model for this data.

   *Discuss: what would **X** and **y** in the least squares solution be?*

9. Plot your best fitting polynomail from Q8.

10. What would the model be if you *generalise* it further?

11. What would the model be if you attempt to *overfit* the training data?

    *Discuss: write down your answers for Q10 and Q11 and discuss your written explanation with a TA*