

# 21112 (Calculus 2) Linear Regression

Albert Cohen

April 23 2003

Whenever we attempt to model a process, our first attempt is to fit it to a linear model. In other words, if we have an input variable (experimental parameter)  $x$  and an output (experimental parameter) variable  $y$ , then we assume that  $y = mx + b$ . But, if we look at experimental data, it very rarely occurs that all of the data points fall on a straight line. Still, we can fit a **best fit** line where we minimize the **mean square error** of the data points to the line. This will require multivariable calculus techniques that we now possess.

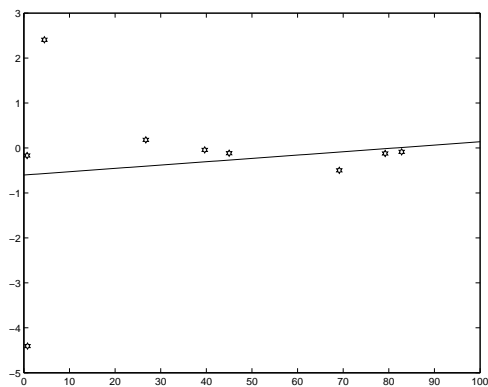


Figure 1: Data for an experimental auction design

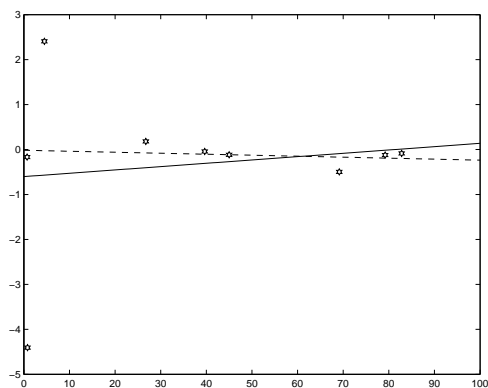


Figure 2: Same data as above, but now we've ignored some data points...

## 1 How do we see data?

Look at the data points in figure 1: Somehow, the line we've drawn doesn't seem visually correct....

After discarding two “outliers”, we fit a line that seems visually more correct. This is the advantage of being a human (versus a computer :), in that we can visually interpret this data and make conclusions. Still, there is a technology that enables us to fit a “best fit” line to our data.

## 2 Mean Square Error

For experimental data such as

Input	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>	x <sub>6</sub>	x <sub>7</sub>	x <sub>8</sub>	x <sub>9</sub>
Output	y <sub>1</sub>	y <sub>2</sub>	y <sub>3</sub>	y <sub>4</sub>	y <sub>5</sub>	y <sub>6</sub>	y <sub>7</sub>	y <sub>8</sub>	y <sub>9</sub>

we wish to fit to the line  $y = mx + b$ , we define the Mean Square Error as

$$E(m, b) := \sum_1^9 (y_i - mx_i - b)^2 \quad (1)$$

If we have  $n$  data points, then we have

$$E(m, b) := \sum_1^n (y_i - mx_i - b)^2 \quad (2)$$

For example, the data pictured above corresponds to the following table:

x	0.71	4.51	0.82	69.13	82.81	26.74	45.00	79.20	39.64
y	0.6	50	0.01	42	76	32	40	70	38

So our Error is

$$E(m, b) = (0.6 - 0.71m - b)^2 + (50 - 4.51m - b)^2 + (0.01 - 0.82m - b)^2 + (42 - 69.13m - b)^2 + (76 - 82.81m - b)^2 + (32 - 26.74m - b)^2 + (40 - 45.00m - b)^2 + (70 - 79.20m - b)^2 + (38 - 39.64m - b)^2$$

One method of fitting the best line therefore is **minimizing** the mean square error. This is done by taking the first partials of  $E(m, b)$  and setting them to zero:

$$0 = \frac{\partial E}{\partial m} = 2 \sum_1^n (y_i - mx_i - b) x_i \quad (3)$$

$$0 = \frac{\partial E}{\partial b} = 2 \sum_1^n (y_i - mx_i - b) (-1) \quad (3)$$

After doing *some* algebra and rearrangement, we get that

$$m = \frac{n \sum_1^n x_i y_i - \sum_1^n x_i \sum_1^n y_i}{n \sum_1^n x_i^2 - (\sum_1^n x_i)^2} \quad (3)$$

$$b = \frac{\sum_1^n y_i - m \sum_1^n x_i}{n} \quad (3)$$

This may seem like heavy duty calculation, and in fact for large  $n$  it is! Still, it is important enough to invest time developing software programs to carry out the calculations. As a bonus exercise (worth 1 point on *any* exam), find  $m$  and  $b$  when we have only  $n = 2$  data points. In many cases,  $n$  is at least of the order of 100, so we have no choice but to use a software program (or calculator!) Still, if we have only 3 or 4 data points, then it is okay to do things by hand. In fact, you should expect on your final exam (coming up soon) to compute, from scratch, the best fit coefficients for a sample data set of 3 or 4 data points.