We are going to be doing a lot of work with least-squares fitting to extract a light curve from our images. *So let's do the easiest application of least-squares fitting problem, which is* ***linear regression.*** Because a linear function is such an easy fitting function, you don't need the help of a computer. For the light curve analysis, we need the computer, but you can solve the linear regression problem exactly in one page. Here we go....

Assume you have a bunch of data points $(x_i,\ y_i)$. We want to fit the data to the function $y = m\,x + b$. So for every x_i, we have a prediction and the prediction is $m\,x_i + b$. The error in the prediction (actual minus predicted) is $e_i = y_i - (m\,x_i + b)$. The sum of the squares of the errors is $\Sigma_i[y_i - (m x_i + b)]^2$. From now on, whenever I have a Σ_i, I'll just write Σ. You know that it is over all n of the data points,
$i = 1,\ \ldots\ ,\ n$.

In least squares minimization, we want to choose m and b so that $\Sigma[y_i - (m x_i + b)]^2$ is at a minimum. This is a quadratic function in m and b and it is easy to find the minima of quadratic functions. You take the derivative and set it equal to 0.

Q1. Take $\frac{d}{d\,m}$ of $\Sigma[y_i - (m x_i + b)]^2$ and set it equal to zero.

Q2. Take $\frac{d}{d\,b}$ of $\Sigma[y_i - (m x_i + b)]^2$ and set it equal to zero.

The quadratic functions, which are now linear in m and b after you took the derivatives, still have quadratic and linear terms in y_i and x_i. Define the following sums:

$\sigma_{y\,y} \equiv \Sigma\ y_i^2$

$\sigma_{x\,y} \equiv \Sigma\ x_i\ y_i$

$\sigma_{x\,x} \equiv \Sigma\ x_i^2$

$\sigma_y \equiv \Sigma\ y_i$

$\sigma_x \equiv \Sigma\ x_i$

$\sigma_1 \equiv \Sigma\ 1 = n$ (so there is no need to have defined σ_1, because it is just n)

Q3. Using these definitions, rewrite your answer to Q1.

Q4. Again using these definitions, rewrite your answer to Q2.

Q5. It is now obvious that you have two equations in two unknowns. Solve for m.

Q6. Solve for *b*.

Q7. Compare your answer for *m* and *b* with a standard textbook answer for the linear regression formula. If you can’t reconcile your answer, go back and find the mistakes.

DISCUSSION:

You now know where linear regression comes from! Can you see why outlier data points so strongly affect linear regressions? Because the effect of outliers is so strongly weighted by least squares, people often go through their data and manually eliminate outliers before doing the linear regression.

There is some statistical theory that explains why throwing out outliers might be a good thing to do, but just treat it as a standard practice, because the theory justifying least squares is an entirely separate derivation. The theory is related to the “central limit theorem.”