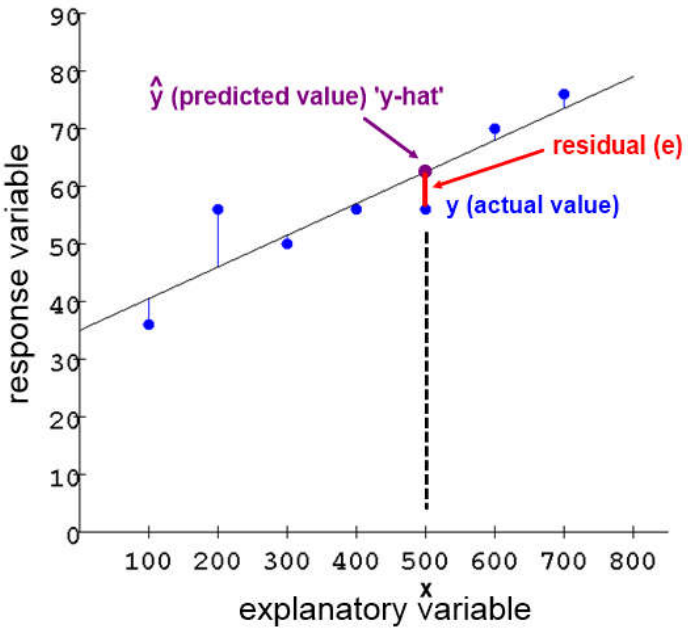
Derivation of equations used for calculating the Least Squares Regression Line (LSRL) for a data set

In statistics, we often need to 'fit a line' to a set of data in such a way that the errors between the predicted y value for each x and the actual y value for that x are as small as possible. We can use multivariable optimization techniques to derive the needed equations for computation.

We start by defining the residual, *e*, as the difference between actual y and predicted y (actual – predicted) at any given x:



A **residual** is the error between what the LSRL line predicts the y value will be (for a given x) and the actual y value.

$$e = y - \hat{y}$$

We then define a Least Squares Regression Line (LSRL) which will be the resulting equation: $\hat{y} = a + bx$
Where y-hat represents the predicted y output from the LSRL equation at input value x.

The error is then given by: $e = y - \hat{y} = y - (a + bx)$

What we are really solving for are the two coefficients, *a* and *b* which result in a minimum total error across all the points. We can define a function for error in terms of *a* and *b*:

$$total\ error\ (squared) = f(a,b) = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2$$

The reason we square is to force all errors both above and below the line to be positive (we don't want the negative errors cancelling out the positive errors).

If we want to optimize this function (find its minimum value) we will take the two partial derivatives, set them equal to zero and solve for critical points, then use the Hessian determinant to verify that this is a minimum.

First, expanding the function:

$$f(a,b)=\sum_{i=1}^{n}\left(y_i-\left(a+bx_i\right)\right)^2$$

$$f(a,b)=\sum_{i=1}^{n}\left({y_i}^2-2y_i\left(a+bx_i\right)+\left(a+bx_i\right)^2\right)$$

$$f(a,b)=\sum_{i=1}^{n}\left({y_i}^2-2y_ia-2y_ibx_i+a^2+2abx_i+b^2{x_i}^2\right)$$

We take the first derivatives and set them equal to zero:

$$\begin{cases} f_a=\sum_{i=1}^{n}\left(-2y_i+2a+2bx_i\right)=0 \\ f_b=\sum_{i=1}^{n}\left(-2y_ix_i+2ax_i+2b{x_i}^2\right)=0 \end{cases}$$

These form a system. Working further with the f_a equation:

$$f_a=\sum_{i=1}^{n}\left(-2y_i+2a+2bx_i\right)=0$$

$$\sum_{i=1}^{n}\left(-2y_i+2a+2bx_i\right)=0$$

$$(-2)\sum_{i=1}^{n}\left(y_i-a-bx_i\right)=0$$

$$\sum_{i=1}^{n}\left(y_i-a-bx_i\right)=0$$

$$\sum_{i=1}^{n}\left(y_i\right)=\sum_{i=1}^{n}\left(a\right)+\sum_{i=1}^{n}\left(bx_i\right)$$

$$\sum_{i=1}^{n}\left(y_i\right)=a\sum_{i=1}^{n}\left(1\right)+b\sum_{i=1}^{n}\left(x_i\right)$$

$$\left(\sum_{i=1}^{n}\left(1\right)\ is\ \ just\ n\right)$$

$$\sum_{i=1}^{n}\left(y_i\right)=na+b\sum_{i=1}^{n}\left(x_i\right)$$

Working further with the f_b equation:

$$f_b = \sum_{i=1}^{n}\left(-2y_i x_i + 2ax_i + 2bx_i^{\,2}\right) = 0$$

$$\sum_{i=1}^{n}\left(-2y_i x_i + 2ax_i + 2bx_i^{\,2}\right) = 0$$

$$(-2)\sum_{i=1}^{n}\left(y_i x_i - ax_i - bx_i^{\,2}\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i x_i - ax_i - bx_i^{\,2}\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i x_i\right) = \sum_{i=1}^{n}\left(ax_i\right) + \sum_{i=1}^{n}\left(bx_i^{\,2}\right)$$

$$\sum_{i=1}^{n}\left(y_i x_i\right) = a\sum_{i=1}^{n}\left(x_i\right) + b\sum_{i=1}^{n}\left(x_i^{\,2}\right)$$

So now we can write the system like this:

$$\begin{cases} f_a = \sum_{i=1}^{n}\left(-2y_i + 2a + 2bx_i\right) = 0 \\ f_b = \sum_{i=1}^{n}\left(-2y_i x_i + 2ax_i + 2bx_i^{\,2}\right) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n}\left(y_i\right) = na + b\sum_{i=1}^{n}\left(x_i\right) \\ \sum_{i=1}^{n}\left(y_i x_i\right) = a\sum_{i=1}^{n}\left(x_i\right) + b\sum_{i=1}^{n}\left(x_i^{\,2}\right) \end{cases}$$

$$\begin{cases} [n]a + \left[\sum_{i=1}^{n}\left(x_i\right)\right]b = \left[\sum_{i=1}^{n}\left(y_i\right)\right] \\ \left[\sum_{i=1}^{n}\left(x_i\right)\right]a + \left[\sum_{i=1}^{n}\left(x_i^{\,2}\right)\right]b = \left[\sum_{i=1}^{n}\left(y_i x_i\right)\right] \end{cases}$$

The expressions in the brackets are numerical values computed from the data set. To make this easier to see, we can replace each with a constant (which would be calculated from the data set):

$$Define:$$

$$Q=\left[\sum_{i=1}^{n}\left(x_i\right)\right]$$

$$R=\left[\sum_{i=1}^{n}\left(y_i\right)\right]$$

$$S=\left[\sum_{i=1}^{n}\left(x_i^{\,2}\right)\right]$$

$$T=\left[\sum_{i=1}^{n}\left(y_i x_i\right)\right]$$

Then the system would become:

$$\begin{cases} [n]a+\left[\sum_{i=1}^{n}\left(x_i\right)\right]b=\left[\sum_{i=1}^{n}\left(y_i\right)\right] \\ \left[\sum_{i=1}^{n}\left(x_i\right)\right]a+\left[\sum_{i=1}^{n}\left(x_i^{\,2}\right)\right]b=\left[\sum_{i=1}^{n}\left(y_i x_i\right)\right] \end{cases}$$

$$\begin{cases} na+Qb=R \\ Qa+Sb=T \end{cases}$$

Which can then be solved using RREF for the LSRL coefficients *a* and *b*.

How do we know this is where total squared error is minimized? (As opposed to maximized or a saddle point)

To verify this is a minimum, we would need to determine the value of the Hessian Determinant, so first we need expressions for the 2nd derivatives:

$$f_a = \sum_{i=1}^{n}\left(-2y_i + 2a + 2bx_i\right)$$

$$f_{aa} = \sum_{i=1}^{n}\left(2\right) = 2n$$

$$f_{ab} = \sum_{i=1}^{n}\left(2x_i\right)$$

$$f_b = \sum_{i=1}^{n}\left(-2y_i x_i + 2ax_i + 2bx_i^{\,2}\right)$$

$$f_{bb} = \sum_{i=1}^{n}\left(2x_i^{\,2}\right)$$

The Hessian determinant would then be:

$$\begin{aligned} D &= f_{aa}f_{bb} - \left(f_{ab}\right)^2 \\ &= \left(2n\right)\left(\sum_{i=1}^{n}\left(2x_i^{\,2}\right)\right) - \left(\sum_{i=1}^{n}\left(2x_i\right)\right)^2 \\ &= 4n\sum_{i=1}^{n}\left(x_i^{\,2}\right) - 4\sum_{i=1}^{n}\left(x_i^{\,2}\right) \\ &= 4\sum_{i=1}^{n}\left(x_i^{\,2}\right)\left(n-1\right) \end{aligned}$$

…which is definitely positive, so a maximum or a minimum. Looking at concavity f_{aa} to check: $f_{aa} = 2n$ is also definitely positive, which means concave up, which verifies that these values of *a* and *b* produce the minimum total square error.