# Regression, Geometry, and Symmetry

Darryl Nester

Bluffton University
Bluffton, Ohio

Mathematics & Statistics Conference
Miami University
October 1, 2004

www.bluffton.edu/mat/seminar/

**Question: Why are the mean and median good representatives for a list of numbers?**

Answer: Our goal is to choose a number which is simultaneously as close as possible to every number in the list (that is, its "total distance" is minimized).

If "distance" means *total squared difference*, then we choose $\bar{x}$. If it means *total magnitude (absolute value) of the differences*, choose the median:

$$x = \bar{x} \quad \text{minimizes} \quad \sum_i (x_i - x)^2$$

$$x = M \quad \text{minimizes} \quad \sum_i |x_i - x|$$

**Why do we prefer $\bar{x}$ over $M$ in many settings?**

In computer terms, the mean is typically easier to compute (for large data sets).

More significantly, it is analytically easier (in terms of taking derivatives, etc.) to deal with $\bar{x}$. (Absolute values are a pain!)

There are also some standard assumptions in statistics that make $\bar{x}$ a natural choice (or at least a more convenient one).

## Why is the usual regression line based on minimizing total squared vertical distance $S_1 = S_1(a,b) = \sum_i (y_i - ax_i - b)^2$ ?

See the previous answers: It is computationally and analytically easier, and makes sense in terms of some standard statistical assumptions.

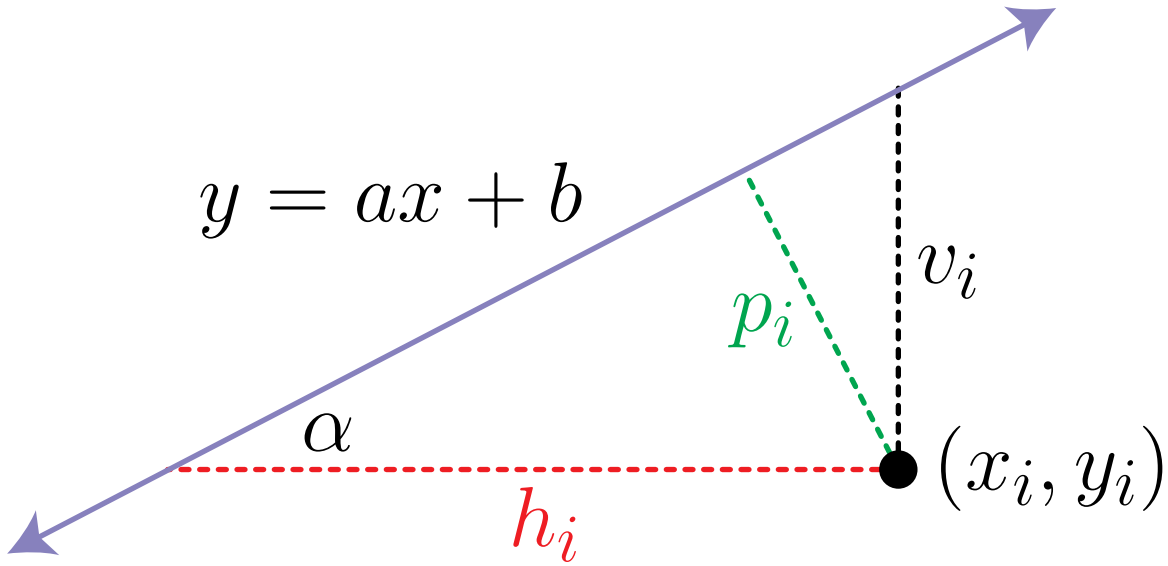
But if we put those reasons aside, we can still explore what happens when we seek to minimize other distance functions, such as ...

$$S_1 = \sum_i v_i^2 \qquad A_1 = \sum_i |v_i|$$

$$S_2 = \sum_i h_i^2 \qquad A_2 = \sum_i |h_i|$$

$$S_3 = \sum_i p_i^2 \qquad A_3 = \sum_i |p_i|$$

$$S_4 = \sum_i |h_i v_i|$$



Note that $a = \tan\alpha$, so that

$$h_i = \frac{v_i}{a} \quad \text{and} \quad p_i = \frac{v_i}{\sqrt{1+a^2}}$$

and therefore

$$S_2 = \frac{S_1}{a^2}, \quad S_3 = \frac{S_1}{1+a^2}, \quad S_4 = \frac{S_1}{|a|},$$

$$A_2 = \frac{A_1}{|a|}, \quad A_3 = \frac{A_1}{\sqrt{1+a^2}}$$

So, given slope $a$, the intercept $b$ which minimizes:

- $S_1$ also minimizes $S_2$, $S_3$, and $S_4$.
- $A_1$ also minimizes $A_2$ and $A_3$.

Given $a$, let $c_i = y_i - ax_i$. How do we choose $b$ to minimize

$$S_1 = \sum_i (y_i - ax_i - b)^2 = \sum_i (c_i - b)^2 \ ?$$

Take $b$ to be the **mean** of $c_1, c_2, \ldots, c_n$; that is, $b = \bar{y} - a\bar{x}$. (This is well-known.)

How do we minimize

$$A_1 = \sum_i |y_i - ax_i - b| = \sum_i |c_i - b| \ ?$$

Take $b$ to be the **median** of $c_1, c_2, \ldots, c_n$. (Not so well-known, but nicely similar to the result for squared distances.)

With these choices of $b$, these distance functions depend on one variable ($a$). For $S_k$, we can optimize with calculus; for $A_k$, we optimize numerically.

The accompanying applet allows us to see:

- the effects of different distance functions on the "best" line,
- how the line changes as we move individual points,
- the relationship between slope and correlation,
- what happens when we swap $x$ and $y$, and
- the effect of standardizing (rescaling and re-centering) $x$ or $y$.

With

$$S_{xy} = \sum\nolimits_i (x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum\nolimits_i (x_i - \bar{x})^2, \quad S_{yy} = \sum\nolimits_i (y_i - \bar{y})^2,$$

the minimizing slopes are

$S_1$: $a = S_{xy}/S_{xx}$

$S_2$: $a = S_{yy}/S_{xy}$

$S_3$: $a = \frac{1}{2}(k \pm \sqrt{k^2 + 4})$ where $k = \dfrac{S_{yy} - S_{xx}}{S_{xy}}$

$S_4$: $a = \pm\sqrt{S_{yy}/S_{xx}}$

For both $S_3$ and $S_4$, take the sign of the square root as the sign of $S_{xy}$.