

## Derivation of the Equation for Least Squares Line of Best Fit

(Excerpted from *Stats: Modeling the World* and the accompanying *Teacher's Guide*.)

**Step 1.** To get at the equation of the line, we'll work with z-scores. We need to remember a few things.

1) The mean of any set of z-scores is 0 ( $\bar{z} = 0$ ), so the sum is also 0:  $\sum z = 0$

2) The standard deviation of a set of z-scores is 1, so the variance is also 1:

$$\frac{\sum (z_y - \bar{z}_y)^2}{n-1} = \frac{\sum (z_y - 0)^2}{n-1} = \frac{\sum z_y^2}{n-1} = 1.$$

3) The correlation is  $r = \frac{\sum z_x z_y}{n-1}$ .

**Step 2.** Standardize the data for each variable and consider the scatterplot of points  $(z_x, z_y)$ . We seek the

line  $\hat{z}_y = a + mz_x$  that minimizes

$$\sum [z_y - \hat{z}_y]^2$$

Substitute the equation for  $\hat{z}_y$ :

$$\sum [z_y - (a + mz_x)]^2$$

Rearrange terms:

$$\sum [(z_y - mz_x) - a]^2$$

Square the binomial:

$$\sum [(z_y - mz_x)^2 - 2a(z_y - mz_x) + a^2]$$

Consider the "middle term"; by (2):

$$\sum [2a(z_y - mz_x)] = 2a \sum z_y - 2am \sum z_x = 0$$

We need to minimize what's left:

$$\sum [(z_y - mz_x)^2 + a^2]$$

By choosing  $a = 0$  we can be sure that the sum will be minimal. (Adding the square of any other value would make it bigger.) Hence the best line must have a  $y$ -intercept of 0 in the standardized plane, proving that the line of regression goes through the origin  $(0,0)$  in the standardized plane and hence through the mean-mean point  $(\bar{x}, \bar{y})$  in the  $xy$ -plane.

**Step 3.** Now we have to find the slope that minimizes the sum of the squared residuals. Because the line passes through the origin, its equation will be of the form  $\hat{z}_y = mz_x$ . We want to find the value for  $m$  that will minimize the sum of the squared residuals. Actually we'll divide that sum by  $n - 1$  and minimize this "mean squared residual", or MSR. Here goes:

Minimize: 
$$MSR = \frac{\sum (z_y - \hat{z}_y)^2}{n-1}.$$

Since  $\hat{z}_y = mz_x$ : 
$$MSR = \frac{\sum (z_y - mz_x)^2}{n-1}$$

Square the binomial: 
$$= \frac{\sum (z_y^2 - 2mz_x z_y + m^2 z_x^2)}{n-1}$$

Rewrite the summation: 
$$= \frac{\sum z_y^2}{n-1} - 2m \frac{\sum z_x z_y}{n-1} + m^2 \frac{\sum z_x^2}{n-1}$$

Substituting from (2) and (3): 
$$MSR = 1 - 2mr + m^2$$

That simplified nicely! And as a bonus, the last expression is quadratic. Remember parabolas from algebra class? A parabola in the form  $y = ax^2 + bx + c$  reaches its minimum at its turning point, which occurs when  $x = \frac{-b}{2a}$ . We can minimize the mean of squared residuals by choosing  $m = \frac{-(-2r)}{2(1)} = r$ .

#### Step 4. CONCLUSIONS

1. In the standardized plane, the slope of the best fit line for z-scores is the correlation,  $r$ .
2. That slope is “over 1, up  $r$ ” for the z-scores, meaning “over 1 standard deviation in  $x$ , up one standard deviation in  $y$ ”. In the  $xy$ -plane, that’s “over  $1s_x$ , up  $rs_y$ ”, making the slope of the regression line  $b = \frac{rs_y}{s_x}$ .
3. In the  $xy$ -plane, the line passes through the mean-mean point  $(\bar{x}, \bar{y})$ .
4. The parabola  $MSR = 1 - 2r + r^2$  has its minimum when  $m = r$ . Substituting into the function we see that the minimum  $MSR = 1 - r^2$ . Since  $MSR$  is the sum of squares, it cannot be negative. Therefore  $1 - r^2 \geq 0$ , proving that  $-1 \leq r \leq 1$ .
5. In Step 1 we saw that the variance of  $z_y$  measured from the mean was  $\frac{\sum (z_y - \bar{z}_y)^2}{n-1} = 1$ . Now we see that the variance of  $z_y$  measured from the line (the residuals) is  $MSR = \frac{\sum (z_y - \hat{z}_y)^2}{n-1} = 1 - r^2$ . That shows the line has removed (accounted for)  $r^2$  of the variability in  $y$ .  
(To examine what this means, try the attached worksheet.)

### What's $R^2$ anyway?

HT(in)	WT(lb)
67	140
71	165
73	168
71	142
74	200
74	175
68	135
73	145
71	150
72	155
69	168
66	106
70	144
71	132
70	140
71	140
70	140
69	130
70	150
74	170
71	175
74	180
72	150
70	150
73	190

- ◇ Enter these height-weight data into L1 (HT =  $x$ ) and L2 (WT =  $y$ ).
- ◇ We're looking for a way to estimate the weight ( $y$ ) of a mystery person. Sometimes we'll guess too high, other times too low. We seek a method that will minimize the errors we make *in the long run*. Absent any information other than this set of weights, what's the best weight to guess? \_\_\_\_\_
- ◇ The errors made using this guess-the-mean strategy would be  $(y - \bar{y})$ . Create a list of these errors: L2 -  $\bar{y}$  STO L3.
- ◇ Now take a look at the errors by making a boxplot of L3.
- ◇ Use TRACE to get a feel for the distribution of these errors. What's that vertical line down the middle of the screen? \_\_\_\_\_ What do the errors to the left of that line represent? \_\_\_\_\_ To the right? \_\_\_\_\_
- ◇ We want a quantitative measure of the overall amount of error. Simply summing the errors won't work. Why not? What *must* that total be? \_\_\_\_ Check by doing  $\text{sum}(L3)$ .
- ◇ You know the usual Statistics trick, of course: sum the *squares* of the errors. Squaring makes them all positive (better for adding) and places greater emphasis on the larger errors we hope to avoid. Find the sum of the squared errors:  $\text{sum}(L3^2) =$  \_\_\_\_\_
- ◇ Now suppose that, rather than just having to guess blindly, we knew something useful about the mystery person. Let's say, um, height (surprise)! Using this additional info should enable us to make a better guess about weight. Find the equation that we'd use to predict weight from height. \_\_\_\_\_
- ◇ How strong is the relationship?  $r =$  \_\_\_\_ (And humor me...  $R^2 =$  \_\_\_\_%)
- ◇ So, how much better is using the regression line than the blind guess-the-average method? Let's look at the resulting errors,  $(y - \hat{y})$ , also known as \_\_\_\_\_.
- ◇ Move the residuals to L4 so we can work with them: (list name) LRESID STO L4. We hope this linear model method's errors are generally smaller than the first batch were. Think: how should the boxplot of these errors compare to the one we looked at before? \_\_\_\_\_ Check that out by creating a parallel boxplot. Happy?
- ◇ The two boxplots show us that taking the mystery person's height into account generally reduces errors in our estimates of weight. Quantitatively, how much better is this method? Find the sum of the squares of these errors:  $\text{sum}(L4^2) =$  \_\_\_\_\_
- ◇ Overall error is now lots smaller! What percent of the original error still remains? \_\_\_\_%
- ◇ What percent of the original error has been removed by using the regression line? \_\_\_\_% Notice anything interesting? Explain: