

SD/Correlation Computations in Easy Steps

Math 1070-1, Spring 2003 (Univ. of Utah)

Example: (SL-County unemployment data)

Year	(x)	1997	1998	1999	2000	2001
Rate in %	(y)	2.7	3.4	3.4	3	4.3

Sample size: $n = 5$.

$$\begin{aligned}\text{Means: } \bar{x} &= (1997 + \dots + 2001)/5 = 1999, \\ \bar{y} &= (2.7 + \dots + 4.3)/5 = 3.36.\end{aligned}$$

The standard Deviation of x : (First the square)

$$S_x^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

What does this mean?

$x - \bar{x}$	-2	-1	0	1	2
$(x - \bar{x})^2$	4	1	0	1	4

So: $S_x^2 = \frac{1}{4} (4 + 1 + 0 + 1 + 4) = 2.5$. Therefore, $S_x = \sqrt{2.5} \approx 1.581139$ years (without rounding).

Also,

$y - \bar{y}$	-0.66	0.04	0.04	-0.36	0.94
$(y - \bar{y})^2$	0.44	0	0	0.13	0.88

So:

$$S_y^2 \approx \frac{1}{4} (0.44 + 0 + 0 + 0.13 + 0.88) \approx 0.363.$$

Therefore, $S_y \approx \sqrt{0.363} \approx 0.6024948\%$ (without rounding). For correlation, let me start by reminding you of the formula:

$$\begin{aligned} r &= \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{S_x} \right) \left(\frac{y - \bar{y}}{S_y} \right) \\ &= \frac{1}{n-1} \sum SU_x SU_y, \end{aligned}$$

where SU means "in standard units." In other words, the above says, "*first compute a column of x in standard units and one for y . Then cross-multiply and add. Finally, divide by $n - 1$.*" Now we are off to work out the details which I will take pains to do very meticulously so as to avoid those silly—and unacceptable—errors.

Recall that $SU_x = (x - \bar{x})/S_x$. So:

x	1997	1998	1999	2000	2001
$x - \bar{x}$	-2	-1	0	1	2
SU_x	-1.3	-0.6	0	0.6	1.3

Ditto for the y 's:

y	2.7	3.4	3.4	3	4.3
$y - \bar{y}$	-0.66	0.04	0.04	-0.36	0.94
SU_y	-1.1	0.07	0.07	-0.6	1.56

SU_x	-1.3	-0.6	0	0.6	1.3
SU_y	-1.1	0.07	0.07	-0.6	1.56
$SU_x SU_y$	1.39	-0.04	0	-0.38	1.97

So

$$r = \frac{1}{4} \left[1.39 + (-0.04) + 0 + (-0.38) + 1.97 \right]$$
$$\approx 0.7348094 \text{ (no rounding).}$$

Regression The equation of the regression line is: $SU_y = rSU_x$. I.e.,

$$\frac{y - \bar{y}}{S_y} = r \left(\frac{x - \bar{x}}{S_x} \right).$$

Solve for y (DO IT!) to obtain:

$$\begin{aligned} y &= rS_y \left(\frac{x - \bar{x}}{S_x} \right) + \bar{y} \\ &= \underbrace{\left(\frac{rS_y}{S_x} \right)}_{\text{(slope)}} x + \underbrace{\left[\bar{y} - \left(\frac{rS_y}{S_x} \right) \bar{x} \right]}_{\text{(intercept)}}. \end{aligned}$$

In our Example above, we had

$$\bar{x} = 1999$$

$$S_x \approx 1.58$$

$$\bar{y} = 3.36$$

$$S_y \approx 0.6$$

$$r \approx 0.73.$$

So slope = $(rS_y/S_x) \approx (0.73 \times 0.6/1.58) = 0.28$ (without rounding). Similarly, intercept = -556.36 (without rounding; check this!) So, the regression line—in the previous Example—is:

$$y = 0.28x - 556.36.$$

The regression-prediction for the unemployment in SL-county in the year $x = 2001$ (based on the above data):

$$y \approx 0.28 \times 2001 - 556.36 = 3.92\%.$$