Derivation of the equation $b = r\frac{s_y}{s_x}$

This is a key equation used when we don't have a complete set of data, but still want to know things about the slope of the LSRL or to create an LSRL using summary data only. The derivation is based upon thinking about where the equation of the line with minimum residuals comes from. To write that line, which we know goes through the point $\left(\overline{x}, \overline{y}\right)$ we need the following supporting facts:

a) The mean of any distribution of z-scores is 0. This tells us that if we standardized all the data values into z-scores, the LSRL for the z-scores would pass through (0, 0).

b) The standard deviation of any distribution of z-scores is 1, so the variance $\left(Var = \sqrt{s}\right)$ is also 1.

For the y distribution (of z-scores): $Var = \frac{\sum\left(z_y - \overline{z_y}\right)^2}{n-1} = \frac{\sum\left(z_y - 0\right)^2}{n-1} = \frac{\sum z_y{}^2}{n-1} = 1$

c) The correlation coefficient (for z-scores), $r = \frac{\sum z_x z_y}{n-1}$

Now, to derive the slope formula. Our objective is to find the slope of the LSRL, initially for a scatterplot where the data has been standardized into y z-scores vs. x z-scores.

1) The LSRL equation would be a line: $y = a + mx$ and here would be: $\widehat{z_y} = a + mz_x$

2) But because the LSRL will pass through (0,0) there is no y-intercept (for the z-scores): $\widehat{z_y} = mz_x$

We are looking for the value of the slope, m, that will minimize the sum of the squared residuals.

3) So are are minimizing the sum of the squared residuals $= \sum\left(z_y - \widehat{z_y}\right)^2$

It makes the computation easier if we go ahead and divide this by n-1 (which is something called the 'mean squared residual' or MSR).

4) So we'll actually find m to minimize the $MSR = \frac{\sum\left(z_y - \widehat{z_y}\right)^2}{n-1}$

5) Since $\widehat{z_y} = mz_x$, we can replace that in the MSR equation: $MSR = \frac{\sum\left(z_y - mz_x\right)^2}{n-1}$

6) Expanding the binomial we get: $MSR = \frac{\sum\left(z_y{}^2 - 2mz_x z_y + m^2 z_x{}^2\right)}{n-1}$

7) Splitting the summation for each term and moving everything that is constant outside the summations:

$$MSR = \frac{\sum z_y{}^2}{n-1} - 2m\frac{\sum z_x z_y}{n-1} + m^2\frac{\sum z_x{}^2}{n-1}$$

8) But from (b) $\frac{\sum z_y{}^2}{n-1} = 1$, and from (c) $\frac{\sum z_x z_y}{n-1} = r$. Substituting these:

$$MSR = 1 - 2mr + m^2 \quad or \quad MSR = m^2 - (2r)m + 1$$

9) This is a quadratic. To minimize MSR, we need to find the value of m at the vertex. A quadratic of form $ax^2 + bx + c$ has its vertex where $x = \frac{-b}{2a}$ so the MSR (residuals) will be minimized when:

$$m = \frac{-(-2r)}{2(1)}$$

$$m = r$$

This means that when a set of data is standardized into z-scores for both the x and y variables, the LSRL (line where residuals is minimized) occurs when the **slope of the LSRL equals the correlation coefficient**.

10) To find the slope of the LSRL with the original data (not standardized into z-scores) we can recall the z-score equation for x and y-hat:

$$z_x = \frac{x - \bar{x}}{s_x}, \qquad \widehat{z}_y = \frac{\hat{y} - \bar{y}}{s_y}$$

11) Starting with the LSRL equation for z-scores, we can replace the z-score terms, and use r for the slope:

$$\frac{\hat{y} - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right)$$

12) Multiplying by s_y and distributing:

$$\hat{y} - \bar{y} = rs_y\left(\frac{x - \bar{x}}{s_x}\right)$$

$$\hat{y} - \bar{y} = r\frac{s_y}{s_x}x - r\frac{s_y}{s_x}\bar{x}$$

13) Rearranging:

$$\hat{y} = \left(y - r\frac{s_y}{s_x}\bar{x}\right) + \left(r\frac{s_y}{s_x}\right)x$$

14) This means we can identify the slope (and y-intercept) of the LSRL for the original data:

$$\hat{y} = a + bx$$

$$a = y - r\frac{s_y}{s_x}\overline{x}$$

$$b = r\frac{s_y}{s_x}$$

So the slope of the LSRL will be: $b = r\frac{s_y}{s_x}$