Analysis: Why does r² represent the percentage of the variability explained by the LSRL?

The beginning of this analysis is the same as the analysis for deriving and equation for the slope of the LSRL. We know that the LSRL goes through the point $\left(\overline{x}, \overline{y}\right)$, and 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.

c) The correlation coefficient (for z-scores) would be: $r = \dfrac{\sum z_x z_y}{n-1}$

Our objective with an LSRL is to find the slope and intercept for the LSRL which minimizes the sum of the square of the residuals. Here, we will standardize the data into z-scores.

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

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

3) We are looking for the value of the slope, m, that will minimize the sum of the square residuals, so we are minimizing $\sum\left(z_y - \hat{z}_y\right)^2$

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

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

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

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

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) the variance of x is 1, so $\frac{\sum z_x{}^2}{n-1} = \frac{\sum z_y{}^2}{n-1} = 1$. Also, from (c) $r = \frac{\sum z_x z_y}{n-1}$. Substituting these we get: $MSR = 1 - 2mr + m^2 1$
or: $MSR = m^2 - (2r)m + 1$

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

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

$$m = r$$

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 are minimizes) occurs when the **<u>slope of the LSRL equals the correlation coefficient</u>**.

10) So now we can ask, "what is this minimum mean square of the residuals?" Setting m = r:

$$MSR = r^2 - (2r)r + 1$$

$$MSR = r^2 - 2r^2 + 1$$

$$MSR = 1 - r^2$$

This is the part of the variability around the LSRL, so it represents a measure of the variability which remains once we've accounted for as much variability as possible that can be explained by the LSRL relating y to x. (The amount of the variability *not* explained by the LSRL).

11) That means that 1 – this amount represents the amount of the variability which *is* explained by the LSRL:

$$amount\ of\ variability\ explained\ by\ LSRL = 1 - (1 - r^2)$$

$$amount\ of\ variability\ explained\ by\ LSRL = r^2$$

12) If we had perfect correlation (r=1) that would mean that 100% of the variation in the data is being explained by the LSRL relating y to x. This is why the maximum possible value of r is 1 (or -1) and also explains why r^2 (which is also called the coefficient of determination) is a measure of the percentage of the y variation which is explained by the LSRL which relates y to x.