The Sums of Squares

Regression, correlation, Analysis of Variance, and other important statistical models all rely on a single key concept, the sum of the squared deviations of a quantity from its mean. You saw this in elementary statistics as the numerator for the variance of a variable.

I will discuss them in the context of simple linear regression.

The "Total Sum of Squares" SST = SSyy = \[ \sum_{i=1}^{n} (y_i - \bar{y})^2 \] measures the variability of the dependent variable, It is equal to n-1 times the sample variance of y: \[ s_y^2 = \frac{SSyy}{n-1}. \]

SSxx = \[ \sum_{i=1}^{n} (x_i - \bar{x})^2 \] measures the variability of the independent variable. It is equal to n-1 times the sample variance of x: \[ s_x^2 = \frac{SSxx}{n-1}. \]

SSxy \[ = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \] measures the tendency of x and y to vary together. It can be negative if high x goes with low y, positive if high x goes with high y, or zero if x and y are unrelated. It is n-1 times the covariance of x and y: Cov[x,y] = \[ \frac{SSxy}{n-1}. \] Note that the covariance of any variable with itself is its variance: Cov[x,x] = \[ s_x^2 \]

The estimated slope of the regression line, \( b_1 \), is the covariance divided by the variance:
\[ b_1 = \frac{SSxy}{SSxx}. \]
(Th e book also gives alternate forms that get the same answer with less arithmetic, but our goal here is to master the concepts, and the shortcut forms obscure those.)

The estimated y-intercept of the regression line is \( b_0 = \bar{y} - b_1 \bar{x} \)

The point estimate corresponding to each specific \( y_i = b_0 + b_1 x_i \).
The "Residual" for each specific \( y_i \) is \( y_i - (b_0 + b_1 x_i) \)

The Error Sum of Squares SSE is the sum of the squared residuals: \[ SSE = \sum_{i=1}^{n} (y_i - \frac{b_0 + b_1 x_i}{})^2 \]
The variance of the residuals is \( s^2 = \frac{SSE}{n-2} \); the standard error s is the square root of this quantity.

The Regression Sum of Squares SSR measures the total amount of variation in y that is accounted for ("explained") by the variation in x: \[ SSR = SST - SSE = SSyy - SSE \]

The Simple Coefficient of Determination \( R^2 = \frac{SSR}{SST} = \frac{SSyy-SSE}{SSyy} \).