Linear Regression

www.FairlyNerdy.com

With linear regression, you are trying to reduce a set of data into a line. The equation for a line is

$$y = \underset{\text{Intercept}}{a} + \underset{\text{Slope}}{b}x$$

The first thing to calculate is the slope of the regression line. That value is the correlation of the two data sets, multiplied by the ratio of their standard deviations. Note that those are sample standard deviations, not population

$$\overset{\text{Slope}}{b} = \underset{\text{Pearson's Correlation}}{r} * \frac{\overset{\text{Sample Standard Deviation of y}}{s_y}}{\underset{\text{Sample Standard Deviation of x}}{s_x}}$$

Correlation shows how much two sets of data change together. Correlation is always between -1 and 1, and is unit-less. Correlation is frequently around the average x, average y, but if you want to force the line through a specific point, you can use that point instead of average x, y in all the

$$\underset{\text{Pearson's Correlation}}{r} = \frac{\overset{\text{Sum Over All Data Points}}{\sum}\left(\overset{\text{x \& y values of each point minus x \& y mean values}}{(x - \bar{x}) * (y - \bar{y})} \right)}{(\underset{\text{\# of Data Points}}{n} - 1) * \underset{\text{Standard Deviation of x \& y}}{s_x * s_y}}$$

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}} \qquad s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{(n-1)}}$$

The sample standard deviation of x and y measure how spread out the x and y values are around their mean. They have units

Once you have done those calculations you have the slope. With the slope and a point the line goes through you can calculate the intercept. If you used the average x, y before use it here again. Otherwise use the same x, y that you specified before

$$\overset{\text{Intercept}}{a} = \bar{y} - \overset{\text{Slope}}{b}\bar{x}$$

Y & X Average

When you have done the regression, one way of evaluating it's quality is R-squared. R-squared is a measure of the summed squared error in the regression vs the error if you didn't do a regression. Summed Squared error is shown below

$$R^2 = 1 - \frac{\overset{\text{Sum Squared Regression Error}}{SS_{Regression}}}{\underset{\text{Sum Squared Total Error}}{SS_{Total}}}$$

$$\underset{\text{Sum Squared Total Error}}{SS_{Total}} = \overset{\text{Sum Over All Data Points}}{\sum} (\underset{\text{Each Data Point}}{y_i} - \underset{\text{Mean Value}}{\bar{y}})^{\overset{\text{Square The Result}}{2}}$$

$$\underset{\text{Sum Squared Regression Error}}{SS_{Regression}} = \overset{\text{Sum Over All Data Points}}{\sum} (\underset{\text{Each Data Point}}{y_i} - \underset{\text{Regression Value}}{y_{Regression}})^{\overset{\text{Square The Result}}{2}}$$

Visit www.FairlyNerdy.com for more FREE Engineering Cheat Sheets

Print This, Save It, Or Share With A Friend!