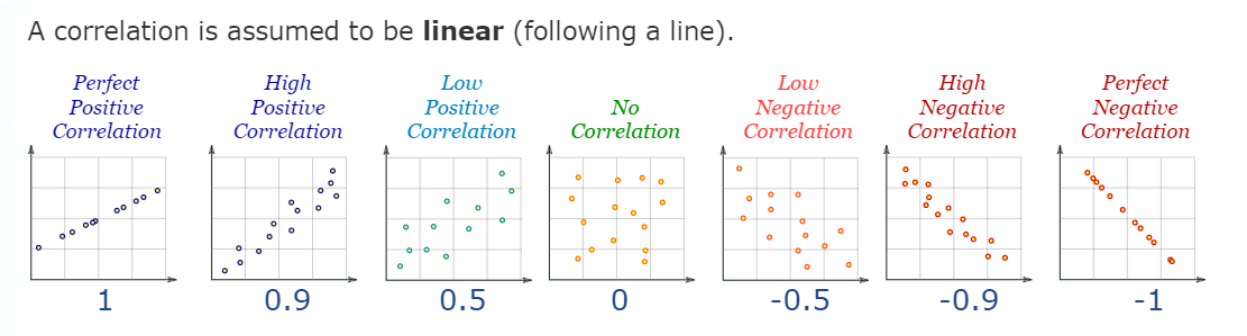
4/22/2023

Scatterplots
Correlation and Coefficient of Determination
Linear Regression

Scatterplots: are used to show the relationship between two numerical (paired) variables.
Typically used with continuous data.

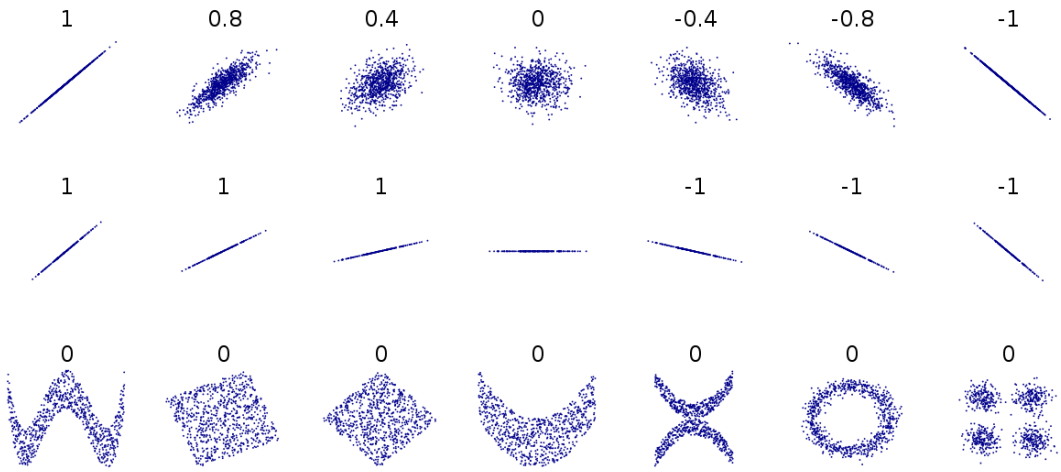
We think of the variable on the horizontal axis as the explanatory variable (typically it's the one we know first, or the one that is easier to measure). The one on the vertical axis is the response variable, and that is the one that is known second, or is harder to measure. For this reason, we want to be able to use the first variable to predict the second one.

Initially, we will use the scatterplot to determine the strength of the relationship (if there is one). And they type of relationship: is the relationship linear or not?



Correlation is a measure of the relationship between two variables (typically in a scatterplot). Correlation measures linear correlation: the strength of the linear relationship between the variables.

Correlation ranges from -1 to 1. -1 is a strong linear relationship with a negative slope (down to the right). And 1 is a strong linear relationship with a positive slope (up to the right). 0 means there is no linear relationship: could mean no relationship at all, or a strongly non-linear relationship.



Correlations that are negative go with negative slopes, but the values are different.
Correlations that are positive go with positive slopes, but the values are different.
Non-linear relationships can produce 0 correlations… so a 0 correlation does not mean there is no relationship, only that if there is one, it's nonlinear. We can only tell which one from looking at the scatterplot.

https://www.guessthecorrelation.com/

The variable name for correlation is r (for a sample), and ρ (rho) for the population.

The strength of the correlation is generally put into three broad groups: strong, moderate and weak.

A strong correlation has $|r| > 0.7$.
A moderate correlation has $0.4 < |r| < 0.7$
A weak correlation has $|r| < 0.4$

Coefficient of Determination, r^2 or ρ^2, or R^2

The percent of the variability in the y-variable is reduced (or accounted for) by the relationship between the y-variable and the x-variable.

If we only consider the y-variable, it has a certain amount of variation (variance). When we account for the relationship to x (via regression), we end with a smaller amount of variation.

The higher R^2 is the better. Can only be values between 0 and 1.

Can calculate correlation from the R-squared value, by taking the square root… but watch the sign, must match the slope.

Linear Regression equation – line of best fit, ordinary least squares line

We want to model the trend with a straight line. We can think of the predicted value we will get from the line as the mean (average) of observations if taken repeatedly with the same x-value.

We represent the line with an intercept and a slope. $y = mx + b$

What does the regression equation we get mean?

$$y = 34.702x + 48621$$

y is overhead, and x is machine hours.

How do we interpret the intercept? The intercept is the value of overhead when machine hours is zero. It's worth noting that the intercept cannot always be interpreted easily.

How do we interpret the slope. The slope is a rate of change of the y-value relative the x-value. Units of the slope are always y-units/x-units. Units are dollars of overhead per machine hour.

In this example we'd say that for each additional machine hour, the overhead increases by $34.70. The overhead increases by $34.70 for each additional machine hour, or per machine hour.

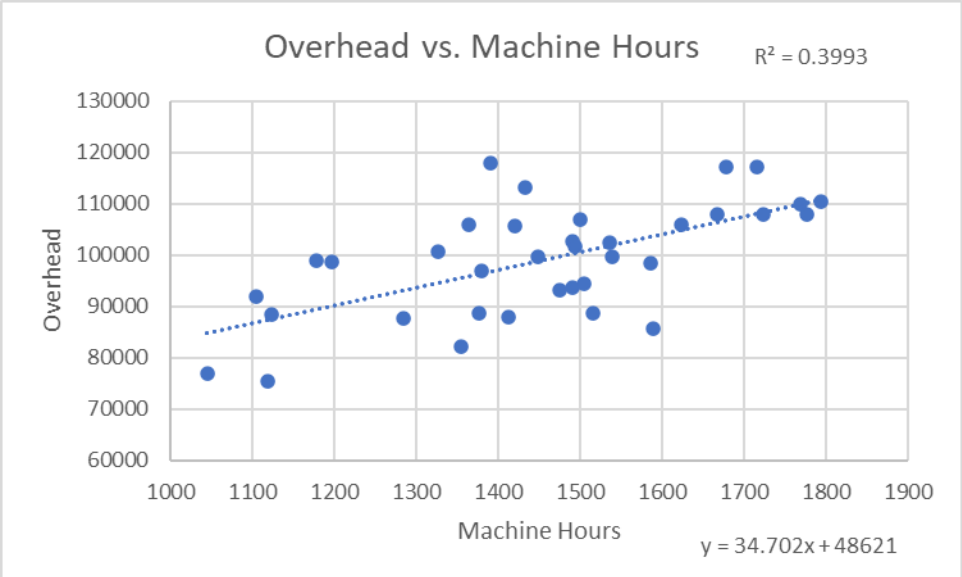
The y-value increases (or decreases) by the value of the slope for each one unit increase in x.

Using the regression equation to make prediction.

The prediction we get will be an average… there is some variation and technically we should have a prediction interval to describe the range of possibilities… but we won't worry about that right now.

Use the trendline equation/regression line/line of best fit…

Suppose I want to predict the overhead for 1700 machine hours. What is the average overhead likely to be? (the x-values we are predicting for should be inside (or very close to) the range of the original data). Predicting inside the range of the original data is called interpolating. Predicting outside the range of the original data is called extrapolating…. Extrapolating is very risky.



$$y = 34.702x + 48621$$
$$y = 34.702(1700) + 48621$$

$$y = \$107{,}614.40$$