Correlation

ingebrigt.kjaereng

September 2024

1 Correlation 4a, 4b and 4c

a.)

Correlation means the degree to which the data in one quantitative variable is associated with data in another quantitative variable. We assume the association is linear, that one variable increases or decreases a fixed amount for a unit increase/decrease in another. The degree of association is determined by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used. The correlation coefficient is measured on a scale that varies from +1 through 0 to -1. Complete correlation between two variables is expressed by either +1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. It is sometimes not clear what is dependent on what. Often each variable is dependent on some third variable, which may or may not be mentioned. The calculation of the correlation coefficient, with x representing the values of the independent variable and y representing the values of the dependent variable, $r = \frac{\sigma(x-x^-)(y-y^-)}{\sqrt{[\sigma(x-x^-)^2(y-y^-)^2]}}$, which is $= r = \frac{\sigma(xy)-n^-x^-y^-}{(n-1)SD(x)SD(y)}$. To determine whether association is merely apparent, or might have arisen due to chance using the $t-test$ in the following: $t = r * \sqrt{\frac{n-2}{1-r^2}}$, when we have $n-2$ degrees of freedom. Correlation describes the strength of the association between two variables, and is completely symmetrical, as the correlation between A and B is the same as the correlation between B and A. But if the two are related, when one changes by a certain amount the other changes on average by a certain amount. The relationship can be represented by the regression equation. In this context regression means that the average of y is a "function" of x, meaning it changes with x. The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together the line slopes upwards from left

to right, when one set decreases as the other increases the line slopes downward from left to right, and this regression equation is often more useful than the correlation coefficient. When creating two new variables, $W1 = \alpha_1 X$ and $W2 = \alpha_2 Y$, with scaling coefficients α_1, α_2, the correlation between these two is given by: $\rho * W1W2 = \frac{cov(\alpha_1 * X, \alpha_2 * Y)}{\sqrt{var(\alpha_1 * X) * var(\alpha_2 * Y)}}$. The covariance gives some information about how X and Y are statistically related. It measures how the values of X and Y move relative to each other. As in correlation, if large values of X tend to happen with large values of Y, then the covariance is positive and we say X and Y are positively correlated. If X tends to be small when Y is large, then the covariance is negative and we say X and Y are negatively correlated. If X and Y are independent, then $Cov(X, Y) = 0$. Also, $Cov(aX, Y) = aCov(X, Y)$, which is what happens when expanding covariance and variances in the formula: $cov(\alpha_1 * X, \alpha_2 * Y) = \alpha_1 * \alpha_2 * cov(X, Y)$, and $var(\alpha_1 * X) = \alpha_1^2 * var(X)$, $var(\alpha_2 * Y) = \alpha_2^2 * var(Y)$, which is $*W1 * W2 = \frac{\alpha_1 * \alpha_2 cov(X,Y)}{\sqrt{\alpha_1^2 * var(X) * \alpha_2^2 * var(Y)}}$. The denominator is simplified to $|\alpha_1| * |\alpha_2| * \sqrt{Var(X) * Var(Y)}$. The ratio $\frac{\alpha_1 * \alpha_2}{|\alpha_1| * |\alpha_2|}$ is the sign of the scaling coefficients $\alpha_1 * \alpha_2$, which can be written as $sgn(\alpha_1 * \alpha_2)$, and the equation becomes $\rho * W1 * W2 = sgn(\alpha_1 * \alpha_2)\rho * XY$. The correlation between the scaled variables $W1$ and $W2$ is exactly the same as the original correlation $*W1W2$, but multiplied by the sign of the product $\alpha_1 * \alpha_2$. If α_1 and α_2 have the same sign, the sign of the correlation remains the same, but if α_1 and α_2 have opposite signs, the correlation flips(changes sign). The difference here is that by introducing more factors, there are more factors with signs to consider in the formula for correlation, so changing one of or both of these might change the sign of the correlation $\rho * XY$. If both scaling coefficients α_1 and α_2 are positive or both negative, the sign of the product will be positive, whereas if one is positive and the other negative, the product of the scaling coefficients is negative, and the result, the correlation itself $\rho * XY$ flips sign. The magnitude of the correlation, meaning how much correlation there is between the two variables, remains unaffected by the scaling, and the absolute value of the correlation remains the same as the original correlation between X and Y. This happens because in the definition of correlation, $\rho * XY = \frac{cov(X,Y)}{\sqrt{var(X) * var(Y)}}$, the covariance is normalized, meaning it is adjusted so that it lies within a specific range. This removes the influence of the variables' units or scales, so correlation becomes independent of how large or small the variables are in absolute terms. The division of the product of the standard deviations $\sqrt{var(X)}, \sqrt{var(Y)}$, is what specifically makes it scale-independent, as when X and Y is scaled by constants, the covariance and variance change, but in a way that cancels out in the formula for correlation. *b.*)

In the formula $\rho * XYZ = \frac{\rho * XY - \rho * XZ \rho * ZY}{\sqrt{1 - \rho^2 * XZ}\sqrt{1 - \rho^2 * ZY}}$, it is assumed Z is a single variable, as the formula $\rho * XY * \phi$ is the $Z =$ 1st order partial correlation. The 0th-order partial correlation is the correlation coefficient $\rho * XY$, the same as in a. The 1st order partial correlation is the difference between a correlation and the product of the removable correlations divided by the product of the coefficients

of alienation(isolation) of the removable coefficients. $\rho * XY$ is the Pearson coefficient between X and Y. $\rho * XZ$ is the Pearson correlation coefficient between X and Z, and $\rho * YZ$ is the Pearson correlation coefficient between Y and Z, so $\rho * XY * Z$ gives the correlation between X and Y, controlling for the effect of Z, adjusting the original correlation $\rho * XY$ by accounting for the influence of Z. The product of the removable correlations in this case is $\rho * XZ$ * $\rho * YZ$. The difference is $\rho * XY - \rho * XZ * \rho * YZ$, which represents how much of the original correlation between X and Y remains after accounting for the linear relationships between X and Z, and Y and Z. This removes the indirect influence of Z on the correlation between X and Y. The coefficients of alienation are derived from the square of the correlations between variables. The coefficient of alienation is given by $\sqrt{1-\rho}^2$, which measures how much variance in a variable is not explained by another variable; it is isolating the influence of the variable to be controlled for (Z) w.r.t to X and Y. Dividing by $\sqrt{(1-\rho^2 XZ)(1-\rho^2 YZ)}$ normalizes the residual correlations between X and Z and between Y and Z so what remains is the unexplained variation in X and Y after controlling for/removing the influence of Z which we call the difference, giving a cleaner measure of the direct relationship between X and Y while controlling for Z.

In studying the relationship between exercise and weight loss, controlling for diet, the problem is that people's diet plays a crucial role in weight loss. Without controlling for Z(diet), the observed correlation between exercise and weight loss can be biased, as diet can influence both exercise and weight loss. By controlling for Z, I can calculate the partial correlation between exercise and weight loss. This isolates the direct effect of exercise on weight loss, excluding the impact of diet. When dividing by the product of coefficients of alienation(isolation) for the correlations between exercise and diet, $\rho * XZ = \sqrt{1-\rho^2 XZ}$ and between weight loss and diet, $\rho * YZ = \sqrt{1-\rho^2 YZ}$, leaving only the direct effect of exercise on weight loss.

In examining the relationship between rank and job performance, controlling for years of service, the problem is that years of service is a potentially confounding variable. As service members with more years of service tend to accumulate experience, they are more likely to be promoted and may also have better performance ratings simply due to experience, not necessarily rank. Without controlling for years of service, the relationship between rank and job performance may be distorted by the effect of experience. Controlling for Z(years of service), I can divide by $\sqrt{(1-\rho^2 XY)(1-\rho`2YZ)}$, where $\rho * XZ$ is the correlation between X(rank) and Z(years of service), and $\rho * YZ$ is the correlation between Y(job performance) and Z(years of service) are the products of alienation/isolation. This leaves only the direct influence of rank on job performance, that is, the relationship between rank and job performance.

Another example is exploring the link between religious involvement and community engagement, controlling for age. The problem is that older people might have more experience and opportunities for community involvement, leading to higher community engagement regardless of their level of religious

involvement. Using partial correlation to control for age means $\rho * XY$ is the correlation between religious involvement and age. $\rho * YZ$ would be the correlation between community engagement and age. Isolating the direct relationship between religious involvement and community engagement while controlling for age is done by division by the product of the coefficients of alienation: $\sqrt{(1-\rho^2 XZ)(1-\rho^2 YZ)}$.

c.) Does the following hold for partial correlation?:

If both scaling coefficients α_1 and α_2 are positive or both negative, the sign of the product will be positive, whereas if one is positive and the other negative, the product of the scaling coefficients is negative, and the result, the correlation itself $\rho * XY$ flips sign. The magnitude of the correlation, meaning how much correlation there is between the two variables, remains unaffected by the scaling, and the absolute value of the correlation remains the same as the original correlation between X and Y. This happens because in the definition of correlation, $\rho * XY = \frac{cov(X,Y)}{\sqrt{var(X)*var(Y)}}$, the covariance is normalized, meaning it is adjusted so that it lies within a specific range. This removes the influence of the variables' units or scales, so correlation becomes independent of how large or small the variables are in absolute terms. The division of the product of the standard deviations $\sqrt{var(X)}, \sqrt{var(Y)}$, is what specifically makes it scale-independent, as when X and Y is scaled by constants, the covariance and variance change, but in a way that cancels out in the formula for correlation.

In partial correlation, $\rho * XYZ = \frac{\rho*XY-\rho*XZ*\rho*ZY}{\sqrt{1-\rho^2 XZ}*\sqrt{1-\rho^2 ZY}}$, controlling for Z, meaning that we count out the influence of Z occurs in the denominator

Dividing by $\rho(1-\rho^2 * XZ)(1\rho^2 * YZ)$ normalizes the residual correlations between X and Z and between Y and Z so what remains is the unexplained variation in X and Y after controlling for/removing the influence of Z which we call the difference, giving a cleaner measure of the direct relationship between X and Y while controlling for Z.

The 1st order partial correlation is the difference between a correlation and the product of the removable correlations divided by the product of the coefficients of alienation(isolation) of the removable coefficient, and as it is the correlation between the residuals eX and eY resulting from the linear regression of X with Z and Y with Z, so it is a measure of the correlation between the error terms of X and Y from linear regressions involving Z. The n-th order partial correlation $\rho * XYZ^n$ means controlling for n variables, and measures the correlation between the error terms of X and Y from linear regressions with the n variables Z,..,n.

The property of the scaling coefficients determining only the sign, and not the magnitude of the correlation, applied to the partial correlation also holds, because in the regular correlation, the property that keeps the scaling coefficients from interfering with the magnitude is the fact that the covariance is normalized, removing the influence of the variables' inputs or scales, and when divided by the square root of product of the variances of X and Y, it becomes scale-independent as the variance and covariance changes in a way that cancels out the other. In partial correlation, the numerator involves correlation coefficients, which are

ratios of covariances to standard deviations, making them scale-independent for the same reason regular correlation is scale-independent: If X, Y or Z is scaled by constants, it increases the number of terms with different signs in the formula and a change in any of them can change the overall term of the expression. The correlations $\rho * ZX, \rho * ZY, \rho * XY$ are themselves scale-independent, so the scaling properties hold for partial correlations of any order, because the order of the partial correlation may change the size of Z but does not change the relationship between $\rho * XZ, \rho * ZY, \rho * XY$ themselves, only the signs.