Correlation: Beginning History through Development of Function

Billie Sue Arndt

Follow this and additional works at: https://mosaic.messiah.edu/honors

Part of the History Commons, and the Medicine and Health Sciences Commons

Recommended Citation

https://mosaic.messiah.edu/honors/148

This Thesis is brought to you for free and open access by Mosaic. It has been accepted for inclusion in Honors Projects and Presentations: Undergraduate by an authorized administrator of Mosaic. For more information, please contact mosaic@messiah.edu.
Correlation: Beginning History through Development of Function

What is correlation? This short question has the ability to either confuse or make one question their understanding of statistics. When posing this question to individuals, one can get a wide range of answers. “That is something in math or statistics, I think,” is one possible answer one may find. When talking to statistics students you may get their short answer, “Correlation is the relationship between two random variables.” These answers are very different, but at the same time, they are both correct. The degree to which one understands the concept of correlation can vary widely depending on one’s experience working with this concept. According to Merriam-Webster Dictionary, correlation is “The degree of association between two random variables ... The correlation between the graphs of two data sets is the degree to which they resemble each other” [1]. This is also a valid way to explain the concept of correlation in relation to statistics, but is it the best way? Correlation is a vast statistical topic varying in definition, as well as representation. This paper seeks to explore the concept of correlation in depth from its formulation in history up through the varying methods of correlation that are used in statistics today.

When first beginning to understand correlation and how it functions, one must first learn how correlation was developed. The discovery of correlation can be attributed to the 19th century scientist Sir Francis Galton, and as Morris Kline writes, it is the “most valuable feature of Galton’s work” [2]. He discovered the concept of correlation in the late fall of 1888, making it relatively recent, but it is one of the well documented discoveries in the last few centuries [4]. In Galton’s account of this discovery, he writes that the idea came to him while he was taking shelter from the rain in a rocky recess along a pathway at Naworth Castle [4]. Karl Pearson called this the “birthplace of the true conception of
correlations,” however later this account was discredited upon further evidence in the writings of Galton leaving the true story hidden in time [4]. Despite this confusion as to how the idea of correlation first came to Galton, he was still a major part of developing this topic to its fruition as a common statistical concept in the arsenal of statistics principles used today by statisticians on a regular basis.

Correlation was the concluding step in a 20 year research project where the major components were in place by the year 1886 [4]. Galton’s work originated with his interest in heredity, and to deal with problems that he encountered in his work he developed an analogue computer, the Quincunx [4]. This Quincunx was then used later on, along with a body of empirical work, to lead to the development of regression, which is needed in correlation as will become evident [4]. This collaboration described the results of Galton’s work by a bivariate normal distribution. It was also noted that there were two regression lines which formed a relationship between the constants of the normal surface [4]. This did not yet lead to a clear development of correlation. At this point Galton was still missing three realizations which were necessary to develop correlation: 1. two regression lines had the same slopes, 2. a common slope could be used as a single numerical summary of the strength of the relationship between two variables, and 3. the idea being applicable much more generally than just for heredity [4]. The first two of these realizations were already curiosities to Galton, but he needed the third, the generality of the problem, to fully realize what he was on the verge of discovering. This third idea dawned on Galton then in 1888.

It was later during an anthropological study which Galton was working on that he was lead to the realization that this current study was similar to his previous studies, which he determined were special cases of a general problem which he finally identified as correlation [4]. “To Galton, correlation meant what we might call today intraclass correlation – two variables are correlated because they share a common set of influences” [4]. Galton however only ever conceived correlation to be a positive relationship; negative correlation never playing a role in his discussion. Today though we understand
the correlation coefficient to range from -1 to 1 when studied in statistics, making this discussion from Galton one of interest to see how his belief has changed over time to allow for the negative correlations in work today.

Galton gave three examples to illustrate his concept of correlation: kinship, the trip time home for two clerks who took the same bus for part of their journeys, and stock portfolios for two investors who hold shares in the same ventures [4]. He used these examples to illustrate aspects about this new correlation he discovered. He first used the examples to underscore the fact that correlation did not depend upon the choice of origin, as well as he informs us that the concept will only apply to variables that have a “quasi-normal” distribution [4]. He also defined the coefficient of correlation, also called the index of correlation, using one of his examples. This work by Galton was later taken over and expanded up by Francis Edgeworth, Karl Pearson, and G. Udny Yule [4].

Karl Pearson was a colleague and researcher in Galton’s lab, and the efforts of both of these great men brought about the product-moment correlation coefficient [3]. As we have found, Galton was the original conceiver of the notion of correlation and regression, but later scientists took his work and expanded upon it. Pearson was one of these great men whose work has made an impact in statistics. Pearson used Galton’s discovery of correlation to develop a rigorous treatment of the mathematics for the Pearson Product-Moment Correlation (PPMC), which Stanton writes was a result of Galton’s first interest in heredity and genetics [3]. The thoughts that were needed by Pearson to create the PPCM began with the problems of heredity, mainly understanding how strongly the characteristics of one generation manifested themselves in the next generation [3]. The two-dimensional diagram of the data from the sweet pea plant example was a good illustration for the basic foundation of regression which was generalized into the PPCM [3]. The letter “r” was originally designated by Galton for regression, but Pearson used it to denote the correlation coefficient which is what we recognize today.
To have a further understanding of how scientists furthered the work started by Galton, let us take a deeper look into how Pearson developed correlation. In 1896, Pearson published his first rigorous treatment of correlation and regression which he credited Bravais with the initial mathematical formulae for correlation [3]. Pearson had stumbled upon the product-moment method for calculating the correlation coefficient; however he was not able to prove it provided the best fit for the data [3]. He was able to support his work by using an advanced statistical proof involving the Taylor expansion. Using the Taylor expansion, Pearson was able to demonstrate that optimum values of both the regression slope and the correlation coefficient could be calculated from the product moment (1) [3].

\[ \sum \frac{x'y}{n} \]  

Here \( x \) and \( y \) are the deviations of the observed values from the respective means used, and \( n \) is the number of pairs of values used by Pearson.

Since we have now examined some of the history detailing where correlation began, let us now examine the aspects of correlation in general. According to David Stockburger, the Pearson Product Moment Correlation Coefficient \( r \), or simply the correlation coefficient, equals a measure of the degree of the linear relationship present between two variables, usually represented as \( x \) and \( y \) [5]. When considering regression, the emphasis is usually on predicting variables from one another; in correlation the emphasis is on the degree to which linear models describe the relationship between two variables [5]. When working with correlation, one is also interested in this relationship as the critical aspect to understand. The correlation coefficient itself can take on any value ranging from -1.00 through +1.00. The sign (positive, negative) is what defines the relationship between the two variables. When the value shows a positive relationship it is showing that as the value of one variable increases, the value of the other variable is also increasing; vice versa when one variable is decreasing the other also decreases [5]. However, when there is a negative correlation, one of the variables in the relationship is increasing while the other decreases and vice versa. To measure the strength of the relationship one just takes the
absolute value of the correlation coefficient. This means that an $r = .50$ indicates a stronger relationship than if $r = .40$, and the same is so for an $r = -.50$ and $r = -.40$ due to the absolute value. When the correlation coefficient $r = 0.00$, this will indicate the absence of a linear relationship, whereas when $r = \pm 1.00$ a perfect relationship is indicated [5].

The correlation coefficient itself can be understood through various methods of interpretation. The first method to examine is the scatter plot. This method best illustrates how the correlation coefficient changes as the relationship between two variables is altered [5]. When $r$ is equal to zero, the points in the plot will be widely-shaped rough circle. As the relationship increases, this circle will become more elliptical in formation until the limiting case of $r = \pm 1.00$ is reached and the points then create a straight line. The second method where the correlation coefficient is noted is the regression line where the correlation coefficient is the slope [5]. The correlation coefficient makes up the slope, $b$, when $x$ and $y$ are converted to $z$-scores using formulas (2) and (3) respectively.

$$Z_x = \frac{x - \bar{x}}{s_x} \quad (2)$$

$$Z_y = \frac{y - \bar{y}}{s_y} \quad (3)$$

This $z$-score transformation subtracts the mean from the raw scores in the collected data and then divides by the standard deviation. From this transformation it is evident that (1) the correlation coefficient is invariant under a linear transformation of either $x$ and/or $y$, meaning that changing the scale of either variable will not change the size of the correlation coefficient as long as one is dealing with a linear relationship, and (2) the slope of the regression line is the correlation coefficient when converted to $z$-scores [5]. This can also be seen by computing the regression parameters (4).

$$Z'_y = rz_x \quad Z'_x = rz_y \quad (4)$$

Variance interpretation is another method of looking at the correlation coefficient. Here the squared correlation coefficient ($r^2$) is “the proportion of variance in $y$ that can be accounted for by knowing $x$ and vice versa” [5]. An important property of variance to keep in mind is that it can be
partitioned into separate additive parts. Therefore, the total variance is the sum of the predicted variance and the error variance which is unpredictable [5]. This relationship can be illustrated using equations (5) and (6) respectively.

\[ s_{\text{total}}^2 = s_{\text{pred}}^2 + s_{\text{error}}^2 \]  
\[ s_{\text{pred}}^2 = s_{\text{total}}^2 - s_{\text{error}}^2 \]  

The squared correlation coefficient is then equal to the ratio of the predicted total variance (7). The error variance, \( s_{\text{error}}^2 \), can also be estimated by the standard error of estimate squared, \( s_{\text{est}}^2 \), which then makes the total variance of \( y \), \( s_y^2 \), as shown in equation (8) [5].

\[ r^2 = \frac{s_{\text{pred}}^2}{s_{\text{total}}^2} \]  
\[ r^2 = 1 - \frac{s_{\text{est}}^2}{s_y^2} \]  
\[ S_{yx} = \sqrt{\frac{(N-1)}{(N-2)} s_y^2 (1 - r^2)} \]  

The standard error of estimation is then represented by equation (9), and from this estimate one is able to discover more about the correlation coefficient [5]. The relationship between the correlation coefficient, variance of \( y \), and the standard error of estimate is evident through equation (9). From this relationship one is also able to discern that the correlation coefficient becomes smaller as the standard error increases relative to the total variance, which shows that the correlation coefficient is a function of the standard error of estimate and the total variance of \( y \) [5].

The actual correlation coefficient can be calculated using the z-scores that were previously found when working with the regression line. The formula for finding the coefficient, \( r \), is as follows in equation 10 [5].

\[ r = \frac{\sum_{i=1}^{N} z_x z_y}{N-1} \]
Calculating the coefficient is not hard in practice, however, there are two things that one must also consider when working with the correlation coefficient. Outliers can appear in data and they can have an effect on $r$. When an outlier falls near where the regression line normally falls, it can increase the size of $r$, but when the outlier falls a distance away from the regression line it can decrease the value of $r$ [5]. This could potentially cause problems with calculations, so realizing how any outliers will affect the correlation coefficient is important. Another key to working with correlation is that correlation does not mean causation. It is a natural inclination that when one sees that two variables are correlated and show a strong relationship to one another, one’s mind could make the conclusion that one variable is causing the other. However, this is not necessarily true. It is possible for two variables to be related to one another but neither actually causes the other. In order for correlation to imply causation, two things must be satisfied before this conclusion can be considered: 1. the causal variable must temporally precede the variable it causes and 2. certain relationships between the causal variable and the second variable must be met [5].

As has hopefully become evident correlation is a very useful measure in statistics. Correlation in a general sense deals with the relationship between two variables, but it is not limited to only looking at this one type of correlation. There are actually many different methods of correlation that are used in the field of statistics. Robert Thorndike talks about many of these varying types of correlation in his book *Correlational Procedures for Research*. For all of the types of correlation that Thorndike covers, normality of data is assumed in generalizations and the product moment correlation, although it is not always certain how important it is that this condition is met [6]. The data is also linear when all points fall in an elliptical pattern scattered around a straight line, which gives a good description of the overall set of points [6]. Nonlinear data on the other hand will not form this ellipse resulting in no line. Thorndike discusses how in the differing types of correlation in his book a linear relationship is assumed to prevail among all the variables. The three types of correlation from Thorndike that will be discussed
Bivariate relationships are ones that are dealing with two variables and summarize relationships. This is a type of “index” in statistics that is used when reporting the results of tests [6]. An index such as this is needed because a scatter plot is often too inefficient for reporting results making a bivariate comparison a better option. This type of relationship involves the principle of least squares which is the basic concept that correlation is founded upon [6]. This principle is an outgrowth of the concept of variance in a distribution and it dictates that the mean is the point to use when a single value is used to describe an entire collection. According to Thorndike, “It is the point that gives the most accurate prediction on the average for the group because the average squared difference between actual and predicted scores will be smallest, and it is most descriptive of the group as a whole” [6]. It also minimizes errors of prediction and description. This is often seen used in the regression line of a bivariate correlation, regression weights in multiple and canonical correlation, factor analysis, and other correlation procedures [6].

The Product Moment Correlation Coefficient is an example of a bivariate relationship. It can be defined by the equation,

$$r_{xy} = \frac{\sum_{i=1}^{N} z_{x_i} z_{y_i}}{N}$$  (11)

which uses the mean of the cross products of the standard scores [6]. The form that most individuals are more familiar with is shown in equation (12).

$$r = \frac{\sum xy}{Nsx sy}$$  (12)

This replaces $Z_{xi}$ with $\frac{x}{sx}$ and $Z_{yi}$ with $\frac{y}{sy}$. The PMCC is then defined using the sum of the deviation score cross-products in the numerator and the two standard deviations in the denominator, but this definition
requires that one is working with equal variances. The computational form for the PMCC is then represented by equation (13), and all the work done with this equation is achieved just with raw scores.

\[ r = \frac{N \sum xy - \sum x \sum y}{\sqrt{N \sum x^2 - (\sum x)^2} \sqrt{N \sum y^2 - (\sum y)^2}} \quad (13) \]

When looking at the correlation coefficient for a bivariate relationship there are seven interpretations available for analysis. These interpretations to consider are the slope of the regression line, the proportion of variance, accuracy of prediction, proportion of common elements, expectation of success, normal correlation surface, and geometric vector interpretation. We will further examine three of these interpretations: the slope of the regression line, proportion of variance, and geometric vector interpretation. To begin, the first aspect of the correlation coefficient develops from the least-squares regression line where \( r \) is the slope of the best fit line for predicting \( y \) from \( x \) [6]. The slope \( B \) is given the value, \( B = \frac{\sum xy}{\sum x^2} \), when working with the regression line, and this value changes depending on which variable one considers as the predicting variable [6]. Standard scores are used here to make the sum of the squared deviations equal to the number of observations, \( N \), which makes the denominator value of \( B \) the same regardless of which variable is chosen as the predictor (14).

\[ \sum Z^2_x = \sum Z^2_y = N \quad (14) \]

This conversion to standard scores allows the correlation to be used for purposes of prediction, giving the correlation coefficient a precise mathematical relationship to the regression given by, \( B = r \frac{S_y}{S_x} \). It is important Thorndike points out, “The only time the two regression coefficients become equal is when the variance of the two variables are equal” [6]. This will occur when both \( B \)’s will equal the correlation coefficient because the ratio of the two variables will be unity, and this occurs because the variables are given in standard score form [6]. In this case then the correlation coefficient is also a regression coefficient.
The next interpretation of the correlation coefficient to consider is the proportion of variance. This is closely related to the correlation ratio $\eta^2$, which is the index of the degree to which knowing a score on $x$ will reduce the error when predicting a score on $y$ [6]. The regression coefficient is only the slope of the best prediction line as we have seen, and it does not give information about the proportion of variance in $y$ that is predicted from $x$, meaning that a second interpretation of $r$ is needed to make statements about the accuracy of this prediction [6]. Thorndike writes, “[The] standard error of estimate, $s_{y|x}$, is defined with $\eta^2$ as the standard deviation of the distribution of $y$ scores within a category of $x$ about the mean of those $y$ scores.” The regression line runs through $y$-means of an infinite number of continuous $x$ categories following a linear relationship, making the standard deviation of the deviations from the regression line the standard error of estimate [6]. In other words, this standard error of estimate is the standard deviation of the distribution of errors which are present in the prediction. If an error is present in the prediction and is defined by $Y - Y'$, the variance of the distribution is then represented by equation (15).

$$S_{y|x}^2 = \frac{\sum(Y-Y')^2}{N} \quad (15)$$

Working with this equation to find how the correlation coefficient is related, a few substitutions must first be made. When scores are expressed as deviation scores and substituting $r (\frac{s_y}{s_x}) x$ for $y'$ gives equation (16) which through the use of algebra gives us equation (17). If one uses principles from the correlation ratio, one is able to find that equation (18) results. This shows that the square of the correlation coefficient can be viewed as the proportion of variances in $y$ which are predictable from $x$ [6].

$$S_{y|x}^2 = \frac{\sum(Y-r(\frac{s_y}{s_x})x)}{N} \quad (16)$$

$$r_{xy}^2 = 1 - \frac{s_{xy}^2}{s_y^2} \quad (17)$$
\[ r_{xy}^2 = \frac{s^2_y}{s^2_x} \]  

(18)

The third type of correlation interpretation that will be analyzed is a geometric interpretation of the correlation coefficient, geometric vector interpretation. This interpretation is not well known and is rarely seen in statistics books according to Thorndike, but it forms a backbone for approaching multivariate statistics [6]. Normally, a person could be represented graphically as a point at the intersection of their scores for two variables when the variables are represented as the orthogonal axes of the space, which are orthogonal to the ordinary scatter plot [6]. This representation of the variables has allowed the idea of correlation to be developed along with its applications. Working with standard scores allows us to simplify our discussions in this process. However, with the geometric interpretation we are now working with a “reverse scatter plot.” This means that we are letting the people be defined as the axes of the space and the variables are represented by the points, the scores the people obtained on the variables [6]. This becomes a problem when trying to visualize the plot of the data. When there were only two variables as the axes, it was simple to show graphically a scatter plot of data for any number of people. In statistics though, one must have a fairly large number of people to work with, and this presents a problem with the reverse scatter plot and needing to visualize in higher dimensions [6]. One is still able to visualize three people represented as axes and simultaneously any number of variables, but when the number of people changes to four or more, visualizing this graph becomes impossible for human comprehension.

Terminology is often very important when working with statistics, and geometric vector interpretation is no different. Here the key term to understand is “vector,” and there are two meanings which are important to know. The first meaning when referring to a vector is that it can be a linear array of numbers, usually the set of standard scores obtained by a group, and these scores are then represented by a row or column vector. The second meaning is that a vector is a line segment in space which has both direction and length, and the vector can be represented by an arrow with an origin and
termination point. A vector can express the scores for a group of people on a variable, and these scores can be raw scores, standard scores, or deviation scores [6]. A variable vector is a vector representing the variable in a space defined by the people and it has the same meaning as the point it runs to [6]. Graphically working with vectors in this interpretation yields two results which are important for connecting to the correlation coefficient.

The first result that appears from using the variables as vectors is that if one takes the sum of the cross products of equivalent scores and divides by, $NS_xS_y$, one gets the correlation between the two variables [6]. When using deviation score vectors for $X$ and $Y$, this process multiplies $x_1 * y_1$, $x_2 * y_2$, and so on, summing them for all $N$ scores which yields the correlation coefficient expressed by $\frac{\sum xy}{NS_xS_y}$ [6]. However, there is a more important development that is found from treating variables as vectors. This was that the resulting expression, $\frac{\sum xy}{NS_xS_y}$, also defines “the cosine of the angle between the two vectors” [6]. This means that the cosine of the angle between the two vectors in the space is equal to the mean of the products of the coordinates which define the vectors, as well as it equals the correlation between the two variables. This gives us a way to graphically represent the correlation between two or more variables [6]. When one considers these vectors then and what the cosine angle means, it is evident that when two variables have a high positive correlation, this is shown geometrically by two vector lines close together with a small angle between them [6]. As the angle approaches zero, the cosine will approach 1.00, and this translates into a correlation coefficient of 1.00, which has been referred to as a perfect correlation. This can only occur when one is considering standard scores and the relationship between the two variables meets the condition $Z_{xi} = Z_{yi}$ [6]. The same occurs for a negative relationship when $Z_x = -Z_y$, and the points defined by these coordinates are opposite in direction yielding a cosine of -1.00. Generally then in a graphical representation an acute angle will show a
positive correlation, an obtuse angle will represent a negative correlation, and zero correlation will be shown as a right angle [6].

The second type of correlation presented by Thorndike is Part and Partial Correlation. These are two separate types of correlation, but they are similar. These methods are used when prediction may be a goal and more than two variables are available. We will first examine Part Correlation. This type of correlation is also known as semi-partial correlation and there are two approaches to take to understand what part correlation achieves [6]. The first approach can be discussed when working with a situation involving three variables: A, B, and C. Consider the relationship between A and B with the effect of C on A removed. To find the part correlations for an example such as this, one can find them directly from the original correlations [6]. Equation (19) gives the formula for finding the part correlation between A and B with C removed from A where the term $r_{B(A+C)}$ equals this correlation and the $r$ terms are the product moment correlations between A & B, A & C, and B & C [6].

$$r_{B(A+C)} = \frac{r_{BA} - r_{BC}r_{AC}}{\sqrt{1 - r_{AC}^2}}$$  \hspace{1cm} (19)

The results from this formula can also be found if the A-score deviations from the regression line for predicting A from C were correlated with the B scores, but that is a more tedious method [6]. The same process can be used to find the part correlation between A and B where C is held constant for B, altering the formula to then be represented by equation (20).

$$r_{A(B+C)} = \frac{r_{AC} - r_{BC}r_{BA}}{\sqrt{1 - r_{BC}^2}}$$  \hspace{1cm} (20)

The second approach to part correlation developed out of the geometric interpretation of correlation. This uses the variable vectors in the people space. In this approach, the vector’s length is its standard score so each vector is of unit length and the angles between the vectors reflects the correlation between them [6]. When the vectors are all positively correlated, acute angles are formed. To visualize the relationship between vectors A, B, and C, there is one plane in the space that passes
through the origin of the vectors and is perpendicular to the C vector [6]. Since vector C is orthogonal with the plane, we are able to say the plane is independent of the vector. When vector A is then projected onto the plane, the projection represents the part of A which is independent from variable C, making the projected new vector not of unit variance since the shared variance with C has been removed [6]. These scores are no longer standard scores. This creates the need for a correction component for the correlation because the variables no longer all have unit variance which is necessary for the geometric interpretation. This correction is actually already located in the denominator of equation (20), and it allows the part correlation to still be viewed as the cosine of the angle between the two vectors [6].

In order to discuss Partial Correlation, we will continue with the example using the three variables A, B, and C. Now let us consider when B is a precursor to A and C. Partial correlation is then defined as when one looks to see if what the correlation is between A and C with the effect of variation in B removed from both variables [6]. The logic here is similar to that of part correlation. To find the partial correlation, one basically creates groups according to scores on B, finding the mean scores for A and C for the members of each group, and then assigning each person as their deviation scores for A and C [6]. The partial correlation between A and C is then the correlation between the pairs of deviation scores for both variables, and these come from the regression line for predicting each variable from B. The equation for partial correlation (21) is the correlation between these 2 sets of deviations and AC*B indicates the effect of B has been removed.

\[
\tau_{AC+B} = \frac{\tau_{AC} - \tau_{AB} \tau_{CB}}{\sqrt{1 - \tau_{AB}^2} \sqrt{1 - \tau_{EB}^2}} 
\]  

(21)

Partial correlation is also similar to part correlation geometrically. The difference is that in partial correlation both variable vectors are projected onto the orthogonal plane of the vector representing the removed variable [6]. Here again a correction must be taken which is already shown in the equation for partial correlation.
The final type of correlation that we will endeavor to discuss in the realm of this paper is Multivariate Correlation. These methods are for examining the relationship between one variable and a combination of two or more variables that are considered simultaneously [6]. Multivariate correlation is often also referred to as multiple regression. The three-variable case can be used to illustrate this type of correlation. Continuing with our example using variables $A$, $B$, and $C$, we now wish to obtain the best prediction of $A$ given $B$ and $C$. This type of example is one that needs multiple regression to obtain its solution [6]. What one must determine in this example is the best combination of $B$ and $C$ for predicting $A$, and to see how good that prediction is. These problems can start to be explained using the idea of combining variables.

This problem we are working with requires the combining of two variables so that the combination of both has the maximum relationship with the variables being predicted [6]. The two variables are called the predictors and the variable we are trying to predict is the criterion. One restriction in this problem is that the combination of variables be linear so this will result in no powered terms in the solution. First begin by considering our variables in a space where $B$ and $C$ are lying in the orthogonal plane. $B$ and $C$ are still correlated, but they are used indirectly to represent any line lying in the plane: meaning any line in the plane can be expressed as a linear function of $B$ and $C$ and any line not in the plane cannot be defined in this manner [6]. The combination of $B$ and $C$ needed to predict $A$ with the best accuracy, or least error, must lie in this plane which both contains and defines $B$ and $C$. Because we have seen that the angle between vectors is related to their correlation, the vector in the plane $BC$ which has the smallest angle, i.e. largest correlation, with the $A$ variable vector is the one formed by the perpendicular projection of $A$ onto the plane [6]. This projection is then the best combination of $B$ and $C$ which one is looking for as the best predictor of $A$.

The next step in working with the multiple regression is how to express the projection of $A$, or $A'$, as a combination of $B$ and $C$. When $B$ and $C$ are uncorrelated this is not a problem since for this case
the solution is then that each predictor variable receives a weight equal to its correlation with the
criterion [6]. However, when \( B \) and \( C \) are correlated, only the portion of each predictor variable’s
variance which is independent of the other predictor is able to be used since in the Cartesian plane
there must be independent reference axes [6]. Vector \( A’ \) is referred to as a predictor composite since it
is a combination of several predictor variables [6]. To define what this statement means it is necessary
to define a few terms. Composite means a latent vector that is a combination of observed variable
vectors, and a variable vector is one that is observed directly in the sense that we have a measuring
instrument to get scores directly [6]. Latent vectors/composite then are ones that do not have a
measuring instrument associated with them, but are measured indirectly as a combination of scores
from measured variables and are then expressed as linear function of observed variables [6]. This
means that \( A’ \) can be expressed as a linear function of \( B \) and \( C \) in the form of equation (22) such that if
one took a point on the \( B \) vector and multiplied it by the \( b_B \), then the same for \( C \) and added the two
values together, the resulting value would lie on \( A’ \) [6].

\[
A' = b_B B + b_C C \quad (22)
\]

This process computes \( b_B \) and \( b_C \) in a way to reflect the portion of each variable vector that is
independent of the other [6]. These terms are then the weights which give the coordinates of \( A’ \) with
respect to the observed axes, being the observed variables. It is important to note that these
coordinates are expressed as the projection of \( A’ \) onto the \( B \) axis parallel to \( C \) and the \( C \) axis parallel to \( B \),
unlike previously observed perpendicular projections [6]. Previously we used independent axes in the
space and the projections perpendicular to one axis were parallel to the axis being omitted; now we are
just using the parallel projections to provide the independent coordinates [6].

The cosine as the angle between two variable vectors adjusted for length is still the correlation.
Working with our example, the cosine of the angle \( A’OC \), where \( O \) is the origin, is given by the length of
the side adjacent to the angle divided by the length of the hypotenuse (\( OA’ \)) [6]. A perpendicular line to
OC creates a right triangle. If C is expressed in standard scores, the projection A’ onto C is the correlation between the latent variable A’ and observed variable C, and this will be whatever it is no matter where B is [6]. However, the coordinates expressing the location of A’ does depend on the correlation between B and C where A’C correlation is unchanged but the correlation between C and B, B’, is negative [6]. Since B and C are standard scores for plotting variables in the space, A’ can be expressed better in the form (23).

\[ Z_{A'} = \beta_B Z_B + \beta_C Z_C \] (23)

Using simple algebra then, equation (24) is evident, and these forms resemble the formulas for part correlations with the criterion.

\[ \beta_B = \frac{r_{AB} - r_{AC}r_{BC}}{1 - r_{BC}^2} \quad \text{and} \quad \beta_C = \frac{r_{AC} - r_{AB}r_{BC}}{1 - r_{BC}^2} \] (24)

This similarity makes sense because if \( r_{AB} \) and \( r_{AC} \) are the appropriate weights for B and C when they are uncorrelated, then something close to their part correlation would be logical equivalents when they are correlated [6]. The denominator here again adjusts for the length of the composite and by subtracting \( r_{AC}r_{BC} (r_{AB}r_{BC}) \) we adjust for the effect of the correlation between B and C [6]. Therefore when one knows the correlations among the three variables it is easy to find the weights to give the best composite for prediction.

The last aspect of multivariate correlations that we will examine is the multiple correlation coefficient. When discussing multiple regression from the point of view of using variable vectors, the problem presented was to find a composite of the predictor variables that had the smallest angle with the criterion vector. Based on our understanding of bivariate correlation which was discussed earlier, the angle should still be related to a correlation. The general criterion vector \( Y' \) is the variable vector of composite scores [6]. If you use equation (23) to compute a standard composite score \( Z_{xy} \), for each person in the sample and then plotted that vector in the space, it would be the composite vector and
each score would be on it [6]. This gives the correlation between the criterion variable and the predictor composite when is given by equation (25).

\[ r_{YY'} = \frac{\sum_{i=1}^{N} Z_{Yi}Z_{Y'i}}{N} \]  \hspace{1cm} (25)

This simple product moment correlation between the criterion and the composite is called the multiple correlation coefficient. The formula given by equation (26) is used to compute the multiple correlation directly without computing the composite score for each individual [6].

\[ R_{Yx_1x_2} = \sqrt{\beta_{x_1}r_{Yx_1} + \beta_{x_2}r_{Yx_2}} \] \hspace{1cm} (26)

Since this equation is defined as a square root, it can never be negative which is not a problem since the \( \beta \)'s take care of the negative bivariate relationships [6]. This means that even if the predictors have a negative correlation with the criterion, a composite can still be found to have a positive correlation with the criterion. Then each product term that enters the multiple correlation will be positive. This covers the basic principles involved with multivariate correlation and the multiple correlation coefficient.

Throughout this paper I have endeavored to explore the concept of correlation from its development by Sir Francis Galton through some of the various methods of correlation that are used in the field of statistics today. By no means is the vast topic of correlation covered through this paper, to do that would require many more chapters. Correlation is much more than just the relationship between two variables, and there is certainly more to it than just being a topic in statistics. Correlation takes on many different forms to adapt to many situations as evident from this paper. Hopefully the reader of this paper will now have a better understanding of the concept of correlation and its vastness. Sir Francis Galton did not know how far this concept of correlation he discovered would go, for him it was simple observations and realizations that came about from his work with heredity. Today however the field of statistics is enriched by this concept Galton discovered and as statisticians today we thank him for giving us correlation.
Bibliography


