Correlation and Regression

Correlation
Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, consider the variables family income and family expenditure. It is well known that income and expenditure increase or decrease together. Thus they are related in the sense that change in any one variable is accompanied by change in the other variable. If the change in one variable is accompanied by a change in the other, then the variables are said to be correlated. We can therefore say that family income and family expenditure, price and demand are correlated.

Correlation Types

Positive Correlation
A correlation in the same direction is called a positive correlation. If one variable increases the other also increases and when one variable decreases the other also decreases. For example, the length of an iron bar will increase as the temperature increases.

Negative Correlation
Correlation in the opposite direction is called a negative correlation. Here if one variable increases the other decreases and vice versa. For example, the volume of gas will decrease as the pressure increases, or the demand for a particular commodity increases as the price of such commodity decreases.

No Correlation or Zero Correlation
If there is no relationship between the two variables such that the value of one variable changes and the other variable remains constant, it is called no or zero correlation.

Scatter Diagram Method
The simplest device for ascertaining whether two variables are related is to prepare a dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper in the form of dots, i.e., for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the charts, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship. If all the points lie on a straight line falling
from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive (i.e., \( r = +1 \)) (diagram I). On the other hand, if all the points are lying on a straight line rising from the upper left-hand corner to the lower right-hand corner, correlation is said to be perfectly negative (i.e., \( r = -1 \)) (diagram II).

If the plotted points fall in a narrow band there would be a high degree of correlation between the variables – correlation shall be positive if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram III) and negative if the points show a declining tendency from the upper left-hand corner to the lower right-hand corner (diagram IV).

On the other hand, if the points are widely scattered over the diagram it is the indication of every little relationship between the variables --- correlation shall be positive (Low Degree of Positive Correlation) if the points are rising from the lower left-hand corner to the upper right-hand corner (diagram V) and negative (Low Degree of Negative Correlation) if the points are declining from the upper left-hand side to the lower right-hand side of the diagram (diagram VI).
If the plotted points lie on a straight line parallel to the x-axis, it shows absence of any relationship between the variables (i.e., \( r = 0 \)) as shown by diagram VII.

**Regression**

Simple regression is used to examine the relationship between one dependent and one independent variable. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known. Regression goes beyond correlation by adding prediction capabilities.

For example, a medical researcher might want to use body weight (independent variable) to predict the most appropriate dose for a new drug (dependent variable). The purpose of running the regression is to find a formula that fits the relationship between the two variables. Then you can use that formula to predict values for the dependent variable when only the independent variable is known. A doctor could prescribe the proper dose based on a person's body weight.

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as \( Y \). The variable we are basing our predictions on is called the predictor variable and is referred to as \( X \). When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of \( Y \) when plotted as a function of \( X \) form a straight line.
The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between $X$ and $Y$. If you were going to predict $Y$ from $X$, the higher the value of $X$, the higher your prediction of $Y$.

![Table 1. Example data.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>3.00</td>
<td>1.30</td>
</tr>
<tr>
<td>4.00</td>
<td>3.75</td>
</tr>
<tr>
<td>5.00</td>
<td>2.25</td>
</tr>
</tbody>
</table>

*Figure 1. A scatter plot of the example data.*

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure 2 is the regression line and consists of the predicted score on $Y$ for each possible value of $X$. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.
The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values ($Y'$) and the errors of prediction ($Y - Y'$). For example, the first point has a $Y$ of 1.00 and a predicted $Y$ (called $Y'$) of 1.21. Therefore, its error of prediction is -0.21.

**Table 2. Example data.**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Y'</th>
<th>Y - Y'</th>
<th>$(Y - Y')^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>1.210</td>
<td>-0.210</td>
<td>0.044</td>
</tr>
<tr>
<td>2.00</td>
<td>2.00</td>
<td>1.635</td>
<td>0.365</td>
<td>0.133</td>
</tr>
<tr>
<td>3.00</td>
<td>1.30</td>
<td>2.060</td>
<td>-0.760</td>
<td>0.578</td>
</tr>
<tr>
<td>4.00</td>
<td>3.75</td>
<td>2.485</td>
<td>1.265</td>
<td>1.600</td>
</tr>
<tr>
<td>5.00</td>
<td>2.25</td>
<td>2.910</td>
<td>-0.660</td>
<td>0.436</td>
</tr>
</tbody>
</table>

Ref: (Link)

**Bivariate Data**

Statistical data are often classified according to the number of variables being studied.
Univariate data
When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.

Bivariate data
When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data.

Some examples:
- Height (X) and weight (Y) are measured for each individual in a sample.
- Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample.
- A cell culture is treated with varying concentrations of a drug, and the growth rate (X) and drug concentration (Y) are recorded for each trial.
- Temperature (X) and precipitation (Y) are measured on a given day at a set of weather stations.

Karl Pearson Correlation Coefficient

Correlation Coefficient, \( r \):

The quantity \( r \), called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honour of its developer Karl Pearson.

The mathematical formula for computing \( r \) is:

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

where \( n \) is the number of pairs of data.

The value of \( r \) is such that \(-1 \leq r \leq +1\). The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

Positive correlation: If \( x \) and \( y \) have a strong positive linear correlation, \( r \) is close to +1. An \( r \) value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between \( x \) and \( y \) variables, such that as values for \( x \) increases, values for \( y \) also increase.

Negative correlation: If \( x \) and \( y \) have a strong negative linear correlation, \( r \) is close to -1. An \( r \) value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between \( x \) and \( y \) such that as values for \( x \) increase, values for \( y \) decrease.

No correlation: If there is no linear correlation or a weak linear correlation, \( r \) is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables. Note that \( r \) is a dimensionless quantity; that is, it does not depend on the units employed.
A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If \( r = +1 \), the slope of this line is positive. If \( r = -1 \), the slope of this line is negative.

A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Ref: (Link)

**Spearman’s Rank Coefficient Correlation**

In mathematics and statistics, Spearman's rank correlation coefficient is a measure of correlation, named after its maker, Charles Spearman. It is written in short as the Greek letter rho (\( \rho \)) or sometimes as \( r_s \). It is a number that shows how closely two sets of data are linked. It only can be used for data which can be put in order, such as highest to lowest. Variables that are either ordinal, interval or ratio.

The general formula for \( r_s \):

\[
\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}
\]

For example, if you have data for how expensive different computers are, and data for how fast the computers are, you could see if they are linked, and how closely they are linked, using \( r_s \).

**Step one**

To work out \( r_s \) you first have to rank each piece of data. We are going to use the example from the intro of computers and their speed.

So, the computer with the lowest price would be rank 1. The one higher than that would have 2. Then, it goes up until it is all ranked. You have to do this to both sets of data.

<table>
<thead>
<tr>
<th>PC</th>
<th>Price ($)</th>
<th>Rank₁</th>
<th>Speed (GHz)</th>
<th>Rank₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>200</td>
<td>1</td>
<td>1.80</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>275</td>
<td>2</td>
<td>1.60</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>300</td>
<td>3</td>
<td>2.20</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>350</td>
<td>4</td>
<td>2.10</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>600</td>
<td>5</td>
<td>4.00</td>
<td>5</td>
</tr>
</tbody>
</table>

**Step two**

Next, we have to find the difference between the two ranks. Then, you multiply the difference by itself, which is called squaring. The difference is called \( d \), and the number you get when you square \( d \) is called \( d^2 \).
Step three
Count how much data we have. This data has ranks 1 to 5, so we have 5 pieces of data. This number is called $n$.

Step four

Finally, use everything we have worked out so far in this formula: $r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$.

\[
\sum d^2 \text{ means that we take the total of all the numbers that were in the column } d^2. \text{ This is because } \sum \text{ means total.}
\]

So, $\sum d^2$ is $1 + 1 + 1 + 1 + 1$ which is 4. The formula says multiply it by 6, which is 24.

$n(n^2 - 1)$ is $5 \times (25 - 1)$ which is 120.

So, to find out $r_s$, we simply do $1 - \frac{24}{120} = 0.8$.

Therefore, Spearman's rank correlation coefficient is 0.8 for this set of data.

What the numbers mean

$r_s$ always gives an answer between $-1$ and 1. The numbers between are like a scale, where $-1$ is a very strong link, 0 is no link, and 1 is also a very strong link. The difference between 1 and $-1$ is that 1 is a positive correlation, and $-1$ is a negative correlation. A graph of data with a $r_s$ value of $-1$ would look like the graph shown except the line and points would be going from top left to bottom right.
For example, for the data that we did above, $r_s$ was 0.8. So this means that there is a positive correlation. Because it is close to 1, it means that the link is strong between the two sets of data. So, we can say that those two sets of data are linked, and go up together. If it was $-0.8$, we could say it was linked and as one goes up, the other goes down.

Ref: ([Link](http://example.com))