# A9Wn Interpretation of the correlation coefficient (r), coefficient of determination (r-squared) and standard error of the estimate (SEE)

To ensure the clarity of explanations, we are using short variables of only 30 observations that were artificially created. As appropriate, we will also switch to some real-life variables from the UK Office for National Statistics (ONS).

We start with the coefficient of correlation r. We do not show calculations here, just focus on the interpretation of the three coefficients.

Imagine two variables with identical observation values. You can say that they are moving exactly in the same direction and showing identical variability of movements. The graph depicting these two variables is in Figure 1.



Figure 1. Two identical variables shown in a scatter diagram

When the value of variable x=1 in Figure 1, then the value of y=1. When x=25, y=25, etc. They form a straight diagonal line and their correlation coefficient r=1. This is an example of the perfect positive correlation. It is positive because in front of 1 we have the implicit positive sign, and it is perfect because the value is 1, which is the maximum the correlation coefficient can achieve.

Now imagine the opposite, i.e. the two variables moving in the opposite direction. In other words, when x=30, imagine that y=1, when x=25, y=6, etc., until x=1 and y=30. These two variables shown in a scatter diagram look as in Figure 2.



Figure 2. Two variables moving in the opposite direction from one another

Figure 2 shows an example of perfect negative correlation and r=-1. When one variable goes up, the other one goes down proportionally.

Both example above were artificially created variables with the values of observations in the range of 1 to 30. Let’s now create variables x and y, both as random numbers between 1 to 30. They form a relationship as given in Figure 3.



Figure 3. Scatter diagram of the two random variables

Theoretically, they should show no correlation and the value of r should be zero. However, as we used a short series of only 30 observations (and Excel random algorithm might not be perfect!), the value of r we got for this example is r=-0.12. In other words, a weak case of negative correlation.

Let’s take an example where x shows only three very close values, for example 4.9, 5.0 and 5.1. The other variable y will take values increasing in simple steps from 1 to 30. The graph looks as Figure 4.



Figure 4. Scatter diagram of two variable, one with only three fixed values, and the second one continuous discrete

The value of the correlation coefficient for Figure 4 is r= -0.00273, which is virtually zero correlation. Note that this indicates that changes in x has virtually no impact on changes in y. In other words, y varies independently of x, hence r=0.

Now let’s look at two real life variables, which are the number of couples divorcing in England and Wales between 1988-2017 and the prices of diesel in the same period. The graph is as in Figure 5.



Figure 5. Scatter diagram of the number of couples divorcing in England and Wales and the prices of diesel between 1988-2017

The correlation coefficient for Figure 5 is very high, r=-0.89. This is a high negative correlation indicating that when the number of divorces drops, the price of diesel goes up. This is clearly non-sensical and confirms our assertion from the textbook that correlation and causation are two separate issue. Correlation is a prerequisite for causation, but it is not a sufficient element to draw the conclusions about the causation.

In summary, the correlation coefficient will describe the direction (plus or minus sign) and it will indicate the strength of the relationship between the two variables, where 1 is very strong and 0 is no relationship at all.

We now focus on the coefficient of determination, or as it is often called the r-squared statistic. This statistic (r2) is the squared value of the correlation coefficient (r).

Let’s look at the two variables from Figure 1 and calculate the equation that fits the movements of y, if expressed as being dependent on x. The equation is y=x, because they both grow from 1 to 30 in exactly the same steps. The coefficient of determination, or r-squared is r2=1. This is shown in Figure 6.



Figure 6. Regression line between two identical variables and their r-squared value

What r-squared tells us is that our model can perfectly predict the changes in y as x changes (read r2=1 as 100%).

Our non-sensical example about the number of couples divorcing in England and Wales between 1988-2017 and the prices of diesel in the same period can also be modelled and the value of r-squared calculated. This is given in Figure 7.



Figure 7. Regression line for the number of couples divorcing in England and Wales between 1988-2017 and the prices of diesel in the same period and their r-squared value

We can see that the equation for straight line in Figure 7 is y=-0.0015+293.58 and that this equation, or model, provides r-squared of 0.79. IF you subtract 0.79 from 1, this give you the value of 21. You can use the following phrase: our model states that 79% of the variability in the movements of prices of diesel can be explained by the variability in number of divorces. The remaining 21% is not explained by this model and is a result of some other factors not included in this model. The explanation is technically correct, but it makes no sense as these two variables are not related. We used it here just to demonstrate how to use the correct wording.

Let’s now artificially minimise the variability in the graph above, just to show what impact it has on the r-squared value. Figure 8 shows this case.



Figure 8. Regression line for the same two variables as in Figure 7., but artificially reduced y values to make a closer fit

As we can see the equation is still the same as in Figure 7, but the actual observations are much closer to the line of regression. The value of r-squared is 0.9381. It has gone up. This means that the smaller the variations of the actual y variable around the line of regression, the higher the r-squared will be.

We can tighten the variations around the regression line even more (do not forget we are only artificially doing this to demonstrate the point). Figure 9 shows this case.



Figure 9 Same as Figure 8, but with even artificially moved y values even closer to the regression line

The equation remains the same, but the r-squared has gone up even more to 0.9974.

Because r-squared is the squared value of the correlation coefficient, we can calculate the correlation coefficients for Figures 7-9 as the square root values of the coefficient of determination. They are 0.7913, 0.9381 and 0.9973 respectively.

The lesson here is that both the correlation coefficient r and the coefficient of determination r-squared respond to how much actual observation deviate from the straight line (linear regression line). In other words, the closer they are to this line, the higher the value of both r and r-squared. The more deviations from the line they show, the lower the value of both coefficients.

What we must remember is that what we said here is only applicable for linear relationships. In other words, a low value of r and r-squared might imply that there is some other type of relationship between the variables (any form of non-liner or any form of periodic relationship, for example).

In Figure 10 we show the same data set fitted by two different curves. The left one was fitted by a linear equation and it shows r-squared of 0.799. The right one was fitted with a polynomial equation (non-liner) and it shows r-squared of 0.913.

 

Figure 10. One variable fitted by the linear model (left) and the same variable fitted by a non-linear polynomial model (right)

If you analysed the differences between the actual and fitted values for the left model vs the actual and fitted values for the right model, you would discover that for the left model the total sum of all these differences is 2.3 time larger that for the model on the right. This means that the model on the right fits the actual data much better than the model on the left.

This also confirms our earlier statement that the coefficient of determination, or r-squared, will depend on how closely the actual values of y are to their approximation (fitted or predicted values) $\hat{y}$, as produced by the model. The closer they are, the higher the value of r-squared.

R-squared is one of the measures of goodness of fit for a model. In other words, how well does the model fit the actual data. The better the fit, the higher the r-squared value.

In the textbook we used Figure 8.38 to show deviations of every point y from the mean value $\overbar{y}$ and from the model predicted value $\hat{y}$. Figure 11 below is the reproduction of Figure 8.38. It shows just one point, but the regression principles apply to the sum all the points in the model.



Figure 11. Depiction of how to visualise differences between every point y from their estimated value $\hat{y}$ and their mean value $\overbar{y}$. The squared sum of all these differences is called SSR, SSE and SST

The squared sum of all these differences creates three different types of variations. We call SSR the explained variations, SSE the unexplained variations and SST the total variations. From there, we declared in the textbook that R-squared is effectively the ratio of explained and total variations:

R-squared = Explained variation / Total variation, or r2 = SSR / SST

We can conclude that the more the explained variations (SSR) participate in total variations (SST), the higher the r-squared value.

To better understand this, let’s imagine an extreme case where all the values that a regression model produces are identical to the mean value. In this case SSR=0 and SSE=SST. This case states that the model cannot account for (or, predict) any variation in y and the r-squared value will be 0. Effectively this implies that x does not tell us anything about y, as clearly the predicted values $\hat{y}$ do not covary with y. According to this model, every value of $\hat{y}=\overbar{y}$.

If, on the other hand, we had perfect predictions where $\hat{y}$=y for all observations, then SSE=0 and SSR=SST, which will return r-squared=1. In this case, $\hat{y}$ responds to changes in x just as does y. The implication of this statement is that the smaller the variations of y, when measured from the line $\hat{y}$ (or closer $\hat{y}$ to y), the higher the r-squared value.

Classical definition of the r-squared statistic is that it is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The higher this value, the better are the variations of the dependent variable predicted by the independent variable. This also implies that r-squared is a statistic that measures the goodness of fit for the model. The higher the value of the r-squared, the more precisely our model fits the actual data.

We used the word “precisely” in the previous paragraph, but what is the meaning of this word? It needs to be understood in the context of accuracy. An accurate model is the one whose fitted values $\hat{y}$ are identical to y. A precise model is the one whose fitted values $\hat{y}$ are close to y. The closer $\hat{y}$ to y, the more precise the model is. In a way, this implies that r-squared not only provides the measure of the goodness of fit, but it also helps us determine the precision level for the model. However, in order to quantify this precision, calculating the r-squared is not sufficient. We need another statistic, which is called the standard error of the estimate, or SEE.

We will now explain the standard error of the estimate (SEE).

As shown in the textbook, SEE is effectively an average value of SSE (unexplained variations). Think of it also as the root means square error, or by how much on average the fitted values $\hat{y}$ deviate from the actual values y.

SSE measures by how much the actual values y are scattered around the corresponding estimated values $\hat{y}$. This statistic is the standard deviation of the fitted values $\hat{y}$, but not from their mean, but from the actual values y.

If SEE is effectively a standard deviation measuring how the actual values y are scattered around their corresponding model estimated values $\hat{y}$, then using this statistic we can construct a confidence interval around our model values. The more actual values fall within this confidence interval, the more precise our model is. This is another goodness of fit measure, but it just takes a different spin on the same topic.

We know from the standard error of the mean (SE) that this value, when combined with the z-value (or t-value), estimates the interval in which the true mean is likely to reside. So, for example, $\overbar{x}$±1.96SE produces a 95% confidence interval within which a true value of μ is likely to reside. Comparably, $\overbar{x}$±1.65×SE provides 90% confidence interval and $\overbar{x}$±2.58×SE provides 99% interval, for example.

The same principle applies to SEE. In other words, $\hat{y}$±1.96SEE will provide 95% confidence interval where the true value of y is likely to be. As this is used in the context of regression analysis, the expression confidence interval is replaced with the phrase prediction interval. So, prediction interval states with a certain percentage a confidence that the actual values are in the region (interval) predicted by the model.

If SEE is small, the prediction interval will be narrower (or, closer to the regression line), which means that the actual observations are more closely scattered around the regression line. This indicates that the model is more precise. A wider prediction interval implies less precise model and means that its SEE is larger.

In summary, R-squared is determined by how close are the actual values y to the line of regression ŷ. Why? Because it measures how well the changes in y (or, variations in y) respond to the changes in x (or variations in x). If they move exactly as the regression model ŷ predicted, then this gives us the high value of r-squared. If they do not, i.e. the changes in x generate no change in y, then the r-squared value is zero. Essentially, r-squared value is the measure of the goodness of fit for regression models, i.e. how well the given model fits (or approximates) the actual observations.

SEE, on the other hand, quantifies how precise is the model by establishing the prediction interval within which we expect, with certain confidence, that the actual values to reside. The narrower this interval, the more precise our model is. Or to put it differently, a more precise model will have the actual values much closer to the regression line. However, unlike the r-squared which is a single coefficient, SEE helps us build an interval that provides a confidence level for the model.

To conclude, both r-squared (coefficient of determination) and SEE (standard error of the estimate) are goodness of fit measures. R-squared indicates how well the model approximates the data and SEE quantifies how precise the model is.