Covariance, Correlation, R-Squared

Deepak Khandelwal Follow
Dec 14, 2019 · 12 min read ★

This article is dedicated to the explanation of the three closely related but still different concepts namely — covariance, correlation and R-Squared. This also describes about the interpretation of the three.

$$\mathbf{VAR(X)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))^2 \tag{1}$$

$$\mathbf{COV(X,\ Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y)) \tag{2}$$

$$\mathbf{COR(X,\ Y)} = \frac{\mathbf{COV(X,\ Y)}}{\sqrt{\mathbf{VAR(X)VAR(Y)}}} \tag{3}$$

$$\mathbf{R}^2 = 1 - \frac{\mathbf{VAR(X,\ Y)}_{FittedLine}}{\mathbf{VAR(X,\ Y)}_{Mean}} \tag{4}$$

> ***What is variance?***

Random variables, by definition, can take different values. The range of values a random variable takes and the variation among them is determined by the distribution of that random variable. Next, the expected value (aka mean, average, expectation) of a random variable tells us what value we can expect from a random variable on an average. For example, throwing a six sided die is a random variable (D) which can take the following values — [1, 2, 3, 4, 5, 6]. The probability of D taking any of the six values is 1/6 for all the values. Hence D follows a uniform distribution. The expected value of D, denoted as E[D], is 3.5.

But what we are interested in is what maximum and minimum values it can take and what is the variation from the expected value. This is defined by the variance. The variance gives us a measure of how much the random variable X deviates from its expected value E(X). It measures the variability among data, i.e., how much the quantity, the random variable represents, fluctuates.

Furthermore, it is entirely possible that two random variables X and Y, having different distributions, have same mean but different variances. For example
X takes the values — [3, 3, 3, 3] , Y takes the values — [1, 3, 3, 5]

```
E(X) = 3, E(Y) = 3
VAR(X) = 0, VAR(Y) = 2
```

We can see that both X and Y have same mean, but different variations — 0 and 2 respectively.

The importance of variance can be understood by the following real life example — suppose an investment is made in real estate sector. Although, the expected amount for the return might be a profit, but we want to actually know how much it can vary, how much low and high it can go so that we can calculate the risk and plan our investment accordingly.

Mathematically, the variance of a random variable X is defined as —

$$\mathbf{VAR(X)} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - E(X))^2 \qquad \text{where n is the number of observations}$$

$$x_i \text{ is the individual observation}$$

Figure 1: Variance

> ***What is covariance?***

Now we extend and generalize the idea of variance. We have seen the variance of one random variable as shown in figure 1. We can rewrite it as —

$$\frac{1}{n-1}\sum_{i=1}^{n} (x_i - E(X))(x_i - E(X))$$

Now consider instead of one random variable, we have two random variables X and Y, we can exploit the above formula and we can write —

$$\mathbf{COV(X,Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

Figure 2: Covariance

This explains how much X varies from its mean when Y varies from its own mean. It is a statistical measure used to analyze how two random variables behave as a pair. To illustrate the concept with examples, I have taken three datasets to explain three different trends represented by covariance.

Positive Trend —

The Figure 3 shows the ice-cream dataset which has two random variables — Temperature and Revenue. It has the information of how income of ice cream sale varies as temperature varies. Using this dataset, we try to find the covariance between Temperature and Revenue.

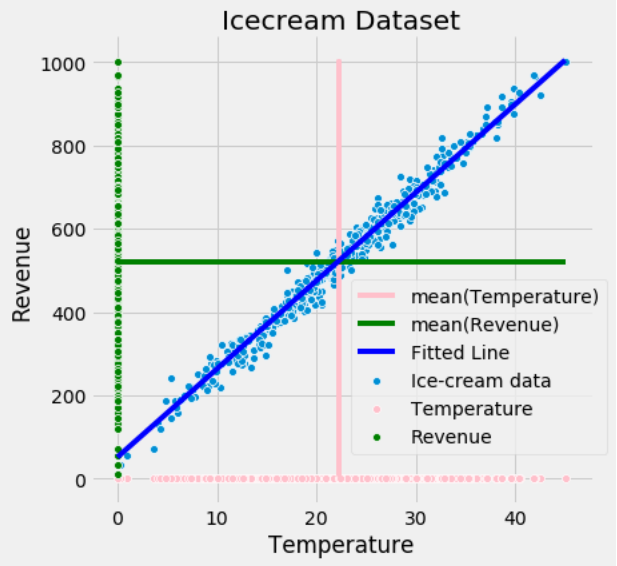


Figure 3: Ice-cream dataset

These observations have been taken from the same ice-cream shop and have been taken in pairs i.e. we have pairs of (Temperature, Revenue). Now we are interested to know whether observations taken in pair (Temperature, Revenue) possess some information which individual observation (Temperature or Revenue) does not.

We have the following distributions of Temperature and Revenue.

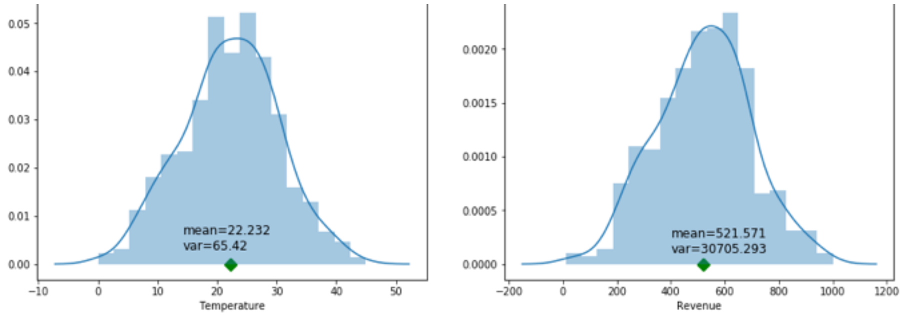


Figure 4: Distribution of Random variables — Temperature and Revenue

```
+-------------------------+-------------------------+
| E(Temperature) = 22.232  | E(Revenue) = 521.571     |
+-------------------------+-------------------------+
| VAR(Temperature) = 65.42 | VAR(Revenue) = 30705.293 |
+-------------------------+-------------------------+
|     COV(Temperature, Revenue) = 1405.66162842       |
+-------------------------+-------------------------+
```

ice_cream_dataset.txt hosted with ❤ by GitHub view raw

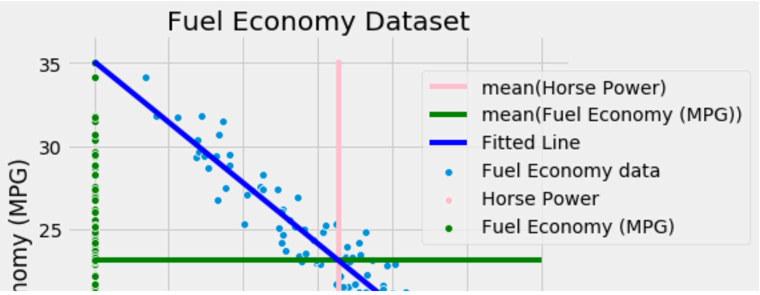
Table 1: Ice-cream dataset — Covariance

Generally speaking, the pairs with relatively low values of Temperature also have relatively low values for Revenue and the pairs with relatively high values of Temperature also have relatively high values for Revenue. The blue line in Figure 3 represents this relationship. Note that this line has positive slope and it represents positive trend where values for Revenue increases with the values for Temperature.

Now we will measure the covariance between the two. We can observe in figure 3 that if temperature increase, revenue also increase.

Negative Trend —

Now I have taken fuel economy dataset which has two random variables — HorsePower and FuelEconomy as shown in Figure 4.. Here, we can see that blue line has fitted the data and has negative slope. This means that for relatively higher values for HorsePower, we have relatively lower values for FuelEconomy and vice-versa. This represents negative trend in the data.



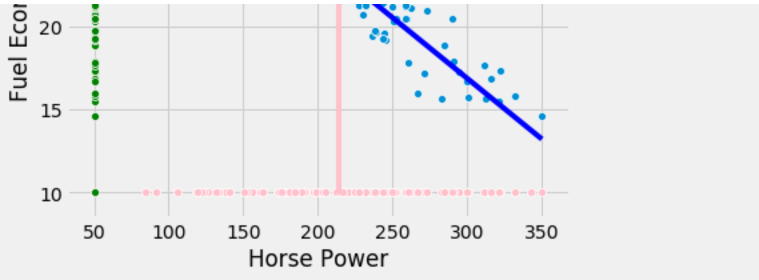


Figure 5: Fuel Economy Dataset

We have the following distributions of HorsePower and FuelEconomy.

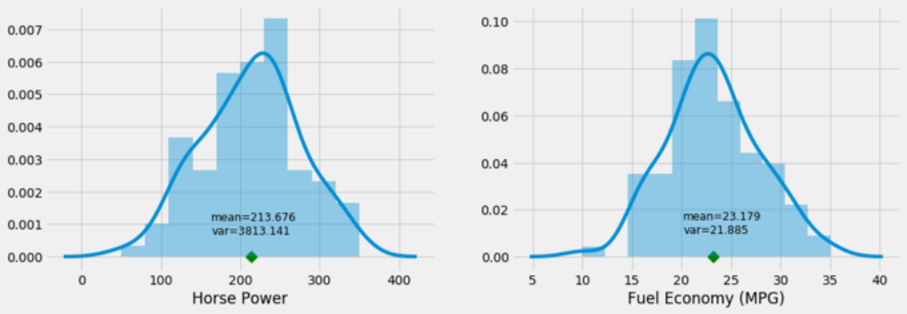


Figure 6: Distribution of Random variables — HorsePower and FuelEconomy

```
+----------------------------+---------------------------+
| E(HorsePower) = 213.676    | E(FuelEconomy) = 23.179   |
+----------------------------+---------------------------+
| VAR(HorsePower) = 3813.141 | VAR(FuelEconomy) = 21.885 |
+----------------------------+---------------------------+
|      COV(HorsePower, FuelEconomy) = -278.28281708      |
+----------------------------+---------------------------+
```

fuel_economy_covariance.txt hosted with ❤ by GitHub view raw

Table 2: Covariance with Negative Trend

No Trend —

If we have data that has following distribution —

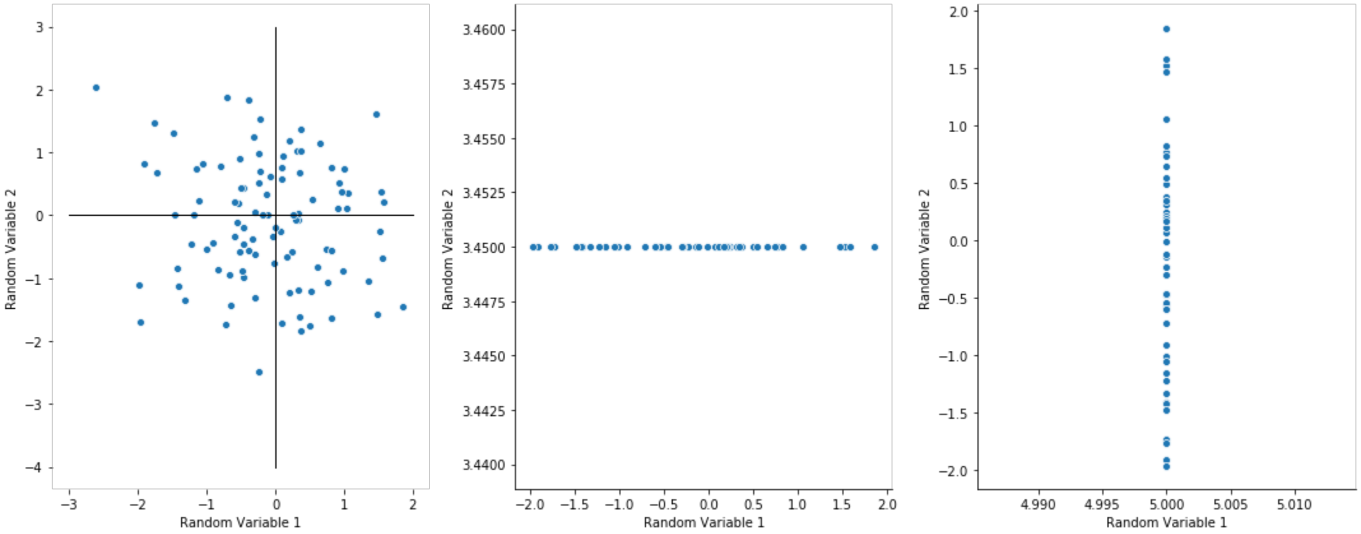


Figure 7: Distribution of Random variables —No trend

In Figure 7, the left plot does not show any relationship between random variable 1 and random variable 2, data is scattered all over the place. There is no defined pattern.
In middle plot, every value of random variable 1 is paired with the same value of random variable 2. The random variable 2 remains constant while decreasing or increasing the value of random variable 1.
In right plot, every value of random variable 2 is paired with the same value of random variable 1. The random variable 1 remains constant while decreasing or increasing the value of random variable 2.
We cannot determine that for relatively smaller or larger values of one random variable, what value another random variable will take. For plots in Figure 7, covariance is close to zero.

Hence, the covariance can define three types of relationship —
1) Relationship with positive trend
2) Relationship with negative trend
3) When there is no relationship because there is no trend in data.

But there are some questions which covariance is not capable of answering.

Limitations of Covariance —

1. How to interpret?
   Covariance is hard to interpret. We will understand this by taking the ice-cream dataset. We multiply each random variable — Temperature

and Revenue, by 2.

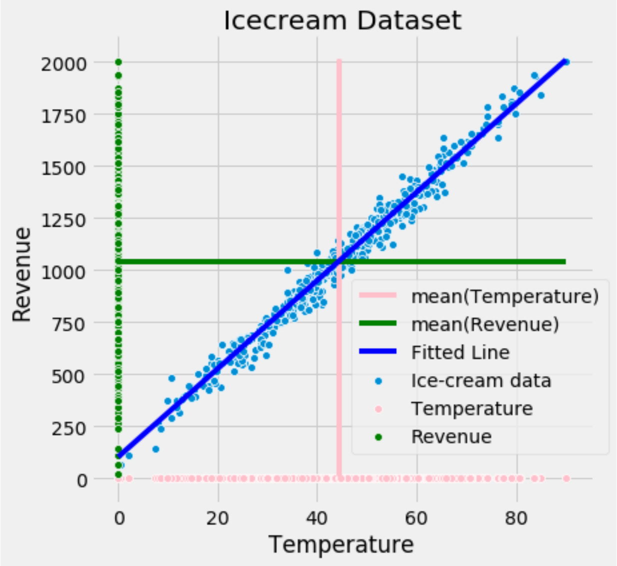


Figure 8: Ice-cream dataset — scaled

Again, we fit a line to the data, the new slope and intercept values are also 2 times the values of the model without scale.
Figure 8— {'Intercept': 1045.199891839289, 'Slope': 342.4126032448046}
Figure 3 — {'Intercept': 522.5999459196445, 'Slope': 171.20630162240232}

The relative position of data does not change and hence each data points stills lie around the same fitted line with positive slope. But covariance between Temperature and Revenue becomes 5622.64652 which is 4 times the original value 1405.66163.
We have only changed the scale of the data, but coefficient value changes even the relationship still remains intact. In other words, covariance value is sensitive to the scale of the data and this makes it difficult to interpret.

2. It tells us that the slope is negative or positive. It does not answer the question — Are data points relatively close or far to the fitted line?
Since covariance is sensitive to scaling, this prevents it from explaining whether data points are close or far to the fitted line that represents the relationship. We will verify this with following example.

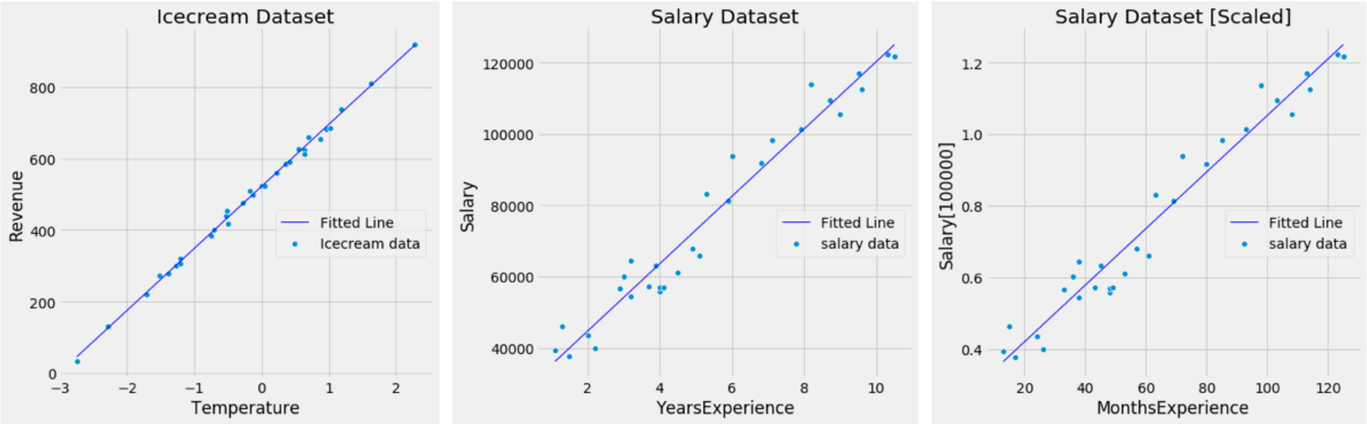


Figure 9: Ice-cream Revenue and Employee Salary Data

In Figure 9, the left plot has data points very close to the fitted line.The covariance between Temperature and Revenue is 221.541. In middle plot, we have salary dataset, here the covariance between YearsExperience and Salary is 76106.303.
We scale the salary data by converting the years of experience into month of experience and dividing the salary by 100000. Then we fit the data in the right plot, the distribution of the data points around the fitted line is same as it is in middle plot, but covariance value goes down to 9.099 and it is quite lesser than the one for the left plot with data points close to the fitted line.

This explains that by just looking at the covariance value we cannot say that data points will be close or far to the fitted line. It may happen that we have vary low magnitude of covariance but data are widely scattered around the fitted line and vice-versa as is the case in this example.

> ***What is correlation (Pearson's Correlation Coefficient — R)?***

We have seen that covariance provides the **direction (positive, negative, near zero)** of the linear relationship between two variables. On the other hand, in addition to the **direction**, correlation also provides the **strength** of the relationship.

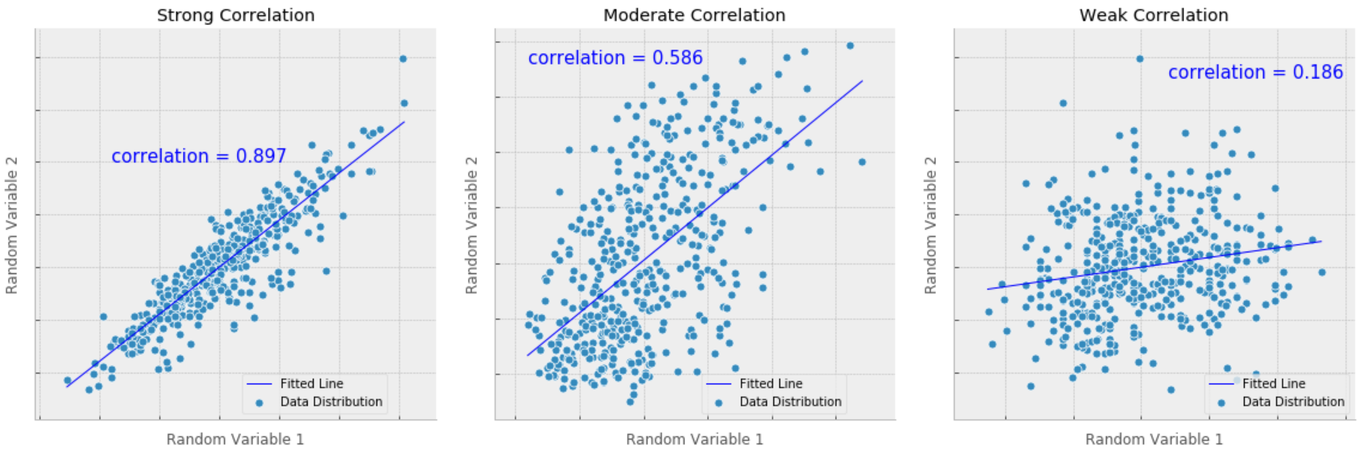


Figure 10: Correlation — Strong, Moderate, Weak

Correlation is not causation.

When we say that the relationship between one random variable is relatively strong with another random variable, this means that we have *observed* that low or high values of one random variable tend to be paired with relatively low or high values of another random variable and this observation suggests a trend and that we can use to make inference.

We are not saying that low or high values for one random variable *causes* another random variable to have low or high values. In other words, we can say that we do not rule out the possibility that something else can cause the trend that we observe. This further means that two completely unrelated factors that may have mathematical correlation but have no sensible relationship in real life.

By far, we have learnt that how correlation gives us the strength of relationship between two random variables. Now we will see how we can quantify this strength of relationship. Mathematically, we have the following formula for correlation —

$$\mathbf{COR(X, Y)} = \frac{\mathbf{COV(X, Y)}}{\sqrt{\mathbf{VAR(X)VAR(Y)}}}$$

As we can see that correlation between X and Y is simply the covariance between them divided by square root of variance of X and variance of Y multiplied. It is analogous to the idea of how standard deviation is calculated by taking square root of the variance. Hence correlation is normalized covariance. Thus we restrict the values of correlation between -1 and 1.

No Correlation [COR(X, Y) = 0]
When correlation between two random variable is zero, the value on x-axis(Random Variable 1) doesn't tell us anything about what to expect on y-axis (Random Variable 2), because there is no reason to choose one value over another.

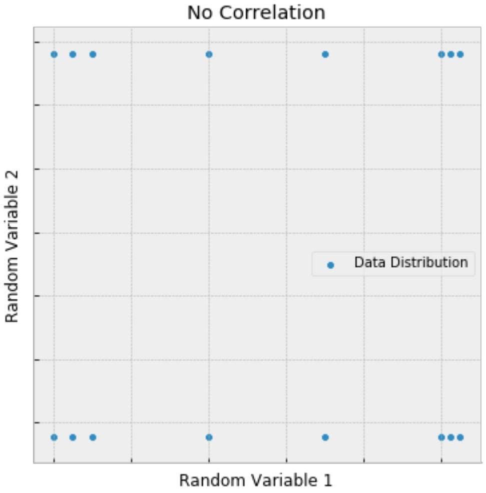


Figure 11: No Correlation

As long as correlation is not exactly zero as in Figure 10(weak correlation), we can fit a line to the data to make predictions, but those predictions would be less confident.
Furthermore, the confidence about the prediction by the fitted line depends on the sample size and the p-value. The smaller the p-value the more confidence we can have in our predictions.

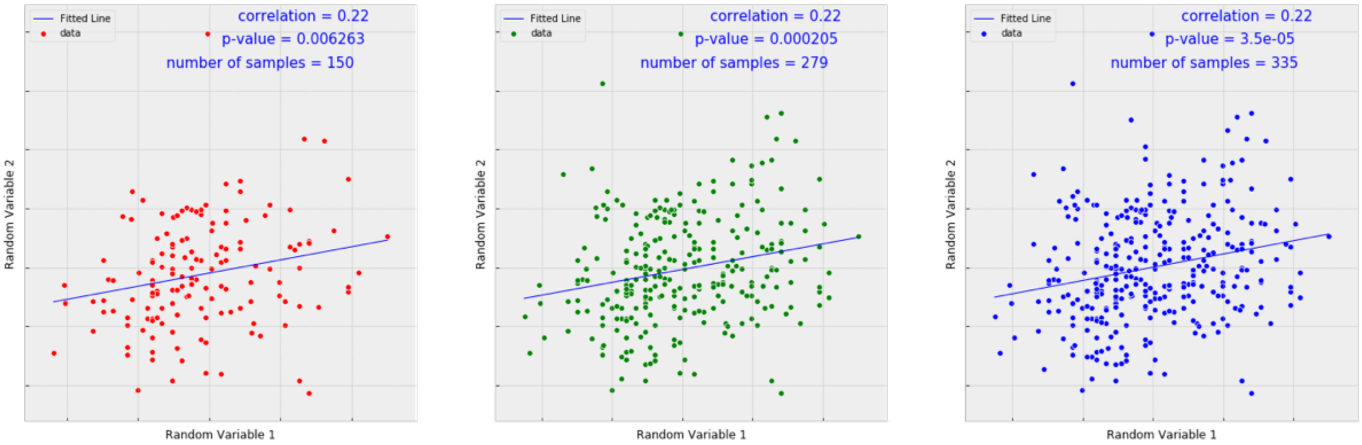


Figure 12: Correlation and P-value

In Figure 12, all three examples have same correlation value. In this case, adding more data points does not increase the correlation and it does not refine the prediction. It has just increased the confidence in the prediction. The predictions will be very inaccurate in all the three examples. In other words, just because we have huge data, we will have pretty strong confidence in prediction, but if the correlation value is small, our predictions will still be very inaccurate.

Limitations of Correlation —

Top highlight

Even correlation is way easier to interpret than covariance value, but it also has some limitation in its interpretation. For example, it is not obvious that strong correlation (= 0.897) in Figure 10 is twice or more as good at making predictions as moderate correlation (= 0.586) is.

> ***What is R squared?***

R-Squared is also known as **coefficient of determination.** The relationship between R and R-Squared is that We can get the value of R-Squared just by squaring the value of R. But they have different interpretation.

R-Squared is used to find the correlation between the predicted and actual values of dependent variable.

R-Squared is a measure of how much of the variance in the actual value of dependent variable (y) can be explained by its relationship to the independent variable (X). In other words, R-Squared is the percentage of variance in y explained by the linear regression equation between X and y. To illustrate the idea behind the R-Squared, I have taken the ice-cream dataset in Figure 3. We starts fitting the data by finding the mean of revenue.

This animation shows that how the value of R^2 changes as the training reaches to the best fitted model. We start to fit the data by calculating the mean of ice-cream revenue and plotting it as a line that spans the data and we calculate the R-Squared using the following formula —

A perfect fit would have a R^2 of 1. R^2 value can also be negative because the model can be arbitrarily worse, in that case it will have variance more than the variance around the mean line, and we will have a negative R^2 value. It need not actually be the square of the quantity R (correlation).

To support this argument, let’s experiment with concrete compressive strenght dataset and try to fit a model using linear regression. The 15-fold cross-validation has been performed on the dataset after scaling it using `StandardScaler` in scikit-learn. The evaluation metric is chosen as `r2`. A working notebook demonstrating the below result is attached. The result of the cross-validated models are as follows —

Negative R-Squared Value on Validation Set

It can be observed that for some models, R^2 value is negative. It indicates that the model can perform strangely on unforeseen data giving negative R^2 value.

Figure 13: Fitting the ice-cream data by taking mean of revenue

Now the question is that is the mean line the best way to predict ice-cream revenue? Obviously not!
Since We have fitted the data by just taking the mean of ice-cream revenues, the fitted line is the mean itself, and R^2 value is 0. The variance around that line is very high. But we can do much better by fitting a line to data as shown in Figure 14.

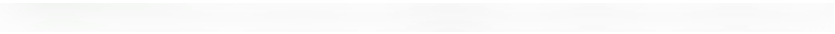


Figure 14: Best Fitted line

By just looking at the Figure 14, we can infer that the new line fits the data better the line in Figure 13. Now our goal is to quantify that difference. We use R² to quantify the difference between predictive power of the line in Figure 13 and the line in Figure 14.

We have R² value as 0.97 for the fitted line in Figure 14. Now we can interpret this value in the following ways —
1. There is around 97% less variation around the line than the mean.
2. The relationship between predictor variable(X) and target variable(y) accounts for 97% of the variation.
3. The 97% of the variation in the data is explained by the relationship between X and y. The remaining 3% is explained by something else or may be it is due to noise in the data.

For regression problems, goodness-of-fit is often determined with the R² aka coefficient of determination. R² can also quantify relationships that are more complicated that a simple straight line.

Limitations of R-Squared —
1. R-Squared tells us how much percentage of variation in y can be explained by the linear model between X and y but it does not tell how much percentage of entire y can be explained by the linear model.
2. Since variance is dataset dependent, R² may not be meaningfully comparable across different datasets.
3. It does not give any indication about the direction of the relationship.

> ***Summary***

1. Covariance value has no upper or lower limit and is sensitive to the scale of the variables. While correlation value is always between -1 and 1 and is insensitive to the scale of the variables.
2. Correlation value tells about the strength as well as direction of the relationship. But correlation strength does not necessarily mean the correlation is statistically significant; will depend on sample size and p-value.
3. The R-Squared can take any value in the range [-∞, 1]. The close the value to 1 the better the explanatory power of the independent variable is. It helps explain the variability in data.

Correlation R Squared Covariance Statistics Mathematics