COORELATION COEFFICIENTS

ARUMUGAM SWETHA-EE20BTECH11005,
YENUGULA HASHISH-EE20BTECH11056

March 2022

Abstract

Correlation is a statistical measure that expresses the extent to which two variables are related. Correlation tests are used as a bias in many applications such as exploratory data analysis, structural modeling, data engineering and many more. To quantitatively assess whether two datasheets are correlated, correlation coefficients, such as Pearson's sample correlation coefficient, Spearman rank correlation coefficient and Kendall Tau are used. The idea of this project is to analyze correlation coefficients and its characteristics.

1 Introduction

Correlation coefficient is a measure that determines the degree to which the movement of two different variables are associated. Its value always ranges between -1 (indicating strong negative correlation) and +1 (indicating strong positive correlation). As the value goes towards zero, the correlation between the variables grows weaker.

1.1 Pearson Correlation Coefficient

Pearson correlation coefficient is the most used coefficient, measures the strength of the linear relationship between two variables.

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum((X-\overline{X})^2).((Y-\overline{Y})^2)}}$$

Assumptions

- Both the variables must be normally distributed which means they have a bell-shaped curve.

- Both variables are continuous, jointly normally distributed, random variables. They follow a bi-variate normal distribution in the population from which they were sampled.

- If there is a relationship between jointly normally distributed data, it is always linear.

1.2 Spearman Correlation Coefficient

Spearman's correlation coefficient measures the strength and direction of monotonic association between two variables. It is a non-parametric test, that does not carry any assumptions about the distribution of the data.

$$\rho = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n}((R(x_i) - \overline{R(x)})^2).\sum_{i=1}^{n}((R(y_i) - \overline{R(y)})^2)}}$$

Assumptions

- It assumes that there is monotonic relationship between the two variables.
- Two variables that are either ordinal, interval or ratio.
- Pair of observations are independent.

1.3 Kendall tau Correlation Coefficient

Kendall's Tau is a non-parametric measure of relationships between columns of ranked data which can be used with ordinal or continuous data.

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

Assumptions

- It assumes that there is monotonic relationship between the two variables.
- Two variables that are either ordinal, interval or ratio.
- Pair of observations are independent.

1.4 P-values

A p-value is the probability that the null hypothesis is true.

- A p-value ≤ 0.05 is statistically significant, and this means there is less than 5% probability the null is correct. Therefore, we reject the null hypothesis, and accept the alternative hypothesis, but this does not mean that there is a 95% probability that the alternative hypothesis is true.

- A p-value $\geqslant 0.05$ is not statistically significant, and this means we retain the null hypothesis and reject the alternative hypothesis. We should note that we can not accept the null hypothesis, we can only reject the null or fail to reject it.

2 Analysis

Correlation analysis in research is a statistical method used to measure the strength of the correlation between two variables and compute their association.

2.1 Based on property of data

In a linear relationship, the variables move in the same direction at a constant rate. In a monotonic relationship, the variables tend to move in the same relative direction, but not necessarily at a constant rate. The research hypothesis is to find whether the data is correlated and find the types of data to which the correlation coefficients are applicable. This is done for linear, monotonic, logarithmic, symmetric U-shaped data.

For codes for the below plots go to Github

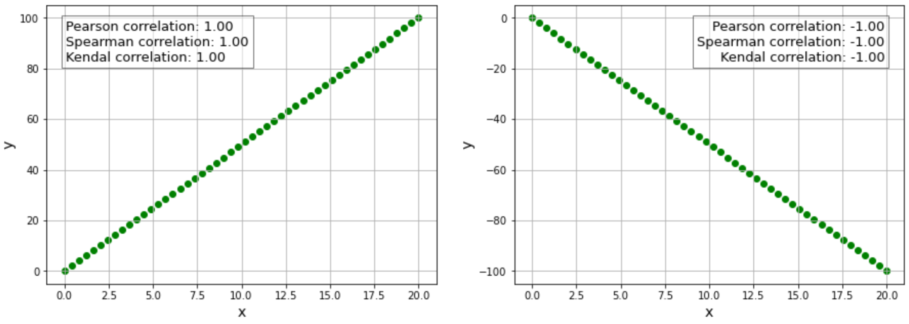

Figure 1: Correlation coefficients for a linear relationship

For figure-1 In the linear relationship, all correlation coefficients are one.
For figure-2, In the logarithmic relationships, only the two non-parametric correlation coefficients are +1 or -1.
For figure-3 In the exponential relationships, only the two non-parametric correlation coefficients are +1 or -1.
For figure-4, In the symmetric U-shaped relationship, all correlation coefficients are zero.

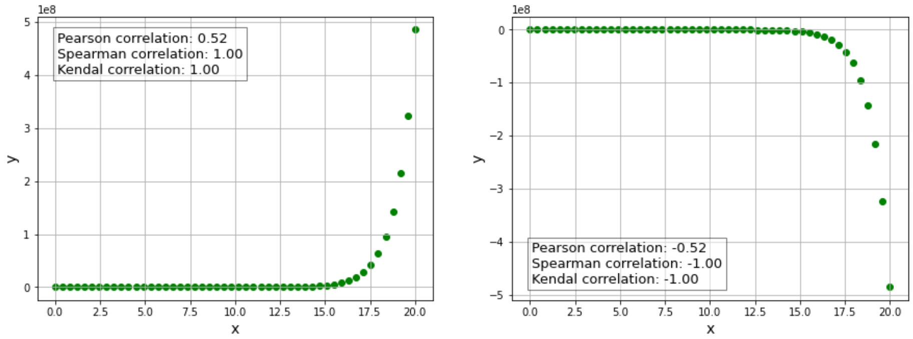

Figure 2: Correlation coefficients for a monotonic relationship

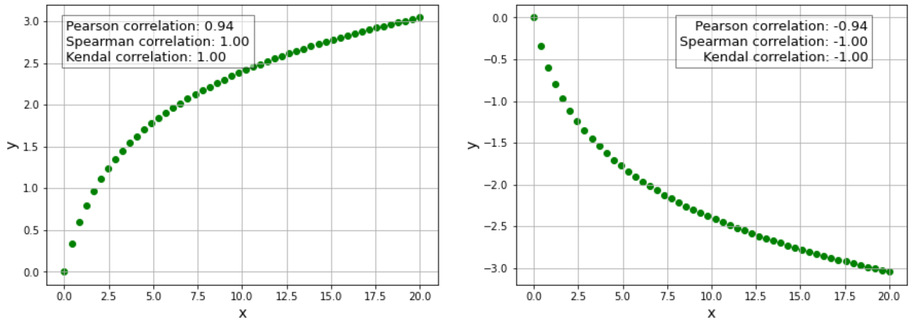

Figure 3: Correlation coefficients for a logarithmic relationship

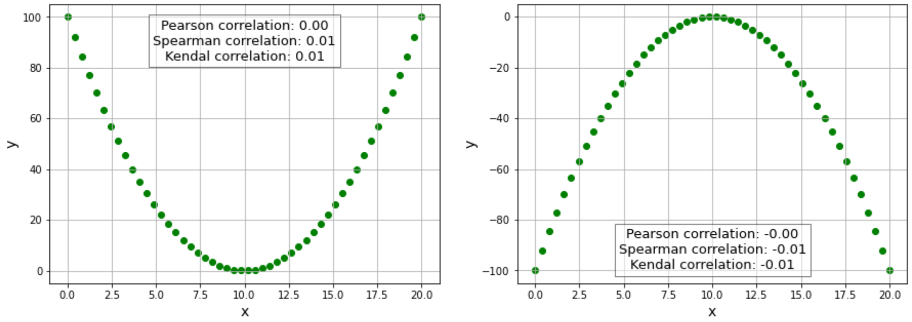

Figure 4: Correlation coefficients for symmetric U-shaped relationship

Result

Correlation coefficients only measure linear (Pearson) or monotonic (Spearman and Kendall) relationships.

2.2 Based on outliers

In most practical circumstances an outlier decreases the value of a correlation coefficient and weakens the regression relationship, but it's also possible that in some circumstances an outlier may increase a correlation value and improve regression. So we must exclude the outlier data points and analyse the data.

The experimental hypothesis is to analyze different data sets and how outliers affect the correlation coefficients.

2.2.1 Anscombe's quartet

Anscombe's quartet comprises of four data sets that the four y variables have the same mean-7.5, variance-4.12, correlation-0.816 and linear regression line: $y = 3 + 0.5x$, but the distributions are different.

For codes for the below plots go to Github

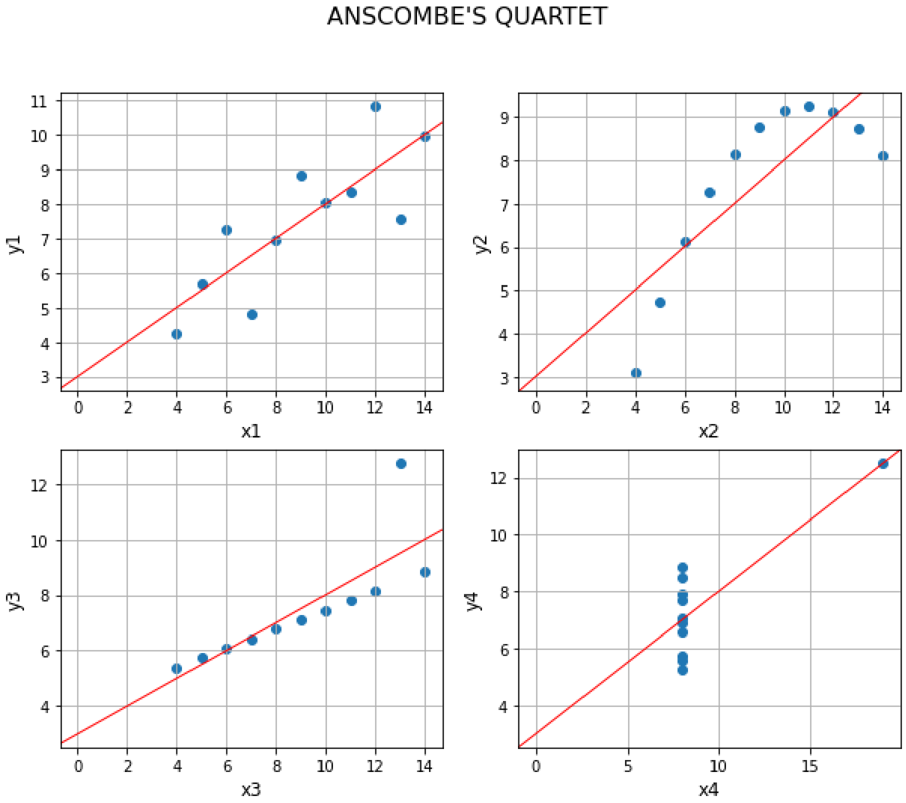

Figure 5: four data sets with same correlation of 0.816

- In the first graph, plot is distributed normally, and corresponds to what one would expect when considering two variables correlated.

  - r = 0.81642 and p_p-value = 0.00217
  - ρ = 0.81818 and p_s-value = 0.00208314
  - τ = 0.63636 and p_k-value = 0.00570717

- In the second graph,plot is not distributed normally and clearly it is not linear. Here in this case, Pearson correlation does not indicate the exact functional relationship.
  - r = 0.81624 and p_p-value = 0.00218
  - ρ = 0.69091 and p_s-value =0.01856503
  - τ = 0.56364 and p_k-value = 0.01654050

- In the third graph, there is a perfect linear relationship, except for one outlier which influence to lower correlation coefficient from 1 to 0.0816.
  - r = 0.81629 and p_p-value = 0.00218
  - ρ = 0.99091 and p_s-value = 0.00000000
  - τ = 0.96364 and p_k-value = 0.00000055

- In the fourth graph is another example of outlier but it is opposite to the third one. It produces a high correlation coefficient, even though the relationship between the two variables is not linear.
  - r = 0.81652 and p_p-value = 0.00216
  - ρ = 0.50000 and p_s-value = 0.11730680
  - τ = 0.42640 and p_k-value = 0.11384630

Result

- The Pearson correlation coefficient indicates the strength of a linear relationship between two variables, but its value does not completely characterize their relationship. In particular, if the conditional mean of Y given X, denoted $E(Y \mid X)$, is not linear in X, the correlation will not fully determine the form of $E(Y \mid X)$.
- Pearson coefficient is therefore sensitive to outliers in data, and it is not robust against them.
- From figure-3 by comparing p_s and p_k values, We can conclude that Kendall tau is more robust and efficient to outliers than the other two coefficients.

2.2.2 Based on distribution of correlation coefficient

The experimental hypothesis is to find how the outliers affects the distribution of correlation coefficients. This is done by plotting bootstrap estimates of the distribution of Pearson's, Spearman's, and Kendall's correlation coefficients based on 2000 re-samplings of the 1000 points.

For codes for the below plots go to Github

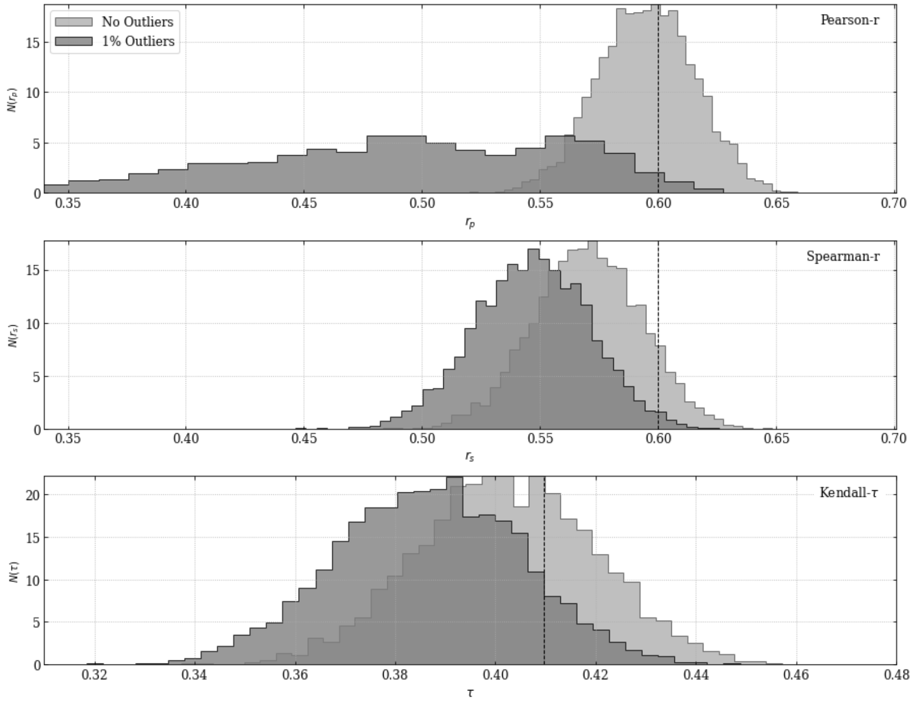

Figure 6: Distribution of correlation coefficients without and with (1%) outliers

2000 bootstrap re-samples of 1000 data points drawn from a bi-variate Gaussian with ρ=0.6 without and with (1%) outliers. The true values are shown by the dashed lines. It is clear that Pearson's correlation coefficient is not robust to contamination.

Observation

- Pearson correlation coefficient is clearly not robust against outliers.
- Spearman and Kendall correlation coefficient have variance which is robust to outliers.

2.2.3 Comparison of Spearman's and Kendall's correlation coefficient

From previous section, we concluded Pearson's correlation coefficient is very sensitive to outliers. Now we do the same analysis as the previous section with 5% outliers.

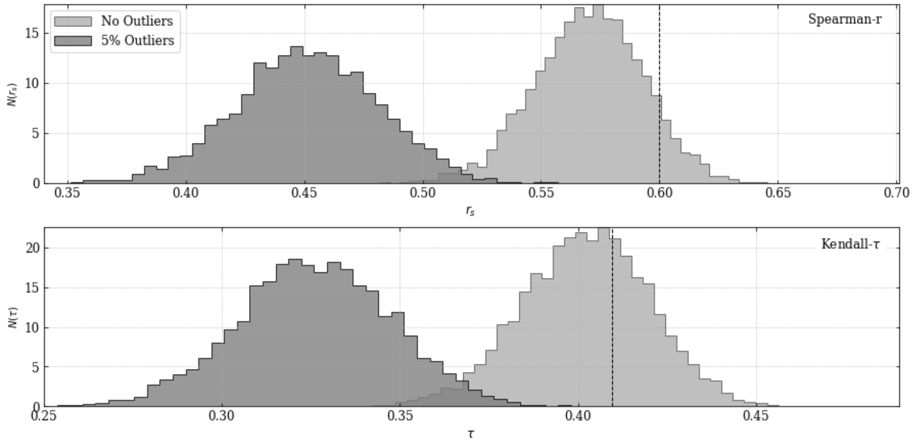

Figure 7: Distribution of correlation coefficients without and with (5%) outliers

From the above plot, we can tell that Kendall correlation coefficient is more robust and effective than Spearman correlation coefficient. This is because Kendall tau correlation has smaller gross error sensitivity(GES) and a smaller asymptotic variance(ASV). Thus, Spearman correlation coefficient is more effected by outliers than Kendall tau correlation coefficient.

From the article* we know that,

$$GES(\tilde{R}_K, \Phi_\rho) = \pi\sqrt{1-\rho^2}\left[\frac{2}{\pi}\arcsin(|\rho|)+1\right]$$

$$GES(\tilde{R}_S, \Phi_\rho) = \pi\sqrt{1-\frac{\rho^2}{4}}\left[\frac{6}{\pi}\arcsin(|\frac{\rho}{2}|)+1\right]$$

Figure 8: where ϕ_ρ is a bi-variate normal distribution with correlation ρ

We observe that GES depends on parameter ρ in a non-linear way. Kendall's correlation coefficient is preferable than Spearman correlation coefficient although the difference GES is negligible for smaller values of ρ.

Also from the article, we know that efficiency of both the correlations is higher than 70% for all possible values of ρ. But on comparing Kendall's correlation coefficient is more efficient than Spearman correlation coefficient by a small amount.

2.3 Normality

Central Limit Theorem says that: given random and independent samples of N observations each, the distribution of sample means approaches normality as the size of N increases, regardless of the shape of the population distribution. Plotting correlation coefficients distribution for large number of samples similar to that of previous plot, we get

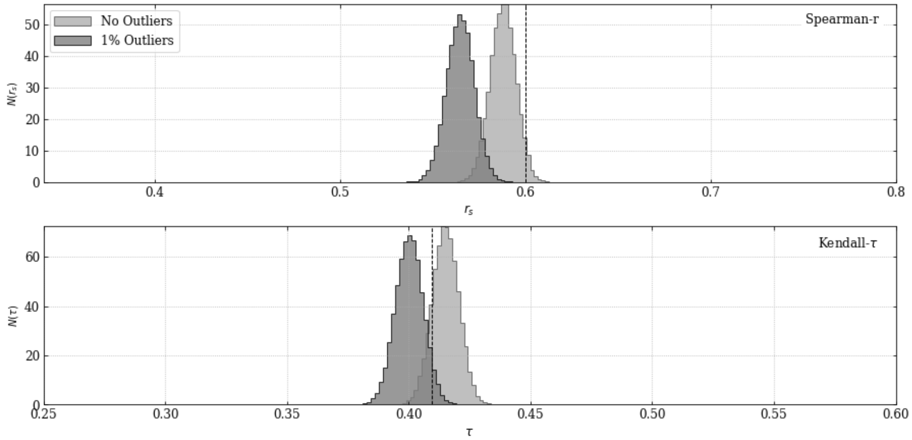

Figure 9: Distribution of correlation coefficients without and with 1% outliers for 10000 samples

From the above plot, we can observe that as the number of samples increases the distribution becomes more normal. And Kendall's tau approaches normality faster than Spearman.

3 Conclusion

In this project, we analyse different correlation coefficients and understand some of their properties. Pearson correlation coefficient considers a full linear dependence (fitting to a straight line) and in requires the variables to be normally distributed. It is parametric test and is very fragile to outliers. Spearman and Kendall Tau correlation coefficients are non-parametric test and are the appropriate tests when data is ordinal.

We showed that though generally, an outlier decreases the value of a correlation coefficient and weakens the regression relationship, it's also possible that in some circumstances an outlier may increase a correlation value and improve regression.

Pearson's correlation coefficent is not robust against outliers. Kendall Tau is more preferable due to its high efficiency and robustness than Spearman. Kedall tau corr coeff approaches normality faster than spearman correlation coefficient.

References

Influence functions of the Spearman and Kendall correlation measures Christophe Croux · Catherine Dehon

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/

https://www.kaggle.com/code/kiyoung1027/correlation-pearson-spearman-and-kendall/report?scriptVersionId=25999032

https://en.wikipedia.org/wiki/Correlation

https://www.astroml.org/book$_figures_1ed/chapter3/fig_correlations.html$

https://www.tandfonline.com/doi/full/10.1080/02626667.2011.586948

https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation$_coefficients_{appropriate_use_an}$