

How to Calculate and Interpret Correlation Measures in Stata

Authored by
stats writer

March 8, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Calculate and Interpret Correlation Measures in Stata*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=134644>

An In-Depth Overview of Correlation Analysis in Stata

In the expansive field of **quantitative research**, **Stata** stands as a premier **statistical software** package, renowned for its efficiency in **data analysis** and data management. One of its most fundamental and frequently utilized capabilities is the assessment of relationships between variables, a statistical concept formally known as **correlation**. Correlation serves as a critical diagnostic tool, allowing researchers to determine how changes in one variable are associated with changes in another. By quantifying these associations, analysts can uncover patterns, validate theoretical frameworks, and refine **econometric** or psychometric models. Within the Stata environment, users have access to a sophisticated suite of correlation measures, each tailored to specific data types and distributional assumptions.

The utility of **correlation** analysis in **Stata** extends beyond mere descriptive statistics; it is often a preliminary step in more complex modeling, such as **linear regression** or **factor analysis**. Understanding the nuances of each correlation type--ranging from **Pearson** to **Kendall**--is essential for avoiding common pitfalls like **spurious correlations** or the misinterpretation of non-linear associations. Stata's command-driven architecture provides a transparent and reproducible workflow for these calculations, making it a favorite among academics and industry professionals alike. Whether one is dealing with large-scale **panel data** or small experimental samples, the choice of a correlation coefficient remains a pivotal decision in the analytical process.

This comprehensive guide explores the various **correlation** measures available in **Stata**, detailing their mathematical foundations, appropriate use cases, and practical implementation. By mastering these tools, researchers can ensure their findings are both statistically robust and theoretically sound. In the following sections, we will delve into the specific coefficients that accommodate continuous, ordinal, and binary data, while also addressing how to control for confounding factors using multivariate techniques. The goal is to provide a clear roadmap for navigating the complexities of **bivariate analysis** within one of the world's most powerful statistical environments.

The Foundations of Pearson's Product-Moment Correlation

The **Pearson correlation coefficient**, often denoted as 'r', is the gold standard for measuring the linear relationship between two **continuous variables**. In **Stata**, this measure is the default for most correlation commands because it effectively captures how well a straight line can describe the relationship between data points. The coefficient ranges from -1 to +1, where a value of +1 signifies a perfect positive **linearity**, -1 indicates a perfect negative linearity, and 0 suggests the absence of any linear association. It is important to note that Pearson's 'r' is sensitive to the magnitude of the variables and assumes that the data follows a **normal distribution**, making it a **parametric** statistic.

When applying the **Pearson correlation coefficient**, researchers must be mindful of its underlying assumptions. Beyond normality, it requires **homoscedasticity**--the condition where the variance of the residuals is constant across all levels of the independent variable. If these conditions are met, Pearson's 'r' provides a highly efficient and easily interpretable metric. However, it is notoriously sensitive to **outliers**, which can disproportionately pull the correlation coefficient toward a higher or lower value, potentially masking the true underlying relationship between the **continuous variables**. Consequently, visualizing data through scatterplots in Stata is a recommended practice before relying solely on the numerical output.

In **Stata**, calculating the Pearson coefficient is typically done using the **pwcorr** or **correlate** commands. The distinction between these commands is subtle but important: **correlate** uses **listwise deletion** for missing values, while **pwcorr** (pairwise correlation) calculates the coefficient based on all available data for each pair of variables. This flexibility allows analysts to maximize their sample size, especially when working with messy real-world datasets. Furthermore, the ability to append the **sig** option to these commands provides immediate feedback on **statistical significance**, helping researchers determine if the observed **correlation** is likely a result of chance or a genuine population-level phenomenon.

Non-Parametric Alternatives: Spearman's Rank Correlation

When data does not meet the strict assumptions of normality or **linearity**, **Spearman's rank correlation coefficient** (Spearman's rho) provides a robust alternative. Unlike Pearson, Spearman's rho measures the **monotonic** relationship between variables, meaning it assesses whether the variables tend to change in the same relative direction, even if not at a constant rate. It achieves this by converting raw data into ranks before performing the correlation calculation. This makes it a **non-parametric statistics** tool that is particularly effective for **ordinal data** or continuous data that contains significant outliers.

The primary advantage of **Spearman's rank correlation coefficient** is its resilience. Because it operates on the order of the values rather than their absolute magnitudes, it is not skewed by extreme observations in the same way Pearson's 'r' is. For instance, if you are analyzing the relationship between the rank of students in a math competition and their rank in a science competition, Spearman's rho is the ideal metric. It captures the essence of the association--whether high ranks in one subject correspond to high ranks in another--without requiring the intervals between ranks to be equal or the data to be normally distributed. This makes it a staple in social sciences where **Likert scales** and other ranked metrics are common.

In **Stata**, the **spearman** command is used to generate these coefficients. The output provides the rho value along with a **p-value** to test the null hypothesis that there is no **monotonic** association. Stata also handles missing values efficiently within this command, allowing for pairwise

comparisons. For researchers dealing with non-linear but consistent relationships--such as exponential growth or logarithmic decay--Spearman's rho often yields a more accurate representation of the **correlation** strength than a standard linear assessment would. It is a vital tool for any analyst looking to expand their **non-parametric statistics** toolkit.

Kendall's Tau: Precision in Small Samples and Tied Ranks

Kendall's rank correlation coefficient, or Kendall's tau, is another **non-parametric statistics** measure available in **Stata** that is similar in purpose to Spearman's rho. However, Kendall's tau takes a different mathematical approach by analyzing the number of **concordant** and **discordant** pairs within the dataset. A pair of observations is concordant if their ranks for both variables are in the same order and discordant if they are in the opposite order. This method provides a more direct probabilistic interpretation of the **correlation** than Spearman's rho, often making it more intuitive for certain types of behavioral research.

One of the specific strengths of **Kendall's rank correlation coefficient** is its performance with small sample sizes. While Spearman's rho is often preferred for larger datasets due to its computational simplicity, Kendall's tau is considered more robust and provides better estimates of the population **correlation** when the sample is limited. Furthermore, Kendall's tau-b, a specific variant calculated by Stata, includes an adjustment for "tied ranks"--instances where multiple observations have the same value. This adjustment is crucial in **categorical data** analysis where ties are frequent, ensuring that the resulting coefficient is not artificially inflated or deflated.

To execute this analysis in **Stata**, users utilize the **ktau** command. This command is particularly useful when the researcher suspects that the data distribution is far from normal and the sample size is too small to rely on the asymptotic properties of other tests. The **p-value** generated by **ktau** allows for rigorous hypothesis testing. By providing both tau-a (the basic version) and tau-b (the version adjusted for ties), Stata empowers the user to choose the most precise metric for their specific data structure, reinforcing the software's reputation for **statistical significance** and accuracy.

Biserial and Tetrachoric Correlations for Binary Variables

Standard **correlation** measures often fail when one or both variables are **binary variables** (dichotomous). To address this, Stata provides specialized coefficients: the **biserial correlation** and the **tetrachoric correlation**. The biserial correlation is designed to measure the relationship between a **continuous variable** and a binary variable. It is particularly useful in testing environments where one might correlate a continuous test score with a binary outcome like "pass" or "fail." Unlike the point-biserial correlation (which is a special case of Pearson's), the biserial correlation assumes the binary variable reflects an underlying normal distribution that has been

artificially dichotomized.

When both variables in the analysis are **binary variables**, the **tetrachoric correlation** becomes the appropriate tool. This coefficient estimates what the **Pearson correlation coefficient** would be if the two variables were continuous and normally distributed. It is extensively used in **psychometrics** and **item response theory** to understand the latent relationships between categorical responses. By using the **tetrachoric** command in **Stata**, researchers can uncover associations that would be missed by simpler **chi-squared** tests, providing a deeper look into the structure of their categorical data.

Choosing between these specialized measures requires a clear understanding of the data's nature. **Biserial correlation** and **tetrachoric correlation** are powerful because they allow the researcher to bridge the gap between categorical observations and continuous theory. In Stata, these commands are straightforward but require the user to be mindful of the underlying distributional assumptions. When applied correctly, they provide a nuanced view of **correlation** that goes beyond the capabilities of standard linear metrics, ensuring that binary data is treated with the same statistical rigor as continuous data.

Partial and Multiple Correlation: Controlling for Complexity

In many research scenarios, the relationship between two variables is obscured by the influence of a third factor. This is where **partial correlation** becomes essential. A partial **correlation** measures the degree of association between two variables while mathematically removing the effect of one or more additional variables. For example, a researcher might want to examine the correlation between exercise and heart health while controlling for the influence of age. By using **Stata** to calculate the partial correlation, the analyst can isolate the "pure" relationship between the primary variables of interest, free from the confounding effects of **covariates**.

Similarly, **multiple correlation** assessments allow researchers to understand how a single variable relates to a combination of other variables simultaneously. This is often represented by the coefficient of multiple determination (R-squared) in a **regression analysis**, but it can also be viewed as a **correlation** measure in its own right. It quantifies the maximum possible correlation between a dependent variable and a linear combination of independent predictors. Stata's robust regression suite and specific commands like **pcor** (for partial correlations) make it easy to disentangle these complex multivariate webs, providing a clearer picture of causal pathways and associative strengths.

The ability to calculate these coefficients is what elevates **Stata** from a basic calculator to a sophisticated analytical engine. **Partial correlation** and **multiple correlation** are indispensable for **model selection** and theory testing. They allow for a more disciplined approach to **data analysis**, where the researcher can account for the interconnected nature of real-world phenomena. By

mastering these multivariate correlation tools, Stata users can produce findings that are not only statistically **significant** but also contextually accurate and highly persuasive in a professional or academic setting.

Correlations in Stata: Pearson, Spearman, and Kendall

In the field of statistics, **correlation** serves as a quantitative measure of the strength and direction of the relationship between two distinct variables. The resulting correlation coefficient is standardized to range from -1 to 1. In this scale, -1 represents a perfect negative relationship (as one variable increases, the other decreases), 0 represents the total absence of a relationship, and 1 represents a perfect positive relationship (both variables increase together).

There are three primary methodologies utilized to measure correlation in modern **data analysis**:

Pearson Correlation: This is the standard approach for assessing the linear association between two **continuous variables**, such as the relationship between an individual's height and their weight.

Spearman Correlation: This **non-parametric statistics** measure is ideal for **ordinal data** or ranked variables, such as comparing the class rank of a student in mathematics against their rank in science.

Kendall's Correlation: This is an alternative to Spearman's method, frequently preferred when dealing with small sample sizes or datasets characterized by a high frequency of tied ranks.

The following technical tutorial provides a step-by-step guide on how to implement and interpret these three **correlation** types within the **Stata** environment.

Step 1: Loading the Reference Dataset

To demonstrate these procedures, we will utilize the built-in *auto* dataset, a classic **Stata** sample containing various automobile specifications. To initialize the data, enter the following command into your Stata Command window:

use <http://www.stata-press.com/data/r13/auto>

Once the data is loaded, it is good practice to perform a preliminary **descriptive statistics** overview. This allows you to understand the range and distribution of your variables. Execute the following command to see a summary table:

summarize

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
```

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

The output reveals that the dataset consists of 12 variables, including metrics like price, mileage (mpg), weight, and length, which are well-suited for **correlation** testing.

How to Find Pearson Correlation in Stata

To determine the **Pearson correlation coefficient** between the *weight* and *length* of the vehicles, you can use the **pwcorr** command. This provides the pairwise correlation between the specified variables:

```
pwcorr weight length
```

```
. pwcorr weight length
```

	weight	length
weight	1.0000	
length	0.9460	1.0000

To assess the **statistical significance** of this relationship, append the **sig** option to the command. This will display the **p-value** alongside the coefficient:

```
pwcorr weight length, sig
```

```
. pwcorr weigh length, sig
```

	weight	length
weight	1.0000	
length	0.9460 0.0000	1.0000

In this example, the resulting **p-value** is **0.000**. Because this value is well below the standard alpha threshold of 0.05, we conclude that the **correlation** between weight and length is statistically significant. For broader analysis involving multiple variables, you can list them consecutively:

```
pwcorr weight length displacement, sig
```

```
. pwcorr weight length displacement, sig
```

	weight	length	displacement
weight	1.0000		
length	0.9460 0.0000	1.0000	
displacement	0.8949 0.0000	0.8351 0.0000	1.0000

Interpreting the correlation matrix involves looking at the intersections of variables:

The **Pearson Correlation** between weight and length is 0.9460 ($p = 0.000$).

The correlation between weight and displacement is 0.8949 ($p = 0.000$).

The correlation between displacement and length is 0.8351 ($p = 0.000$).

How to Find Spearman Correlation in Stata

For relationships that may be **monotonic** rather than strictly linear, or when working with **ordinal data**, use the **spearman** command. Here, we analyze the relationship between *trunk* space and the repair record (*rep78*):

spearman trunk rep78**. spearman trunk rep78**

Number of obs = 69
 Spearman's rho = -0.2235

Test of Ho: trunk and rep78 are independent
 Prob > |t| = 0.0649

The interpretation of the **Spearman's rank correlation coefficient** output includes several key components:

Number of obs: This indicates the total pairwise observations used. Due to missing values in *rep78*, Stata used 69 observations.

Spearman's rho: The coefficient here is -0.2235, suggesting a weak negative **correlation**.

Prob > |t|: This **p-value** of 0.0649 suggests the relationship is not statistically significant at the 5% level.

To perform a multivariate Spearman analysis with **p-values** for each pair, use the **stats(rho p)** option:

spearman trunk rep78 gear_ratio, stats(rho p)

```
. spearman trunk rep78 gear_ratio, stats(rho p)
(obs=69)
```

Key
<i>rho</i>
<i>Sig. Level</i>

	trunk	rep78	gear_r~o
trunk	1.0000		
rep78	-0.2235 0.0649	1.0000	
gear_ratio	-0.5187 0.0000	0.4275 0.0002	1.0000

Key findings from this matrix include a significant negative **correlation** between trunk space and gear ratio (-0.5187 , $p = 0.0000$) and a significant positive correlation between gear ratio and repair records (0.4275 , $p = 0.0002$).

How to Find Kendall's Correlation in Stata

When your analysis requires a **non-parametric statistics** approach that is more robust to small samples or ties, **Kendall's rank correlation coefficient** is the preferred choice. In **Stata**, this is executed via the **ktau** command:

ktau trunk rep78

. ktau trunk rep78

```

Number of obs =      69
Kendall's tau-a =    -0.1424
Kendall's tau-b =    -0.1752
Kendall's score =   -334
  SE of score =    181.254   (corrected for ties)

Test of Ho: trunk and rep78 are independent
  Prob > |z| =      0.0662   (continuity corrected)

```

The interpretation of **Kendall's rank correlation coefficient** output is as follows:

Number of obs: Consistent with the Spearman test, 69 observations were utilized due to missing data.

Kendall's tau-b: The value of -0.1752 indicates a negative relationship. Tau-b is specifically used because it adjusts for tied ranks in the data.

Prob > |z|: The **p-value** of 0.0662 confirms that the **correlation** does not reach **statistical significance** at the 0.05 level.

For a comprehensive matrix of multiple variables using Kendall's method, apply the **stats(taub p)** option:

ktau trunk rep78 gear_ratio, stats(taub p)

Test of Ho: trunk and rep78 are independent
 Prob > |z| = **0.0662** (continuity corrected)

```
. ktau trunk rep78 gear_ratio, stats(taub p)
(obs=69)
```

Key	
<i>tau_b</i>	
<i>Sig. Level</i>	

	trunk	rep78	gear_r~o
trunk	1.0000		
rep78	-0.1752 0.0662	1.0000	
gear_ratio	-0.3753 0.0000	0.3206 0.0006	1.0000

The results show that while the trunk-to-repair record correlation remains non-significant, the trunk-to-gear ratio (-0.3753, $p = 0.0000$) and gear ratio-to-repair record (0.3206, $p = 0.0006$) associations are highly significant.

Summary and Strategic Considerations

In summary, the diverse array of **correlation** measures available in **Stata** provides researchers with the necessary flexibility to analyze relationships across various data types and distributions. From the **Pearson correlation coefficient** for linear, continuous data to **Spearman** and **Kendall** for ranked and non-parametric sets, Stata ensures that the mathematical tool matches the empirical reality of the dataset. Furthermore, specialized options for **binary variables** and multivariate controls allow for a highly granular exploration of **statistical significance**.

Choosing the correct **correlation** measure is not merely a technical requirement but a strategic one. Each coefficient carries its own set of assumptions regarding **linearity**, normality, and measurement scales. Misapplying a **parametric** test to non-parametric data can lead to misleading results, while ignoring the influence of confounding variables can result in incorrect causal inferences. By carefully considering the research question and the properties of the data, analysts can use Stata to produce insights that are both reliable and actionable.

Ultimately, **Stata's** power lies in its ability to present these complex calculations through simple,

reproducible commands. Whether you are conducting a preliminary data screen or a final, rigorous hypothesis test, the correlation tools discussed here form the bedrock of successful **data analysis**. As you continue your work in Stata, remember that the goal of correlation is to describe the "what" of a relationship, providing a crucial stepping stone toward understanding the "why" through more advanced **econometric** techniques.

ARABPSYCHOLOGY.COM