

Tutorial: Creating and Interpreting Correlation Matrices in Stata

In statistics, we're often interested in understanding the relationship between two variables. For example, we might want to understand the relationship between the number of hours a student studies and the exam score they receive. One way to quantify this relationship is to use the correlation matrix, which is a measure of the...

Authored by
Mohammed loot

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Tutorial: Creating and Interpreting Correlation Matrices in Stata* In statistics, we're often interested in understanding the relationship between two variables. For example, we might want to understand the relationship between the number of hours a student studies and the exam score they receive. One way to quantify this relationship is to use the correlation matrix, which is a measure of the.... PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13725>

Introduction to Correlation and Its Importance in Data Analysis

In the realm of [statistics](#) and data analysis, one of the most fundamental tasks is understanding the dynamic interplay between different variables. Researchers frequently seek to quantify the nature and strength of the relationship between two measurable attributes--for instance, assessing how a student's dedication, measured by hours of study, influences their eventual exam performance. This type of inquiry moves beyond mere observation, aiming to establish a quantifiable link that can inform predictions and theoretical models.

Before diving into the mechanics of generating these relationships in statistical software, it is crucial to appreciate the mathematical concept that underpins this analysis. Correlation provides a succinct summary of whether two variables tend to change together, and if so, in what direction. A strong understanding of this concept is essential for accurately interpreting the results produced by software like [Stata](#), ensuring that the visual and numerical output accurately reflects the underlying data behavior.

Understanding the Correlation Coefficient

The primary metric used to quantify the [linear association](#) between two variables is the [correlation coefficient](#). This value, typically represented by the Greek letter rho (ρ) for a population or 'r' for a sample, is a standardized measure designed to fall strictly within the range of **-1** and **1**. This standardized scale allows for easy comparison of relationship strengths across different datasets and variable types, regardless of their original units of measurement.

Interpreting the correlation coefficient requires careful attention to both its magnitude and its sign. The sign indicates the direction of the relationship, while the absolute value indicates the strength. A positive coefficient means that as one variable increases, the other tends to increase as well (a positive correlation). Conversely, a negative coefficient signifies an inverse relationship, where an increase in one variable corresponds to a decrease in the other (a negative correlation). The further away the correlation coefficient is from zero, the stronger the relationship between the two variables.

The values are interpreted as follows:

-1 indicates a perfectly negative linear correlation between two variables, meaning the data points fall perfectly along a downward-sloping line.

0 indicates no linear correlation between two variables, suggesting the variables are independent of each other in a linear sense.

1 indicates a perfectly positive linear correlation between two variables, where the data points fall perfectly along an upward-sloping line.

It is paramount to remember that the correlation coefficient measures only the strength of the **linear** relationship. Two variables may be strongly related in a non-linear fashion (e.g., quadratic or exponential), yet still yield a correlation coefficient close to zero. Therefore, correlation should always be examined in conjunction with visual aids, such as scatter plots, to ensure the detected linearity truly represents the data's behavior.

The Role of the Correlation Matrix

While analyzing the relationship between just two variables is straightforward, many real-world statistical models involve dozens of potential predictors. Assessing every pairwise combination individually would be cumbersome and inefficient. This is where the [correlation matrix](#) becomes an indispensable tool. A correlation matrix is essentially a square table designed to display the correlation coefficients for multiple variables simultaneously, providing an immediate, comprehensive overview of the interdependence within a dataset.

In a typical correlation matrix, the variables are listed both across the columns and down the rows. Each cell in the matrix represents the correlation coefficient (r) between the variable specified by that row and the variable specified by that column. By definition, the matrix is symmetrical (the correlation of A with B is the same as B with A), and the diagonal elements--where a variable is correlated with itself--are always exactly 1.0000. This structure allows analysts to quickly identify which pairs of variables exhibit strong collinearity, which is essential for tasks like regression modeling where highly correlated predictors can lead to unstable results.

In this tutorial, we explain how to create a correlation matrix using the powerful capabilities of [Stata](#). Stata offers robust, yet simple, commands that streamline the process of calculating and presenting complex correlation data. We will now proceed through the step-by-step process required to execute these commands and accurately interpret the resulting output.

Preparing Data and Generating the Basic Correlation Matrix in Stata

The fundamental command used within the Stata environment to produce a correlation matrix for selected variables or an entire dataset is **corr**. Before executing this command, however, we must ensure that the relevant dataset is loaded and accessible within the current Stata session. For demonstration purposes, we will utilize a widely accessible example dataset--the 1980 US census data--which is conveniently hosted by Stata Press.

To load this census data into Stata, execute the following command in the command window. This action downloads the file directly into memory, preparing it for immediate analysis:

use <http://www.stata-press.com/data/r13/census13>

Once the data is loaded, it is good practice to gain an initial understanding of the variables present and their basic descriptive statistics. The **summarize** command provides a quick yet informative overview of key metrics such as the mean, standard deviation, minimum, and maximum values for all variables in the loaded dataset. This helps confirm data integrity and variable types before proceeding to correlation analysis.

summarize

Executing the summary command yields a table detailing the characteristics of the nine distinct variables included in this particular census dataset, confirming the data structure we will analyze:

```
. use http://www.stata-press.com/data/r13/census13
(1980 Census data by state)
```

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
state	50	25.5	14.57738	1	50
brate	50	167.94	29.34552	125	286
pop	50	4518149	4715038	401851	2.37e+07
medage	50	29.54	1.693445	24.2	34.7
division	50	5.12	2.560612	1	9
region	50	2.66	1.061574	1	4
mrgrate	50	.0186789	.0257327	.0103731	.1955433
dvcrate	50	.0079769	.0031552	.0039954	.0236739
medagesq	50	875.422	99.87562	585.64	1204.09

We see that the dataset contains nine different variables. To generate the full correlation matrix encompassing every possible pairwise combination among these nine variables, we simply issue the **corr** command without specifying any subsequent variable names. This tells Stata to calculate and display the correlation for the entire loaded dataset.

corr

The resulting output is the complete, nine-by-nine correlation matrix:

```
. corr
(obs=50)
```

	state	brate	pop	medage	division	region	mrgrate	dvcrate
state	1.0000							
brate	0.0208	1.0000						
pop	-0.0540	-0.2830	1.0000					
medage	-0.0624	-0.8800	0.3294	1.0000				
division	-0.1345	0.6356	-0.1081	-0.5207	1.0000			
region	-0.1339	0.6086	-0.1515	-0.5292	0.9688	1.0000		
mrgrate	0.0509	0.0677	-0.1502	-0.0177	0.2280	0.2490	1.0000	
dvcrate	-0.0655	0.3508	-0.2064	-0.2229	0.5522	0.5682	0.7700	1.0000
medagesq	-0.0621	-0.8609	0.3324	0.9984	-0.5162	-0.5239	-0.0202	-0.2192

Analyzing Subsets and Interpreting Results

The numerical values presented in the matrix are the [Pearson Correlation Coefficients](#) (r) for the intersecting variables. Interpreting these values is straightforward based on the magnitude and sign. For example, examining the cell corresponding to the intersection of the variables *pop* (population) and *state* reveals a correlation coefficient of **-0.0540**. Because this value is extremely close to zero, it suggests a very weak, slightly negative linear relationship between the two variables. This indicates that these two variables are slightly negatively correlated.

Notice that a crucial feature of any correlation matrix is the diagonal, which runs from the top-left to the bottom-right corner. These diagonal values are uniformly **1.0000**. This is a fundamental property because any variable is perfectly and positively correlated with itself.

Often, researchers are not interested in the correlation among all variables, but rather a specific subset relevant to a focused hypothesis. Stata provides the flexibility to narrow down the analysis by simply listing the desired variable names after the **corr** command. This is highly useful for managing large datasets and focusing computational resources only on the most pertinent relationships.

You can create a correlation matrix for only a certain subset of variables in a dataset by specifying the variables after the **corr** command. For example, here is how to create a correlation matrix for just the variables *pop*, *medage*, and *region*:

```
corr pop medage region
```

This execution yields a simplified, three-by-three matrix, focusing solely on the specified variables:

```
. corr pop medage region
(obs=50)
```

	pop	medage	region
pop	1.0000		
medage	0.3294	1.0000	
region	-0.1515	-0.5292	1.0000

This targeted approach not only cleans up the output but also aids in preventing spurious interpretations that might arise from examining dozens of marginally correlated variables simultaneously. It ensures that the interpretation remains grounded in the theoretical framework guiding the analysis.

Advanced Correlation: Incorporating Statistical Significance

While the correlation coefficient indicates the magnitude and direction of the relationship, it does not inherently tell us whether that relationship is statistically meaningful, or whether the observed correlation is likely due to chance. To address this, Stata provides the **pwcorr** command, which calculates pairwise correlations and offers advanced options, most notably the ability to indicate [statistical significance](#) using an asterisk or "star" notation.

The **pwcorr** command generally produces results identical to **corr** for the correlation coefficients themselves, but it is often preferred when significance testing is required. By appending the **star()** option, we instruct Stata to mark coefficients that meet a specified significance threshold (alpha level). This is crucial for distinguishing robust findings from merely incidental observations in the data.

For example, the following code produces a correlation matrix for every variable in the census dataset and places a star next to the correlation coefficients that are statistically significant at the standard alpha level of 0.05 ($\alpha = 0.05$):

```
pwcorr, star(.05)
```

The resulting matrix now includes asterisks next to the coefficients where the null hypothesis (that the true population correlation is zero) can be rejected at the 5% significance level:

. pwcorr, star(.05)

	state	brate	pop	medage	division	region	mrgrate
state	1.0000						
brate	0.0208	1.0000					
pop	-0.0540	-0.2830*	1.0000				
medage	-0.0624	-0.8800*	0.3294*	1.0000			
division	-0.1345	0.6356*	-0.1081	-0.5207*	1.0000		
region	-0.1339	0.6086*	-0.1515	-0.5292*	0.9688*	1.0000	
mrgrate	0.0509	0.0677	-0.1502	-0.0177	0.2280	0.2490	1.0000
dvcrate	-0.0655	0.3508*	-0.2064	-0.2229	0.5522*	0.5682*	0.7700*
medagesq	-0.0621	-0.8609*	0.3324*	0.9984*	-0.5162*	-0.5239*	-0.0202
	dvcrate medagesq						
dvcrate	1.0000						
medagesq	-0.2192	1.0000					

Analyzing this output reveals which relationships are likely genuine effects in the population. Notice how several of the correlation coefficients in the table are statistically significant at $\alpha = 0.05$. We could set α to be any number we'd like, but common choices are .01, .05, and .10.

Controlling the Alpha Level and Conclusion

The choice of the alpha (α) level directly impacts the number of observed correlations deemed statistically significant. The alpha level represents the probability of committing a Type I error--incorrectly rejecting a true null hypothesis (a false positive). Consequently, lowering the value of α increases the stringency of the test, reducing the likelihood of observing a star next to a correlation coefficient.

In general, the lower we set the value of α , the fewer correlation coefficients will be statistically significant. For example, suppose we set the significance threshold to a more conservative $\alpha = 0.01$. This requires a much stronger correlation to pass the significance test:

pwcorr, star(.01)

As anticipated, comparing this output to the previous matrix confirms that fewer correlation coefficients now display a star, reflecting the stricter criteria imposed by the 0.01 alpha level:

. pwcorr, star(.01)

	state	brate	pop	medage	division	region	mrgrate
state	1.0000						
brate	0.0208	1.0000					
pop	-0.0540	-0.2830	1.0000				
medage	-0.0624	-0.8800*	0.3294	1.0000			
division	-0.1345	0.6356*	-0.1081	-0.5207*	1.0000		
region	-0.1339	0.6086*	-0.1515	-0.5292*	0.9688*	1.0000	
mrgrate	0.0509	0.0677	-0.1502	-0.0177	0.2280	0.2490	1.0000
dvcrate	-0.0655	0.3508	-0.2064	-0.2229	0.5522*	0.5682*	0.7700*
medagesq	-0.0621	-0.8609*	0.3324	0.9984*	-0.5162*	-0.5239*	-0.0202
	dvcrate medagesq						
dvcrate	1.0000						
medagesq	-0.2192	1.0000					

Notice how fewer correlation coefficients have a star next to them. Mastering these commands allows researchers to quickly diagnose linear relationships within their data, identify potential issues like multicollinearity, and ensure that subsequent, more complex modeling efforts are built upon a solid understanding of variable interdependence.