How to Generate a Correlation Matrix in Stata: A Step-by-Step Guide

Authored by

stats writer

December 28, 2025

Generating a correlation matrix is a fundamental step in exploring multivariate data analysis using statistical software like Stata. This process requires calculating the correlation coefficients across multiple pairs of variables simultaneously. Once calculated, the primary command in Stata, `corr`, is utilized to display these coefficients in a clean, tabular format. Furthermore, Stata offers excellent flexibility, allowing analysts to customize the output by selecting specific variables and altering the display format, including methods for flagging statistical significance.

In the field of statistics, much of our work focuses on quantifying the nature and strength of the association between two or more elements. Consider a simple example: examining the link between the amount of time a student spends studying and the final grade they achieve on an exam. Understanding these relationships, which analysts define using measures of association, is crucial for building predictive models and drawing insightful conclusions from complex datasets.

One standard way to quantify this relationship is to use the correlation coefficient, which is a standardized metric of the linear association between two variables. This coefficient is strictly bounded between -1 and 1, providing immediate insight into both the direction and the magnitude of the linear relationship. Interpreting this value is critical for initial data exploration:

-1 indicates a perfectly negative linear correlation, meaning that as one variable increases, the other decreases predictably.
0 indicates no observable linear correlation between the two variables.
1 represents a perfectly positive linear correlation, where both variables increase or decrease together in lockstep.

It is important to remember that the further the calculated correlation coefficient is from zero--that is, the closer it is to either -1 or 1--the stronger the linear relationship between the two variables. Values near zero suggest a weak or non-existent linear connection, though non-linear relationships may still be present.

While analyzing single pairs of variables is necessary, real-world data analysis often involves dozens of interdependent variables. In these complex scenarios, simply running individual correlation tests becomes highly inefficient. This is precisely where the correlation matrix proves its value. It is a fundamental tool--a powerful, square table designed to display all pairwise correlation coefficients among a defined set of variables, offering an immediate and consolidated overview of the entire correlation structure within the dataset.

In this tutorial, we will explain how to efficiently generate, customize, and interpret a correlation matrix using the statistical software package, Stata. We will begin by demonstrating the core command and then progress to methods for subsetting and evaluating statistical significance.

How to Use the Basic Correlation Command in Stata

The primary command used to produce a correlation matrix for a loaded dataset in Stata is simply `corr`. This command automatically calculates the correlation coefficient for every possible pairwise combination of numeric variables available in your active dataset.

To illustrate this functionality, we will use a standard sample dataset provided by StataCorp: the 1980 census data. We load this dataset into the Stata environment by typing the following command into the command box and pressing Enter:

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

Once the data is loaded, it is good practice to get an initial summary of the dataset to understand the variables we are working with. We can do this quickly using the `summarize` command:

summarize

Executing this command produces a descriptive statistical table summarizing the key properties of the variables in the dataset:

```
. 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

Generating the Full Correlation Matrix

As shown in the output above, the census dataset contains nine different variables. To generate a correlation matrix that encompasses every pairwise combination of these nine variables, we simply use the base command `corr` without specifying any variable names. This instructs Stata to run correlations on all applicable variables simultaneously.

Type the following into the command box:

corr

This action produces the comprehensive correlation matrix seen below:

```
. 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

Interpreting the Basic Correlation Matrix

The values presented in the body of the table represent the Pearson Correlation Coefficients for each intersection of variables. When reading the matrix, you look at the row and column intersection to find the correlation between those two specific variables. For example, by examining the intersection of the variable *pop* (population) and *state*, we find a coefficient of **-0.0540**. This indicates a very slight, almost negligible, negative linear correlation between these two variables.

It is crucial to observe the primary diagonal of the table. Notice that the correlation coefficient for every variable correlated with itself (e.g., *pop* with *pop*, *medage* with *medage*) is exactly **1.0000**. This is because a variable is perfectly correlated with itself. Additionally, the correlation matrix is symmetric; the coefficient for *A* correlated with *B* is the same as *B* correlated with *A*, meaning you only need to examine the coefficients above or below the diagonal.

Creating a Matrix for Specific Variables

In large datasets, analysts often only require the correlation coefficients for a small subset of the available variables. Stata allows you to generate a targeted correlation matrix by simply listing the variables of interest after the `corr` command. This is essential for focusing on specific relationships without the clutter of extraneous data.

For instance, if we only wanted to assess the correlations among *pop* (population), *medage*

(median age), and *region*, we would execute the command as follows:

corr pop medage region

This command produces a focused correlation matrix that includes only these three selected 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**

Displaying Statistical Significance

Beyond the simple measure of association, researchers are often concerned with whether the observed correlation is statistically significant--that is, whether the relationship is likely to exist in the underlying population, or if it occurred merely by random chance. To incorporate indicators of statistical significance into the matrix output, we must use the `pwcorr` command instead of `corr`.

The `pwcorr` command is functionally very similar to `corr`, but it includes additional options for handling missing data and, importantly, the ability to specify the `star()` option. The `star()` option allows Stata to place an asterisk (*) next to correlation coefficients that achieve statistical significance at a predefined alpha ($alpha$) level.

For example, the following command generates a correlation matrix for the entire census dataset and flags coefficients that are statistically significant at the commonly used $alpha = 0.05$ level:

pwcorr, star(.05)

This updated matrix now clearly highlights significant relationships:

```
. 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

Observe that several of the correlation coefficients are now marked with an asterisk, indicating that the relationships they represent are statistically significant at the 5% level. This feature helps analysts quickly identify relationships that warrant further investigation based on their probability of being true.

Adjusting the Alpha Level ($alpha$) for Significance

The choice of the alpha level ($alpha$) directly influences how many correlation coefficients are deemed statistically significant. While $alpha = 0.05$ is standard, analysts can set $alpha$ to any desired value, with common alternatives being 0.01 (indicating a higher standard of evidence) and 0.10 (a more lenient standard).

In general, the lower the value we set for $alpha$, the fewer correlation coefficients will pass the stringent threshold required for statistical significance. For instance, suppose we want to increase our confidence level and set the significance threshold at $alpha = 0.01$. We would adjust the `star()` argument accordingly:

pwcorr, star(.01)

This command yields a new matrix:

```
. 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

By comparing this output to the previous $alpha = 0.05$ matrix, it is noticeable that fewer correlation coefficients now feature a star. This demonstrates the direct trade-off between the stringency of the significance level and the number of relationships identified as statistically meaningful within the dataset.