correlate — Correlations of variables

Description

The `correlate` command displays the correlation matrix or covariance matrix for a group of variables. If `varlist` is not specified, the matrix is displayed for all variables in the dataset.

`pwcorr` displays all the pairwise correlation coefficients between the variables in `varlist` or, if `varlist` is not specified, all the variables in the dataset.

Quick start

Correlation matrix for variables v1, v2, and v3
```
correlate v1 v2 v3
```
As above, but display covariances instead of correlations
```
correlate v1 v2 v3, covariance
```
Pairwise correlation coefficients between v1, v2, and v3
```
pwcorr v1 v2 v3
```
Also print significance level of each correlation coefficient
```
pwcorr v1 v2 v3, sig
```
As above, but star correlation coefficients significant at the 5% level
```
pwcorr v1 v2 v3, sig star(.05)
```
As above, but use Bonferroni-adjusted significance levels
```
pwcorr v1 v2 v3, sig star(.05) bonferroni
```

Menu

`correlate`
Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Correlations and covariances

`pwcorr`
Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Pairwise correlations
Syntax

Display correlation matrix or covariance matrix

```
correlate [varlist] [if] [in] [weight] [ , correlate_options ]
```

Display all pairwise correlation coefficients

```
pwcorr [varlist] [if] [in] [weight] [ , pwcorr_options ]
```

**correlate_options**  Description

<table>
<thead>
<tr>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>means</td>
<td>display means, standard deviations, minimums, and maximums with matrix</td>
</tr>
<tr>
<td>noformat</td>
<td>ignore display format associated with variables</td>
</tr>
<tr>
<td>covariance</td>
<td>display covariances</td>
</tr>
<tr>
<td>wrap</td>
<td>allow wide matrices to wrap</td>
</tr>
</tbody>
</table>

**pwcorr_options**  Description

<table>
<thead>
<tr>
<th>Main</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>obs</td>
<td>print number of observations for each entry</td>
</tr>
<tr>
<td>sig</td>
<td>print significance level for each entry</td>
</tr>
<tr>
<td>listwise</td>
<td>use listwise deletion to handle missing values</td>
</tr>
<tr>
<td>casewise</td>
<td>synonym for listwise</td>
</tr>
<tr>
<td>print(#)</td>
<td>significance level for displaying coefficients</td>
</tr>
<tr>
<td>star(#)</td>
<td>significance level for displaying with a star</td>
</tr>
<tr>
<td>bonferroni</td>
<td>use Bonferroni-adjusted significance level</td>
</tr>
<tr>
<td>sidak</td>
<td>use Šidák-adjusted significance level</td>
</tr>
</tbody>
</table>

*varlist* may contain time-series operators; see [U] 11.4.4 Time-series varlists.

by and collect are allowed with correlate and pwcorr; see [U] 11.1.10 Prefix commands.

aweights and fweights are allowed; see [U] 11.1.6 weight.

Options for correlate

**means** displays summary statistics (means, standard deviations, minimums, and maximums) with the matrix.

**noformat** displays the summary statistics requested by the *means* option in *g* format, regardless of the display formats associated with the variables.

**covariance** displays the covariances rather than the correlation coefficients.

**wrap** requests that no action be taken on wide correlation matrices to make them readable. It prevents Stata from breaking wide matrices into pieces to enhance readability. You might want to specify this option if you are displaying results in a window wider than 80 characters. Then you may need to set *linesize* to however many characters you can display across a line; see [R] log.
Options for pwcorr

obs adds a line to each row of the matrix reporting the number of observations used to calculate the correlation coefficient.

sig adds a line to each row of the matrix reporting the significance level of each correlation coefficient.

listwise handles missing values through listwise deletion, meaning that the entire observation is omitted from the estimation sample if any of the variables in varlist is missing for that observation. By default, pwcorr handles missing values by pairwise deletion; all available observations are used to calculate each pairwise correlation without regard to whether variables outside that pair are missing.

correlate uses listwise deletion. Thus, listwise allows users of pwcorr to mimic correlate’s treatment of missing values while retaining access to pwcorr’s features.

casewise is a synonym for listwise.

print(#) specifies the significance level of correlation coefficients to be printed. Correlation coefficients with larger significance levels are left blank in the matrix. Typing pwcorr, print(.10) would list only correlation coefficients significant at the 10% level or better.

star(#) specifies the significance level of correlation coefficients to be starred. Typing pwcorr, star(.05) would star all correlation coefficients significant at the 5% level or better.

bonferroni makes the Bonferroni adjustment to calculated significance levels. This option affects printed significance levels and the print() and star() options. Thus, pwcorr, print(.05) bonferroni prints coefficients with Bonferroni-adjusted significance levels of 0.05 or less.

sidak makes the Šidák adjustment to calculated significance levels. This option affects printed significance levels and the print() and star() options. Thus, pwcorr, print(.05) sidak prints coefficients with Šidák-adjusted significance levels of 0.05 or less.

Remarks and examples

Remarks are presented under the following headings:

correlate
pwcorr
Video example

correlate

Typing correlate by itself produces a correlation matrix for all variables in the dataset. If you specify the varlist, a correlation matrix for just those variables is displayed.
Example 1

We have state data on demographic characteristics of the population. To obtain a correlation matrix, we type

```
. use https://www.stata-press.com/data/r17/census13
(1980 Census data by state)
. correlate
(obs=50)
```

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

Because we did not specify the `wrap` option, Stata did its best to make the result readable by breaking the table into two parts.

To obtain the correlations between `mrgrate`, `dvcrate`, and `medage`, we type

```
. correlate mrgrate dvcrate medage
(obs=50)
```

<table>
<thead>
<tr>
<th></th>
<th>mrgrate</th>
<th>dvcrate</th>
<th>medage</th>
</tr>
</thead>
<tbody>
<tr>
<td>mrgrate</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dvcrate</td>
<td>0.7700</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>medage</td>
<td>-0.0177</td>
<td>-0.2229</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Example 2

The `pop` variable in example 1 represents the total population of the state. Thus, to obtain population-weighted correlations among `mrgrate`, `dvcrate`, and `medage`, we type

```
. correlate mrgrate dvcrate medage [w=pop]
(analytic weights assumed)
(sum of wgt is 225,907,472)
(obs=50)
```

<table>
<thead>
<tr>
<th></th>
<th>mrgrate</th>
<th>dvcrate</th>
<th>medage</th>
</tr>
</thead>
<tbody>
<tr>
<td>mrgrate</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dvcrate</td>
<td>0.5854</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>medage</td>
<td>-0.1316</td>
<td>-0.2833</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

With the covariance option, `correlate` can be used to obtain covariance matrices, as well as correlation matrices, for both weighted and unweighted data.
Example 3

To obtain the matrix of covariances between `mrgrate`, `dvcrate`, and `medage`, we type `correlate mrgrate dvcrate medage, covariance`:

```
. correlate mrgrate dvcrate medage, covariance
(obs=50)

                     |  mrgrate  dvcrate  medage
--------------------+------------------------
    mrgrate         |     0.000662          
    dvcrate         |    0.000063 1.0e-05   
    medage          |   -0.000769 -0.01191 2.86775
```

We could have obtained the pop-weighted covariance matrix by typing `correlate mrgrate dvcrate medage [w=pop], covariance`.

\[ \text{pwcorr} \]

`correlate` calculates correlation coefficients by using casewise deletion; when you request correlations of variables \(x_1, x_2, \ldots, x_k\), any observation for which any of \(x_1, x_2, \ldots, x_k\) is missing is not used. Thus if \(x_3\) and \(x_4\) have no missing values, but \(x_2\) is missing for half the data, the correlation between \(x_3\) and \(x_4\) is calculated using only the half of the data for which \(x_2\) is not missing. Of course, you can obtain the correlation between \(x_3\) and \(x_4\) by using all the data by typing `correlate x3 x4`.

\( \text{pwcorr} \) makes obtaining such pairwise correlation coefficients easier.

Example 4

Using `auto.dta`, we investigate the correlation between several of the variables.

```
. use https://www.stata-press.com/data/r17/auto1
(Automobile models)
. pwcorr mpg price rep78 foreign, obs sig
```

```
                  |    mpg    price   rep78   foreign
------------------+------------------------
     mpg           |     1.0000           
                      |            74
     price          |  -0.4594   1.0000    
                      |          0.0000    74
     rep78          |   0.3739   0.0066    1.0000
                      |       0.0016  0.9574
                      |        69     69
     foreign        |   0.3613   0.0487   0.5922
                      |         0.0016  0.6802
                      |         74     74
```
Technical note

The `correlate` command will report the correlation matrix of the data, but there are occasions when you need the matrix stored as a Stata matrix so that you can further manipulate it. You can obtain the matrix by typing

```stata
.matrix accum R = varlist, noconstant deviations
.matrix R = corr(R)
```

The first line places the cross-product matrix of the data in matrix R. The second line converts that to a correlation matrix. Also see [P] `matrix define` and [P] `matrix accum`.

Video example

Pearson’s correlation coefficient in Stata

Stored results

correlate stores the following in `r()`:

Scalars
- `r(N)` number of observations
- `r(rho)` $\rho$ (first and second variables)
- `r(cov_12)` covariance (covariance only)
- `r(Var_1)` variance of first variable (covariance only)
- `r(Var_2)` variance of second variable (covariance only)
- `r(sum_w)` sum of weights

Matrices
- `r(C)` correlation or covariance matrix
**pwcorr** stores the following in \( r() \):

**Scalars**
- \( r(N) \): number of observations (first and second variables)
- \( r(rho) \): \( \rho \) (first and second variables)

**Matrices**
- \( r(C) \): pairwise correlation matrix
- \( r(sig) \): significance level of each correlation coefficient

**Methods and formulas**

For a discussion of correlation, see, for instance, Snedecor and Cochran (1989, 177–195); for an introductory explanation using Stata examples, see Acock (2018, 205–210).

According to Snedecor and Cochran (1989, 180), the term “co-relation” was first proposed by Galton (1888). The product-moment correlation coefficient is often called the Pearson product-moment correlation coefficient because Pearson (1896) and Pearson and Filon (1898) were partially responsible for popularizing its use. See Stigler (1986) for information on the history of correlation.

The estimate of the product-moment correlation coefficient, \( \hat{\rho} \), is

\[
\hat{\rho} = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}}
\]

where \( w_i \) are the weights, if specified, or \( w_i = 1 \) if weights are not specified. \( \bar{x} = (\sum w_i x_i) / (\sum w_i) \) is the mean of \( x \), and \( \bar{y} \) is similarly defined.

The unadjusted significance level is calculated by **pwcorr** as

\[
p = 2 \times ttail(n - 2, |\hat{\rho}| \sqrt{n - 2} / \sqrt{1 - \hat{\rho}^2})
\]

Let \( v \) be the number of variables specified so that \( k = v(v - 1)/2 \) correlation coefficients are to be estimated. If bonferroni is specified, the adjusted significance level is \( p' = \min(1, kp) \). If sidak is specified, \( p' = \min\{1, 1 - (1 - p)^k\} \). In both cases, see **Methods and formulas** in [R] oneway for a more complete description of the logic behind these adjustments.
Carlo Emilio Bonferroni (1892–1960) was born in Bergamo, Italy, in 1892. Before enrolling in the mathematics department at the University of Turin, he studied conducting and the piano at the Music Conservatory of Turin. Bonferroni, like many men of his generation, fought in World War I, during which he was part of The Engineer Corps of the Italian Army.

After the war, Bonferroni was appointed as an assistant professor at the Polytechnic University of Turin. He taught geometry, mechanics, and analysis. In 1923, he moved to Bari and began teaching at the Economics Institute, where he got to teach financial mathematics, a topic that was of great interest to him. In 1933, Bonferroni moved to Florence, where he was chair of the mathematics department until his death in 1960.

Bonferroni’s interests had a large breadth. He published on actuarial mathematics, probability, statistics, analysis, geometry, and mechanics. His work on probability inequalities has been applied to simultaneous statistical inference. However, the application of Bonferroni’s theory to the construction of confidence intervals is the work of Olive Jean Dunn.

Olive Jean Dunn (1915–2008) was born in the United States in 1915. She obtained her bachelor’s degree in 1936, her master’s degree in 1951, and her PhD in 1956, all from the University of California in Los Angeles (UCLA). After spending one year as an assistant professor at Iowa State College, she returned to UCLA to serve in the biostatistics and preventive medicine and health departments. She remained at UCLA for the rest of her career. Dunn died in 2008.

Dunn is best known for her application of Bonferroni’s inequalities to construct corrections to confidence intervals for multiple comparisons. Although the literature refers to it as the Bonferroni correction, it is Dunn who developed the application we use today.

Dunn is also well known for her textbooks Basic Statistics: A Primer for the Biomedical Sciences, written in 1977 with later editions coauthored with Virginia A. Clark, and Applied Statistics: An Analysis of Variance and Regression, which was also coauthored with Clark.

In 1968, Dunn became a Fellow of the American Statistical Association. She also was a fellow of the American Public Health Association and the American Association for the Advancement of Science. In 1974, she was awarded the honor of UCLA Woman of Science.
Florence Nightingale David (1909–1993) was born in Ivington, England, to parents who were friends with Florence Nightingale, David’s namesake. She began her studies in statistics under the direction of Karl Pearson at University College London and continued her studies under the direction of Jerzy Neyman. After receiving her doctorate in statistics in 1938, David became a senior statistician for various departments within the British military. She developed statistical models to forecast the toll on life and infrastructure that would occur if a large city were bombed. In 1938, she also published her book *Tables of the Correlation Coefficient*, dealing with the distributions of correlation coefficients. After the war, she returned to University College London, serving as a lecturer until her promotion to professor in 1962. In 1967, David joined the University of California–Riverside, eventually becoming chair of the Department of Statistics. One of her most well-known works is the book *Games, Gods and Gambling: The Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era*, a history of statistics. David published over 100 papers on topics including combinatorics, symmetric functions, the history of statistics, and applications of statistics, including ecological diversity. She published under the name F. N. David to avoid revealing her gender in a male-dominated profession.

Karl Pearson (1857–1936) studied mathematics at Cambridge. He was professor of applied mathematics (1884–1911) and eugenics (1911–1933) at University College London. His publications include literary, historical, philosophical, and religious topics. Statistics became his main interest in the early 1890s after he learned about its application to biological problems. His work centered on distribution theory, the method of moments, correlation, and regression. Pearson introduced the $\chi^2$ test and the terms coefficient of variation, contingency table, heteroskedastic, histogram, homoskedastic, kurtosis, mode, random sampling, random walk, skewness, standard deviation, and truncation. Despite many strong qualities, he also fell into prolonged disagreements with others, most notably, William Bateson and R. A. Fisher.

Zbyněk Šidák (1933–1999) was a notable Czech statistician and probabilist. He worked on Markov chains, rank tests, multivariate distribution theory and multiple-comparison methods, and he served as the chief editor of *Applications of Mathematics*.

References


Also see

[R] esize — Effect size based on mean comparison
[R] estat vce — Display covariance matrix estimates
[R] icc — Intraclass correlation coefficients
[R] pcorr — Partial and semipartial correlation coefficients
[R] spearman — Spearman’s and Kendall’s correlations
[R] summarize — Summary statistics
[R] tetrachoric — Tetrachoric correlations for binary variables