Title

correlate — Correlations (covariances) of variables or coefficients

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
Options	
`means`	display means, standard deviations, minimums, and maximums with matrix
`noformat`	ignore display format associated with variables
`covariance`	display covariances
`wrap`	allow wide matrices to wrap

pwcorr_options	Description
Main	
`obs`	print number of observations for each entry
`sig`	print significance level for each entry
`listwise`	use listwise deletion to handle missing values
`casewise`	synonym for `listwise`
`print(#)`	significance level for displaying coefficients
`star(#)`	significance level for displaying with a star
`bonferroni`	use Bonferroni-adjusted significance level
`sidak`	use Šidák-adjusted significance level

varlist may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.
`by` is allowed with `correlate` and `pwcorr`; see **[D] by**.
`aweights` and `fweights` are allowed; see **[U] 11.1.6 weight**.

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

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. Also see the `estat vce` command in [R] **estat vce**.

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

Options for correlate

Options

`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

Main

`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

stata.com

Remarks are presented under the following headings:

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 http://www.stata-press.com/data/r13/census13
(1980 Census data by state)
. correlate
(obs=50)
```

	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

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)
```

	mrgrate	dvcrate	medage
mrgrate	1.0000		
dvcrate	0.7700	1.0000	
medage	-0.0177	-0.2229	1.0000

◁

▷ 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   2.2591e+08)
(obs=50)
```

	mrgrate	dvcrate	medage
mrgrate	1.0000		
dvcrate	0.5854	1.0000	
medage	-0.1316	-0.2833	1.0000

◁

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	.000662		
dvcrate	.000063	1.0e-05	
medage	-.000769	-.001191	2.86775

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

◁

pwcorr

`correlate` calculates correlation coefficients by using casewise deletion; when you request correlations of variables x_1, x_2, ..., x_k, any observation for which any of x_1, x_2, ..., 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` x_3 x_4.

`pwcorr` makes obtaining such pairwise correlation coefficients easier.

▷ Example 4

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

```
. use http://www.stata-press.com/data/r13/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	74		
rep78	0.3739	0.0066	1.0000	
	0.0016	0.9574		
	69	69	69	
foreign	0.3613	0.0487	0.5922	1.0000
	0.0016	0.6802	0.0000	
	74	74	69	74

```
. pwcorr mpg price headroom rear_seat trunk rep78 foreign, print(.05) star(.01)
```

	mpg	price	headroom	rear_s~t	trunk	rep78	foreign
mpg	1.0000						
price	-0.4594*	1.0000					
headroom	-0.4220*		1.0000				
rear_seat	-0.5213*	0.4194*	0.5238*	1.0000			
trunk	-0.5703*	0.3143*	0.6620*	0.6480*	1.0000		
rep78	0.3739*					1.0000	
foreign	0.3613*		-0.2939	-0.2409	-0.3594*	0.5922*	1.0000

```
. pwcorr mpg price headroom rear_seat trunk rep78 foreign, print(.05) bon
```

	mpg	price	headroom	rear_s~t	trunk	rep78	foreign
mpg	1.0000						
price	-0.4594	1.0000					
headroom	-0.4220		1.0000				
rear_seat	-0.5213	0.4194	0.5238	1.0000			
trunk	-0.5703		0.6620	0.6480	1.0000		
rep78	0.3739					1.0000	
foreign	0.3613				-0.3594	0.5922	1.0000

◁

❑ 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

```
. 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)`	ρ (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)

Matrices

`r(C)`	correlation or covariance matrix

`pwcorr` will leave in its wake only the results of the last call that it makes internally to `correlate` for the correlation between the last variable and itself. Only rarely is this feature useful.

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 (2014, 200–206).

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, ρ, is

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

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

The unadjusted significance level is calculated by `pwcorr` as

$$p = 2 * \texttt{ttail}(n-2, |\widehat{\rho}|\sqrt{n-2}/\sqrt{1-\widehat{\rho}^2})$$

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, $p' = \min\left\{1, 1-(1-p)^k\right\}$. 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) studied in Turin and taught there and in Bari and Florence. He published on actuarial mathematics, probability, statistics, analysis, geometry, and mechanics. His work on probability inequalities has been applied to simultaneous statistical inference, although the method known as Bonferroni adjustment usually relies only on an inequality established earlier by Boole.

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-squared 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

Acock, A. C. 2014. *A Gentle Introduction to Stata*. 4th ed. College Station, TX: Stata Press.

Dewey, M. E., and E. Seneta. 2001. Carlo Emilio Bonferroni. In *Statisticians of the Centuries*, ed. C. C. Heyde and E. Seneta, 411–414. New York: Springer.

Eisenhart, C. 1974. Pearson, Karl. In Vol. 10 of *Dictionary of Scientific Biography*, ed. C. C. Gillispie, 447–473. New York: Charles Scribner's Sons.

Galton, F. 1888. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London* 45: 135–145.

Gleason, J. R. 1996. sg51: Inference about correlations using the Fisher z-transform. *Stata Technical Bulletin* 32: 13–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 121–128. College Station, TX: Stata Press.

Goldstein, R. 1996. sg52: Testing dependent correlation coefficients. *Stata Technical Bulletin* 32: 18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 128–129. College Station, TX: Stata Press.

Pearson, K. 1896. Mathematical contributions to the theory of evolution—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London, Series A* 187: 253–318.

Pearson, K., and L. N. G. Filon. 1898. Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London, Series A* 191: 229–311.

Porter, T. M. 2004. *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton, NJ: Princeton University Press.

Rodgers, J. L., and W. A. Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *American Statistician* 42: 59–66.

Rovine, M. J., and A. von Eye. 1997. A 14th way to look at the correlation coefficient: Correlation as the proportion of matches. *American Statistician* 51: 42–46.

Seed, P. T. 2001. sg159: Confidence intervals for correlations. *Stata Technical Bulletin* 59: 27–28. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 267–269. College Station, TX: Stata Press.

Seidler, J., J. Vondráček, and I. Saxl. 2000. The life and work of Zbyněk Šidák (1933–1999). *Applications of Mathematics* 45: 321–336.

Snedecor, G. W., and W. G. Cochran. 1989. *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press.

Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press.

Verardi, V., and C. Dehon. 2010. Multivariate outlier detection in Stata. *Stata Journal* 10: 259–266.

Weber, S. 2010. bacon: An effective way to detect outliers in multivariate data using Stata (and Mata). *Stata Journal* 10: 331–338.

Wolfe, F. 1997. sg64: pwcorrs: Enhanced correlation display. *Stata Technical Bulletin* 35: 22–25. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 163–167. College Station, TX: Stata Press.

——. 1999. sg64.1: Update to pwcorrs. *Stata Technical Bulletin* 49: 17. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, p. 159. College Station, TX: Stata Press.

Also see

[R] **esize** — Effect size based on mean comparison

[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