<div style="border:1px solid black; padding:10px;">

pcorr — Partial and semipartial correlation coefficients

</div>

Description

pcorr displays the partial and semipartial correlation coefficients of a specified variable with each variable in a varlist after removing the effects of all other variables in the varlist. The squared correlations and corresponding significance are also reported.

Quick start

Partial and semipartial correlations of v1 with v2, v3, and v4

    pcorr v1 v2 v3 v4

Same as above, but for each level of categorical variable catvar

    by catvar: pcorr v1 v2 v3 v4

Partial and semipartial correlations of v5 with v6, v7, and v8, and a one-period lag of v5 using [tsset](#) data

    pcorr v5 L.v5 v6 v7 v8

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Partial correlations

Syntax

pcorr *varname varlist* [*if*] [*in*] [*weight*]

varlist may contain factor variables; see [U] 11.4.3 Factor variables.
varname and *varlist* may contain time-series operators; see [U] 11.4.4 Time-series varlists.
by and collect are allowed; see [U] 11.1.10 Prefix commands.
aweights and fweights are allowed; see [U] 11.1.6 weight.

Remarks and examples

Assume that y is determined by $x_1, x_2, \ldots, x_k$. The partial correlation between y and x_1 is an attempt to estimate the correlation that would be observed between y and x_1 if the other x's did not vary. The semipartial correlation, also called part correlation, between y and x_1 is an attempt to estimate the correlation that would be observed between y and x_1 after the effects of all other x's are removed from x_1 but not from y.

Both squared correlations estimate the proportion of the variance of y that is explained by each predictor. The squared semipartial correlation between y and x_1 represents the proportion of variance in y that is explained by x_1 only. This squared correlation can also be interpreted as the decrease in the model's R^2 value that results from removing x_1 from the full model. Thus, one could use the squared semipartial correlations as criteria for model selection. The squared partial correlation between y and x_1 represents the proportion of variance in y not associated with any other x's that is explained by x_1. Thus, the squared partial correlation gives an estimate of how much of the variance of y not explained by the other x's is explained by x_1.

▷ Example 1

Using our automobile dataset (described in [U] **1.2.2 Example datasets**), we can obtain the simple correlations between price, mpg, weight, and foreign from correlate (see [R] **correlate**):

```
. use https://www.stata-press.com/data/r19/auto
(1978 automobile data)

. correlate price mpg weight foreign
(obs=74)
```

	price	mpg	weight	foreign
price	1.0000			
mpg	-0.4686	1.0000		
weight	0.5386	-0.8072	1.0000	
foreign	0.0487	0.3934	-0.5928	1.0000

Although correlate gave us the full correlation matrix, our interest is in just the first column. We find, for instance, that the higher the mpg, the lower the price. We obtain the partial and semipartial correlation coefficients by using pcorr:

```
. pcorr price mpg weight foreign
(obs=74)
```

Partial and semipartial correlations of price with

Variable	Partial corr.	Semipartial corr.	Partial corr.^2	Semipartial corr.^2	Significance value
mpg	0.0352	0.0249	0.0012	0.0006	0.7693
weight	0.5488	0.4644	0.3012	0.2157	0.0000
foreign	0.5402	0.4541	0.2918	0.2062	0.0000

We now find that the partial and semipartial correlations of price with mpg are near 0. In the simple correlations, we found that price and foreign were virtually uncorrelated. In the partial and semipartial correlations, we find that price and foreign are positively correlated. The nonsignificance of mpg tells us that the amount in which R^2 decreases by removing mpg from the model is not significant. We find that removing either weight or foreign results in a significant drop in the R^2 of the model.

◁

❑ Technical note

Use caution when interpreting the above results. As we said at the outset, the partial and semipartial correlation coefficients are an *attempt* to estimate the correlation that would be observed if the effects of all other variables were taken out of both y and x or only x. pcorr makes it too easy to ignore the fact that we are fitting a model. In the example above, the model is

$$\text{price} = \beta_0 + \beta_1 \text{mpg} + \beta_2 \text{weight} + \beta_3 \text{foreign} + \epsilon$$

which is, in all honesty, a rather silly model. Even if we accept the implied economic assumptions of the model—that consumers value mpg, weight, and foreign—do we really believe that consumers place equal value on every extra 1,000 pounds of weight? That is, have we correctly parameterized the model? If we have not, then the estimated partial and semipartial correlation coefficients may not represent what they claim to represent. Partial and semipartial correlation coefficients are a reasonable way to summarize data if we are convinced that the underlying model is reasonable. We should not, however, pretend that there is no underlying model and that these correlation coefficients are unaffected by the assumptions and parameterization.

❑

Stored results

pcorr stores the following in r():

Scalars
    r(N)   number of observations
    r(df)   degrees of freedom

Matrices
    r(p_corr)   partial correlation coefficient vector
    r(sp_corr)   semipartial correlation coefficient vector

Methods and formulas

Results are obtained by fitting a linear regression of *varname* on *varlist*; see [R] **regress**. The partial correlation coefficient between *varname* and each variable in *varlist* is then calculated as

$$\frac{t}{\sqrt{t^2 + n - k}}$$

(Greene 2018, 39), where t is the t statistic, n is the number of observations, and k is the number of independent variables, including the constant but excluding any dropped variables.

The semipartial correlation coefficient between *varname* and each variable in *varlist* is calculated as

$$\text{sign}(t)\sqrt{\frac{t^2(1 - R^2)}{n - k}}$$

(Cohen et al. 2003, 89), where R^2 is the model R^2 value, and t, n, and k are as described above.

The significance is given by $2\Pr(t_{n-k} > |t|)$, where t_{n-k} follows a Student's t distribution with $n-k$ degrees of freedom.

Acknowledgment

The addition of semipartial correlation coefficients to pcorr is based on the pcorr2 command by Richard Williams of the Department of Sociology at the University of Notre Dame.

References

Cohen, J., P. Cohen, S. G. West, and L. S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. Hillsdale, NJ: Erlbaum.

Greene, W. H. 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

Also see