Partial correlation

From Wikipedia, the free encyclopedia

In probability theory and statistics, **partial correlation** measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

Contents

Formal definition

Formally, the partial correlation between X and Y given a set of n controlling variables $\mathbf{Z} = \{Z_1, Z_2, ..., Z_n\}$, written $\rho_{XY \cdot \mathbf{Z}}$, is the correlation between the residuals R_X and R_Y resulting from the linear regression of X with $\mathbf{Z}$ and of Y with $\mathbf{Z}$, respectively. The first-order partial correlation (i.e. when n=1) is the difference between a correlation and the product of the removable correlations divided by the product of the coefficients of alienation of the removable correlations. The coefficient of alienation, and its relation with joint variance through correlation are available in Guilford (1973, pp. 344–345).[1]

Computation

Using linear regression

A simple way to compute the sample partial correlation for some data is to solve the two associated linear regression problems, get the residuals, and calculate the correlation between the residuals. Let X and Y be, as above, random variables taking real values, and let $\mathbf{Z}$ be the n-dimensional vector-valued random variable. If we write x_i, y_i and $\mathbf{z}_i$ to denote the ith of N i.i.d. samples of some joint probability distribution over real random variables X, Y and $\mathbf{Z}$, solving the linear regression problem amounts to finding n-dimensional coefficient vectors $\mathbf{w}_X^*$ and $\mathbf{w}_Y^*$ such that

$$\mathbf{w}_X^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$
$$\mathbf{w}_Y^* = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\}$$

with N being the number of samples and $\langle \mathbf{v}, \mathbf{w} \rangle$ the scalar product between the vectors $\mathbf{v}$ and $\mathbf{w}$. Note that in some formulations the regression includes a constant term, so the matrix $\mathbf{Z}$ would have an additional column of ones.

The residuals are then

$$r_{X,i} = x_i - \langle \mathbf{w}_X^*, \mathbf{z}_i \rangle$$
$$r_{Y,i} = y_i - \langle \mathbf{w}_Y^*, \mathbf{z}_i \rangle$$

and the sample **partial** correlation is then given by the usual formula for sample correlation , but between these new *derived* values.

$$\hat{\rho}_{XY \cdot \mathbf{Z}} = \frac{N \sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \sum_{i=1}^{N} r_{X,i}^2 - \left(\sum_{i=1}^{N} r_{X,i} \right)^2} \sqrt{N \sum_{i=1}^{N} r_{Y,i}^2 - \left(\sum_{i=1}^{N} r_{Y,i} \right)^2}}.$$

Using recursive formula

It can be computationally expensive to solve the linear regression problems. Actually, the *n*th-order partial correlation (i.e., with $|\mathbf{Z}| = n$) can be easily computed from three $(n - 1)$th-order partial correlations. The zeroth-order partial correlation $\rho_{XY\cdot\emptyset}$ is defined to be the regular correlation coefficient ρ_{XY}.

It holds, for any $Z_0 \in \mathbf{Z}$:

$$\rho_{XY\cdot\mathbf{Z}} = \frac{\rho_{XY\cdot\mathbf{Z}\setminus\{Z_0\}} - \rho_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}\rho_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}{\sqrt{1-\rho^2_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}\sqrt{1-\rho^2_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}}.$$

Naïvely implementing this computation as a recursive algorithm yields an exponential time complexity. However, this computation has the overlapping subproblems property, such that using dynamic programming or simply caching the results of the recursive calls yields a complexity of $\mathcal{O}(n^3)$.

Note in the case where Z is a single variable, this reduces to:

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho^2_{XZ}}\sqrt{1-\rho^2_{ZY}}}.$$

Using matrix inversion

In $\mathcal{O}(n^3)$ time, another approach allows *all* partial correlations to be computed between any two variables X_i and X_j of a set **V** of cardinality *n*, given all others, i.e., $\mathbf{V} \setminus \{X_i, X_j\}$, if the correlation matrix (or alternatively covariance matrix) $\mathbf{\Omega} = (\omega_{ij})$, where $\omega_{ij} = \rho_{X_iX_j}$, is positive definite and therefore invertible. If we define $\mathbf{P} = \mathbf{\Omega}^{-1}$, we have:

$$\rho_{X_iX_j\cdot\mathbf{V}\setminus\{X_i,X_j\}} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}.$$

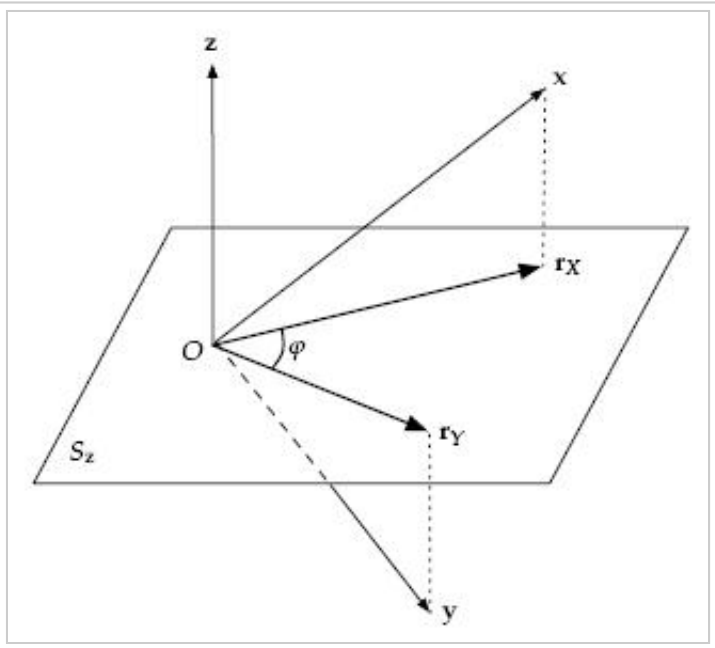
Interpretation

Geometrical

Let three variables X, Y, Z (where Z is the "control" or "extra variable") be chosen from a joint probability distribution over n variables $\mathbf{V}$. Further let $\mathbf{v}_i$, $1 \le i \le N$, be N n-dimensional i.i.d. samples taken from the joint probability distribution over $\mathbf{V}$. We then consider the N-dimensional vectors $\mathbf{x}$ (formed by the successive values of X over the samples), $\mathbf{y}$ (formed by the values of Y) and $\mathbf{z}$ (formed by the values of Z).

It can be shown that the residuals R_X coming from the linear regression of X on Z, if also considered as an N-dimensional vector $\mathbf{r}_X$, have a zero scalar product with the vector $\mathbf{z}$ generated by Z. This means that the residuals vector lies on an $(N–1)$-dimensional hyperplane $S_{\mathbf{z}}$ that is perpendicular to $\mathbf{z}$.

The same also applies to the residuals R_Y generating a vector $\mathbf{r}_Y$. The desired partial correlation is then the cosine of the angle φ between the projections $\mathbf{r}_X$ and $\mathbf{r}_Y$ of $\mathbf{x}$ and $\mathbf{y}$, respectively, onto the hyperplane perpendicular to $\mathbf{z}$.[2]:ch. 7



Geometrical interpretation of partial correlation for the case of N=3 samples and thus a 2-dimensional hyperplane

As conditional independence test

With the assumption that all involved variables are multivariate Gaussian, the partial correlation $\rho_{XY\cdot\mathbf{Z}}$ is zero if and only if X is conditionally independent from Y given $\mathbf{Z}$.[3] This property does not hold in the general case.

To test if a sample partial correlation $\hat{\rho}_{XY\cdot\mathbf{Z}}$ vanishes, Fisher's *z-transform of the partial correlation* can be used:

$$z(\hat{\rho}_{XY\cdot\mathbf{Z}}) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_{XY\cdot\mathbf{Z}}}{1-\hat{\rho}_{XY\cdot\mathbf{Z}}}\right).$$

The null hypothesis is $H_0 : \hat{\rho}_{XY\cdot\mathbf{Z}} = 0$, to be tested against the two-tail alternative $H_A : \hat{\rho}_{XY\cdot\mathbf{Z}} \neq 0$. We reject H_0 with significance level α if:

$$\sqrt{N - |\mathbf{Z}| - 3}\cdot|z(\hat{\rho}_{XY\cdot\mathbf{Z}})| > \Phi^{-1}(1-\alpha/2),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a Gaussian distribution with zero mean and unit standard deviation, and N is the sample size. Note that this z-transform is approximate and that the actual distribution of the sample (partial) correlation coefficient is not straightforward. However, an exact t-test based on a combination of the partial regression coefficient, the partial correlation coefficient and the partial variances is available.[4]

The distribution of the sample partial correlation was described by Fisher.[5]

Semipartial correlation (part correlation)

The semipartial (or part) correlation statistic is similar to the partial correlation statistic. Both compare variations of two variables after certain factors are controlled for, but to calculate the semipartial correlation one holds the third variable constant for either X or Y but not both, whereas for the partial correlation one holds the third variable constant for both.[6] The semipartial correlation compares the unique variation of one variable (having removed variation associated with the Z variable(s)), with the unfiltered variation of the other, while the partial correlation compares the unique variation of one variable to the unique variation of the other.

The semipartial (or part) correlation can be viewed as more practically relevant "because it is scaled to (i.e., relative to) the total variability in the dependent (response) variable." [7] Conversely, it is less theoretically useful because it is less precise about the role of the unique contribution of the independent variable.

The absolute value of the semipartial correlation of X with Y is always less than or equal to that of the partial correlation of X with Y. The reason is this: Suppose the correlation of X with Z has been removed from X, giving the residual vector r_x . In computing the semipartial correlation, Y still contains both unique variance and variance due to its association with Z. But r_x , being uncorrelated with Z, can only explain some of the unique part of the variance of Y and not the part related to Z. In contrast, with the partial correlation, only r_y (the part of the variance of Y that is unrelated to Z) is to be explained, so there is less variance of the type that r_x cannot explain.

Use in time series analysis

In time series analysis, the partial autocorrelation function (sometimes "partial correlation function") of a time series is defined, for lag h, as

$$\phi(h) = \rho_{X_0 X_h \cdot \{X_1,\ldots,X_{h-1}\}}.$$

This function is used to determine the appropriate lag length for an autoregression.

See also

- Linear regression
- Conditional independence
- Multiple correlation

References

1. Guilford J. P., Fruchter B. (1973). *Fundamental statistics in psychology and education*. Tokyo: McGraw-Hill Kogakusha, LTD.
2. Rummel, R. J. (1976). "Understanding Correlation".
3. Baba, Kunihiro; Ritei Shibata; Masaaki Sibuya (2004). "Partial correlation and conditional correlation as measures of conditional independence". *Australian and New Zealand Journal of Statistics* **46** (4): 657–664. doi:10.1111/j.1467-842X.2004.00360.x.
4. Kendall MG, Stuart A. (1973) *The Advanced Theory of Statistics*, Volume 2 (3rd Edition), ISBN 0-85264-215-6, Section 27.22
5. Fisher, R.A. (1924). "The distribution of the partial correlation coefficient". *Metron* **3** (3–4): 329–332.
6. http://luna.cas.usf.edu/~mbrannic/files/regression/Partial.html. Missing or empty `|title=` (help)
7. StatSoft, Inc. (2010). "Semi-Partial (or Part) Correlation" (http://www.statsoft.com/textbook/statistics-glossary/s/?button=0), Electronic Statistics Textbook. Tulsa, OK: StatSoft, accessed January 15, 2011.

External links

- Prokhorov, A.V. (2001), "Partial correlation coefficient", in Hazewinkel, Michiel, *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- What is a partial correlation? (http://www.psychwiki.com/wiki/What_is_a_partial_correlation%3F)
- Mathematical formulae in the "Description" section of the IMSL Numerical Library PCORR routine (http://www.roguewave.com/Portals/0/products/imsl-numerical-libraries/fortran-library/docs/7.0/stat/stat.htm)
- A three-variable example (http://faculty.vassar.edu/lowry/ch3a.html)

Wikiversity has learning materials about ***Partial correlation***

Categories: Covariance and correlation | Time series analysis

---