Multiple and partial regression and correlation

Partial r^2, contribution and fraction [a]

Daniel Borcard
Université de Montréal
Département de sciences biologiques
January 2002

The definitions of the three following notions are often confusing:

1. Contribution of a variable x_j to the explanation of the variation of a dependent variable y.
2. Fraction [a] in variation partitioning.
3. Partial r^2 (partial determination coefficient) between an x_j and a y variable.

This document assumes that the bases of simple and multiple regression, as well as the principle of the variation partitioning in regression (see Legendre & Legendre (1998), p. 528 sq.) are known to the reader.

Let us first define the three notions.

Definitions

1. **Contribution** of a variable x_j to the explanation of the variation of a dependent variable y;

This expression is used by Scherrer (1984) in the framework of the computation of the coefficient of multiple determination in multiple regression (p. 699-700). Note that Scherrer inadequately calls "*coefficient de corrélation multiple*" (multiple correlation coefficient) the coefficient of multiple determination (R^2). The multiple correlation coefficient (R) is the square root of the coefficient of multiple determination.

The coefficient of multiple determination measures the proportion of the variance of a dependent variable y explained by a set of explanatory variables x_{p-1}. It can be computed as

$$R^2 = \sum_{j=1}^{k} a_j^{'} r_{yx_j} \qquad \text{(eq. 1)}$$

where a'_j is the standardized regression coefficient of the j-th explanatory variable and r_{yx_j} is the simple correlation coefficient (Pearson r) between y and x_j [1].

---

[1] In Scherrer's notation: $R^2 = \sum_{j=1}^{p-1} a_j^{'} r_{jp}$, where the p-th variable is the dependent variable.

In this context, Scherrer calls the quantity $a'_j r_{yx_j}$ the "contribution" of the j-th variable to the explanation of the variance of y. The sum of the contributions of all explanatory variables x_k gives R^2. A contribution can be positive or negative.

2. **Fraction [a]** in variation partitioning

This fraction measures the proportion of the variance of y explained by the explanatory variable x_1 (for example) when the other explanatory variables (x_2, x_3...) are held constant **with respect to x_1 only** (and not with respect to y).

Thus, one obtains fraction [a] by examining the r^2 obtained by regressing y on the residuals of a regression of x_1 on x_2, x_3...

Note: one can compute fraction [a] for several explanatory variables simultaneously, but for the sake of simplicity we illustrate here the case where one seeks the fraction with respect to a single explanatory variable only.

3. **Partial r^2** (coefficient of partial determination) between an x_j and a y variable.

Note: when only two variables are involved, one generally uses the lowercase for r or r^2. Uppercase R and R^2 are used for the coefficients of multiple correlation and determination respectively.

First, the **partial *r*** measures the mutual relationship between two variables y and x when other variables (x_1, x_2, x_3...) are held constant **with respect to the two variables involved y and x_j** (contrary to the previous case)

The partial correlation coefficient is very useful in multiple regression, where it "...allows to directly estimate the proportion of unexplained variation of y that becomes explained with the addition of variable x_j. [to the model]" (translated from Scherrer, p.702).

For an explanatory variable x_1 and a variable x_2 held constant, the partial correlation coefficient is computed as follows (Scherrer, eq. 18-50, p. 704):

$$r_{y,x_1|x_2} = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{(1-r^2_{y,x_2})(1-r^2_{x_1,x_2})}} \qquad \text{(eq. 2)}$$

The **partial r^2** is the square of the partial r above. It measures the proportion of the variance of the residuals of y with respect to x_2 that is explained by the residuals of x_1 with respect to x_2. It can thus also be obtained by examining the r^2 of a regression of the residuals of y with respect to x_2 on the residuals of x_1 with respect to x_2.

Let us now examine the properties of these three measures in two different situations, each implying a dependent variable y and two explanatory variables x_1 and x_2.

Example 1. Intercorrelated (i.e., collinear) explanatory variables (most general case)

Data:

y	x_1	x_2
4.000	1.000	8.000
2.000	1.000	7.000
3.000	1.000	7.000
4.000	1.000	9.000
5.000	1.000	5.000
5.000	1.000	4.000
7.000	2.000	3.000
5.000	2.000	6.000
4.000	2.000	7.000
9.000	2.000	2.000
7.000	2.000	3.000
6.000	2.000	2.000

Simple linear correlation matrix:

	x_1	x_2
y	0.677	-0.824
x_1		-0.612

Partial linear correlation matrix:

	x_1	x_2
y	0.385	-0.704
x_1		-0.131

Regression coefficients:

	Raw coeff.	Standardized coeff.
Intercept	6.300	0.000
x_1	1.019	0.276
x_2	-0.523	-0.655

The coefficient of **multiple determination** of the regression can be computed using equation 1 (see above):

$R^2 = (0.276 \times 0.677) + (-0.655 \times -0.824) = 0.187 + 0.540 = 0.727$

The intermediate calculations yield the following informations:

- **contribution** of x_1 to the explanation of the variance of $y = 0.187$
- **contribution** of x_2 to the explanation of the variance of $y = 0.540$

Partial r^2 of y and x_1, holding x_2 constant with respect to y and x_1 (equation 2):

$$r_{y,x_1|x_2} = \frac{0.677-(-0.824\times-0.612)}{\sqrt{[1-(-0.824)^2][1-(-0.612)^2]}} = \frac{0.1727}{\sqrt{0.3210\times0.6255}} = \frac{0.1727}{0.4481} = 0.385$$

$r^2_{y,\,x1|x2} = 0.385^2 = 0.148$

Partial r^2 of y and x_2, holding x_1 constant with respect to y and x_2:

$$r_{y,x_2|x_1} = \frac{-0.824-(0.677\times -0.612)}{\sqrt{[1-(0.677)^2][1-(-0.612)^2]}} = \frac{-0.4097}{\sqrt{0.5417\times 0.6255}} = \frac{-0.4097}{0.5821} = -0.704$$

$r^2_{y,\,x1|x2} = -0.704^2 = 0.496$

Fraction [a]: proportion of the variance of y explained by the explanatory variable x_1 when x_2 is held constant with respect to x_1 only (and not with respect to y). This is the r^2 obtained by regressing y on the residuals of a regression of x_1 on x_2 (details omitted):

[a] = 0.0476

Fraction [c]: proportion of the variance of y explained by the explanatory variable x_2 when x_1 is held constant with respect to x_2 only (and not with respect to y). This is the r^2 obtained by regressing y on the residuals of a regression of x_2 on x_1 (details omitted):

[c] = 0.268

Fraction [b]: R^2 of the multiple regression of y on x_1 and x_2 – [a] – [c] = 0.727 – 0.048 – 0.268 = 0.411

Adding fractions [a], [b], [c] et[d] gives:

0.048 + 0.411 + 0.268 + (1–0.727) = 0.048 + 0.411 + 0.268 + 0.273 = 1.000

Note: using these results, one can also calculate the partial r^2 of y on x_1, controlling for x_2; this partial r^2 can indeed be construed as [a]/[a]+[d]:

$r^2_{y,\,x1|x2}$ = 0.048/(0.048+0.273) = 0.148

The same calculation could be done for the partial r^2 of y on x_2.

Example 2. Explanatory variables orthogonal (linearly independent) with respect to one another

Variables x_1 and x_2 can represent, for instance, two classification criteria describing an experimental design with two orthogonal factors. In that case, the analysis of variance can be computed by multiple regression, as in this example.

Data:

y	x_1	x_2
4.000	1.000	1.000
2.000	1.000	1.000
3.000	1.000	1.000
4.000	1.000	2.000
5.000	1.000	2.000
5.000	1.000	2.000
7.000	2.000	1.000
5.000	2.000	1.000
4.000	2.000	1.000
9.000	2.000	2.000
7.000	2.000	2.000
6.000	2.000	2.000

Simple linear correlation matrix:

	x_1	x_2
y	0.677	0.496
x_1		0.000

Partial linear correlation matrix:

	x_1	x_2
y	0.780	0.674
x_1		-0.526

Regression coefficients:

	Raw coeff.	Standardized coeff.
Intercept	-1.417	0.000
x_1	2.500	0.677
x_2	1.833	0.496

Coefficient of **multiple determination** (equation 1):

$R^2 = (0.677 \times 0.677) + (0.496 \times 0.496) = 0.458 + 0.246 = 0.704$

Therefore:

- **contribution** of x_1 to the explanation of the variance of $y = 0.458$

- **contribution** of x_2 to the explanation of the variance of $y = 0.246$

Partial r^2 of y and x_1, holding x_2 constant with respect to y and x_1 (equation 2):

$$r_{y,x_1|x_2} = \frac{0.677 - 0.496 \times 0}{\sqrt{[1-0.496^2)(1-0^2)}} = \frac{0.677}{\sqrt{0.754 \times 1}} = \frac{0.677}{0.868} = 0.780$$

$r^2_{y,\,x1|x2} = 0.780^2 = 0.608$

Partial r^2 of y and x_2, holding x_1 constant with respect to y and x_2:

$$r_{y,x_1|x_2} = \frac{0.496 - 0.677 \times 0}{\sqrt{[1-0.677^2)(1-0^2)}} = \frac{0.496}{\sqrt{0.542 \times 1}} = \frac{0.496}{0.736} = 0.674$$

$r^2_{y,\, x1|x2} = 0.674^2 = 0.454$

Fraction [a]: proportion of the variance of y explained by the explanatory variable x_1 when x_2 is held constant with respect to x_1 only (and not with respect to y). In this example, x_1 and x_2 are orthogonal (linearly independent), so that fraction [a] is directly equal to the r^2 obtained by regressing y on x_1 (details omitted):

[a] = 0.458

Fraction [c]: proportion of the variance of y explained by the explanatory variable x_2 when x_1 is held constant with respect to x_2 only (and not with respect to y). In this example, x_1 and x_2 are orthogonal (linearly independent), so that fraction [c] is directly equal to the r^2 obtained by regressing y on x_2 (details omitted):

[c] = 0.246

Fraction [b]: R^2 of the multiple regression of y on x_1 and x_2 – [a] – [c] =

= 0.704 – 0.458 – 0.246 = 0.000

The orthogonality of the explanatory variables x_1 and x_2 translates into a fraction [b] equal to 0.

Adding fractions [a], [b], [c] et[d] gives:

0.458 + 0.000 + 0.246 + (1– 0.704) = 0.458 + 0.000 + 0.246 + 0.296 = 1.000

In this case (orthogonality of x_1 and x_2), fractions [a] and [c] are equal to the contributions of variables x_1 and x_2 to the explanation of the variance of y.

Here again one can verify that equation $r^2_{y,\, x1|x2}$ = [a]/[a]+[d] gives the partial r^2 obtained above:

$r^2_{y,\, x1|x2} = 0.458/(0.458+0.296) = 0.607 \approx 0.608$

Comments (and summary of the computations)

1. Contribution (*sensu* Scherrer) of an explanatory variable is equal to fraction [a] of the explained variation (in the sense of a variation partitioning) **in one case only**: when all the explanatory variables are orthogonal to each other (linearly independent, uncorrelated). In that case, fraction [b] of the variation partitioning is equal to zero, since each explanatory variable explains a completely different fraction of the variance of y.

The total R^2 of the multiple regression (coefficient of multiple determination) is then computed as follows:

- either by summing the $a'_j r_{yx_j}$

- or by summing fractions [a] and [c] of the partitioning (since [b] equals zero!).

2. In the general case, i.e., when the explanatory variables are more or less intercorrelated (collinear), each one explains a fraction of the variation of y, but these fractions overlap more or less. Each explanatory variable "does a part of the job" of the other(s), since they are partly intercorrelated. The result is a non-zero fraction [b]. In general, fraction [b] is positive, and "nibbles" thus a part of fractions [a] and [c]. Therefore, [a] and [c] are smaller than their partial contributions which comprise fraction [a] or [c] **plus** a part of fraction [b].

In that case, the total R^2 of the multiple regression (coefficient of multiple determination) is computed as follows:

- either by summing the $a_j^{'} r_{yx_j}$
- or by summing fractions [a], [c] **and [b]** (since the latter does not equal zero).

Note: negative fractions [b] are sometimes observed! This happens when two explanatory variables have strong and opposite effects of the dependent variable, while being strongly intercorrelated. In that case, fractions [a] and [c] are larger than their partial contributions! A detailed explanation of this pattern is given by Legendre & Legendre (1998, p. 533).

Computation

To obtain, for example, fraction [a] of a regression of y explained by x_1 and x_2, you proceed as follows:

Step 1: compute a regression of x_1 explained by x_2 and keep the residuals. By doing this you have removed the effects of the other explanatory variable (x_2) from the explanatory variable of interest (x_1).

Step 2: compute a regression of y explained by the residuals obtained above. The r^2 of this second regression is equal to fraction [a)]. Hence, you have explained y with the part of x_1 that has no relationship with x_2.

If you are not interested in obtaining the fitted values, but only in the values of the fractions, you can perform the whole partitioning without resorting to partial regression. The steps are the following:

1. Regression of y on x_1 yields [a] + [b].
2. Regression of y on x_2 yields [b] + [c].
3. Multiple regression of y on x_1 and x_2 yields [a] + [b] + [c].
4. [a] is obtained by subtracting the result of step 2 from that of step 3.
5. [b] is obtained by subtracting the result of step 3 from that of step 1.
6. [d] is obtained by subtracting from the value 1.0 the result of step 3.

3. As explained above, the partial r^2 measures the (square of the) mutual relationship between two variables when other variables are held constant with respect to **both** variables involved. Remember that model I regression quantifies the effect of an explanatory variable on a dependent variable, while correlation measures their mutual relationship. This is also true in the partial case. There are two possible ways of computing it. The steps below show the computation of a partial r^2 between y and an x_1 variable, removing the effect of a variable x_2 as in the case of fraction [a], while emphasising the differences between the two methods:

• Computation using residuals

1. Compute a regression of y explained by x_2 and keep the residuals. The effect of x_2 has been removed from y.

2. Compute a regression of x_1 explained by x_2 and keep the residuals. The effect of x_2 has been removed from x_1.

3a. First variant: run a regression of the residuals of step 1 explained by the residuals of step 2 above.

3b. Second variant: compute the linear correlation between the residuals of step 1 and the residuals of step 2 above. This is another way of computing the partial correlation coefficient.

Contrary to the example given above for fraction [a], this operation completely evacuates the influence of x_2, from y as well as from x_1. Consequently, the r^2 obtained is the partial r^2 of y and x_1, controlling for x_2! The standardized slope (standardized regression coefficient) of this simple linear regression is the partial correlation coefficient between y and x_1.

Note that, if x_1 and x_2 are orthogonal, step 2 is useless, since x_2 explains nothing of x_1. Therefore, the residuals are equal to (centred) x_1.

• Computation through the fractions of variation

The partial r^2 can also be computed as [a]/[a]+[d]. Examination of this equation shows that one compares:

- in the numerator, the fraction of the variance of y explained only by x_1;
- in the denominator, this same fraction **plus** the unexplained variance, but **not** fractions [b] and [c] corresponding to the effect of variable x_2.

4. Finally, I remind the reader that fraction [b] **has nothing to do with an interaction in ANOVA**. In ANOVA, **an interaction measures the effect of an explanatory variable (a factor) on the influence of the other explanatory variable(s) on the dependent variable**. An interaction can have a non-zero value when the two explanatory variables are orthogonal, which is the situation where fraction [b] is equal to zero.

Acknowledgements

Many thanks to Pierre Legendre for critical reading and advice on this text, and help in the English translation.

Literature cited

Legendre, P., & L. Legendre. 1998. Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Scherrer, B. 1984. Biostatistique. Gaëtan Morin, Chicoutimi.