Partial regression and variation partitioning

Pierre Legendre
Département de sciences biologiques
Université de Montréal
http://www.NumericalEcology.com/

Outline of the presentation

1. Definition and objectives of partial regression.
2. Partial regression: two calculation methods
3. Statistics in partial regression
4. Variation partitioning
5. R software
6. References

1. Partial regression: definition and objectives

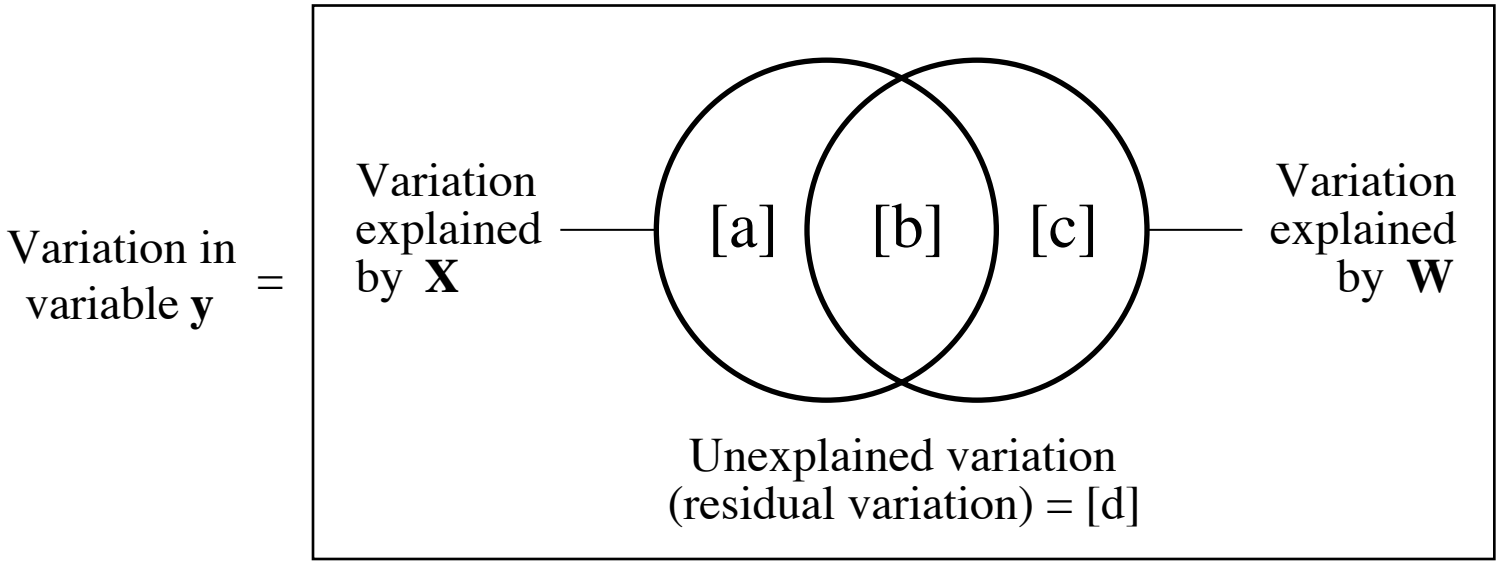
Partial linear regression is a statistical method that computes a linear model of the relationship $\mathbf{y} \sim \mathbf{X}$ while controlling for the linear effects of a second explanatory matrix $\mathbf{W}$:

$$(\mathbf{y} \sim \mathbf{X}) \mid \mathbf{W}$$

where $\mathbf{y}$ is the response variable, $\mathbf{X}$ is the matrix of explanatory variables and $\mathbf{W}$ is the matrix of covariables.

$\mathbf{X}$ and $\mathbf{W}$ may each contain a single variable.

Objectives of partial linear regression



- Estimate how much of the variation of **y** can be attributed –
  - exclusively to **X** once the effect of **W** has been controlled for,
  - exclusively to **W** once the effect of **X** has been controlled for.
- Estimate the vectors of fitted values corresponding to the exclusive effect of **X**, or to that of **W**.
- Partition the variation of **y** between **X** and **W** (see section 4).

Note *(from the multiple regression presentation)* –

The regression coefficients obtained in a multiple regression are *partial regression coefficients*.

Hence, to assess the unique contribution of each explanatory variable in a multiple regression model, it is not necessary to carry out partial regression analysis. The *partial regression coefficients* provide that information.

=> Standardizing the variables in matrix **X** makes them dimensionally homogeneous. The coefficients of multiple regression of standardized variables are called *standard partial regression coefficients*. They are directly comparable to one another because all **X** variables are dimensionless.[1]

[1] If **y** is not standardized, standard partial regression coefficients have the physical dimension of **y**. If **y** is also standardized before the calculation, standard partial regression coefficients are dimensionless. Unstandardized partial regression coefficients have the following physical dimension: dimension(**y**)/dimension(**x**).

2. Partial regression: two calculation methods

1. Compute residuals of **y** on **W**: $\mathbf{y}_{res|W} = \mathbf{y} - \mathbf{W}[\mathbf{W}'\mathbf{W}]^{-1}\,\mathbf{W}'\,\mathbf{y}$

   Compute residuals of **X** on **W**: $\mathbf{X}_{res|W} = \mathbf{X} - \mathbf{W}[\mathbf{W}'\mathbf{W}]^{-1}\,\mathbf{W}'\,\mathbf{X}$

2a. Regress $\mathbf{y}_{res|W}$ on $\mathbf{X}_{res|W}$ to obtain the *partial* R^2.

2b. Regress **y** on $\mathbf{X}_{res|W}$ to obtain the *semipartial* R^2.

Reminders about OLS regression –

1. The regression coefficients (vector **b**) are found by solving the eq.

$$\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{y}$$

The vector of fitted values is found by solving: $\mathbf{y}_{fit} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{y}$

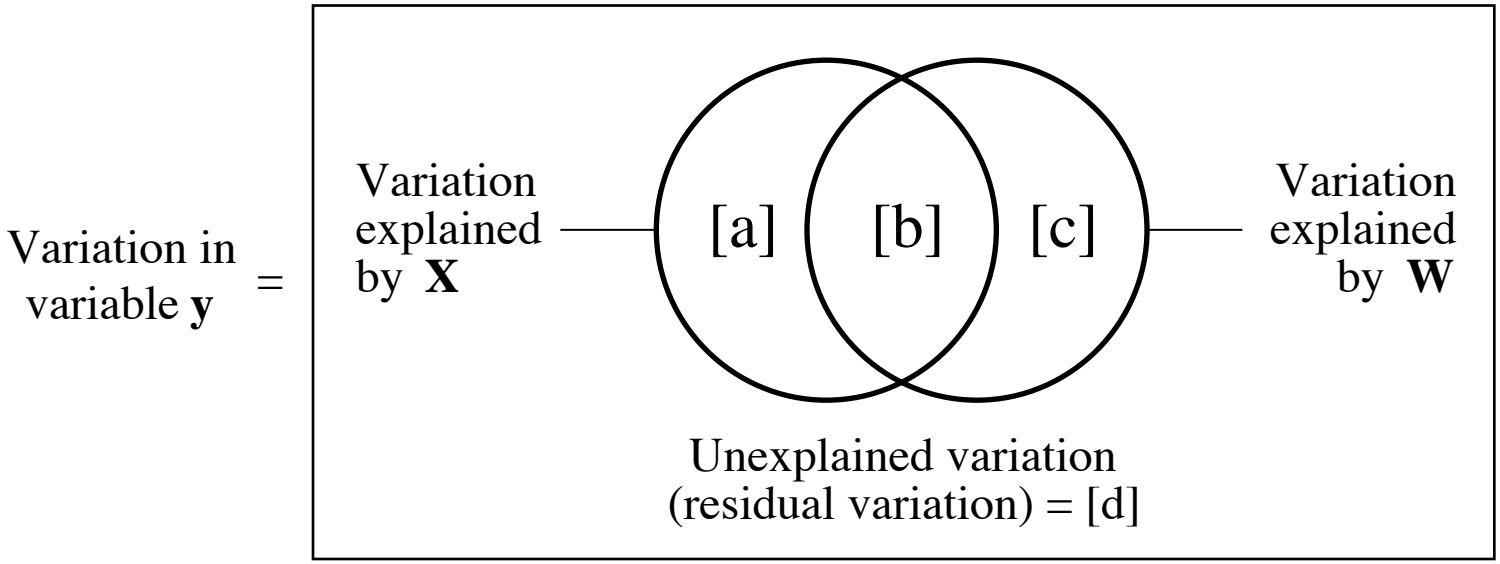
The vector of residual is found by solving: $\mathbf{y}_{res} = \mathbf{y} - \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{y}$

2. A column of '1' must be added to the matrix of regressors to estimate the intercept.

Comparison of the fitted values $\hat{\mathbf{y}}$ of methods 2a and 2b –

- The fitted values of regression method 2a are centred on 0; thus they have a mean of 0. The mean of the fitted values of method 2b is a constant that may differ from 0.

- The two methods produce fitted values with the same variance.



The explanatory data sets are designated by **X** and **W** in this figure (2 explanatory data sets).

Vegan's plot.varpart() function labels the matrices of explanatory variables as $\{\mathbf{X}_1, \mathbf{X}_2\}$ when there are two, $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ when there are three, and $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4\}$ when there are four.

3. Statistics in partial regression (**y** ~ **X** | **W**)

For test of fraction [a]:
$SS(\mathbf{y}) = [a+b+c+d]$
$SS(\mathbf{y}_{fit}) = [a]$
$SS(\mathbf{y}_{res}) = [d]$
$SS(\mathbf{y}_{res|\mathbf{W}}) = [a+d]$

$$R^2_{\mathbf{y}_{res|\mathbf{W}}|\mathbf{X}_{res|\mathbf{W}}} = \frac{SS(\mathbf{y}_{fit})}{SS(\mathbf{y}_{res|\mathbf{W}})}$$

Partial R^2 = [a]/[a+d]

$$R^2_{\mathbf{y}|\mathbf{X}_{res|\mathbf{W}}} = \frac{SS(\mathbf{y}_{fit})}{SS(\mathbf{y})}$$

Semipartial R^2
= [a]/[a+b+c+d]

$$F = \frac{SS(\mathbf{y}_{fit})/m}{SS(\mathbf{y}_{res})/(n-m-q-1)}$$

- Parametric test of F-stat. with d.f. m and $(n–m–q–1)$
- Permutation test of F-stat: See the course on canonical analysis

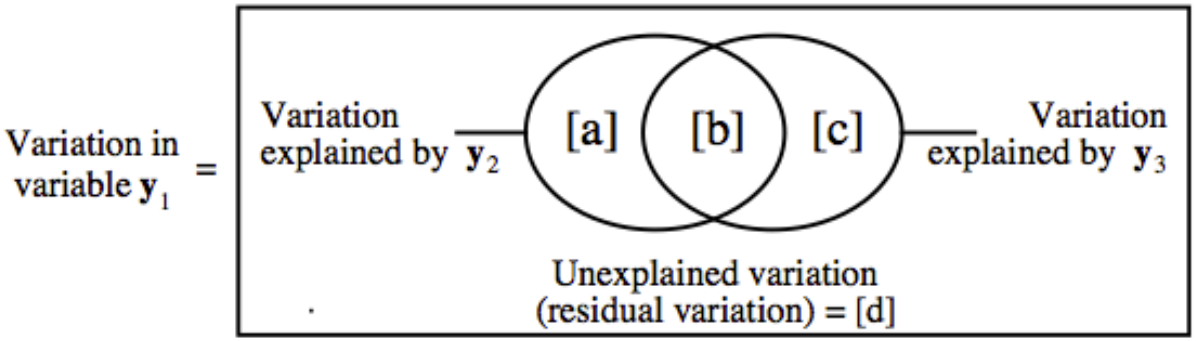
→ p-value

R^2_{adj} is obtained from the variation partitioning table

where n = number of objects (e.g. sites), m = rank $\mathbf{X}_{cent}$ and q = rank of $\mathbf{W}_{cent}$

Multiple, partial and semipartial R^2

Three variables only, $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$, are considered in this example. In the following Venn diagram, the rectangle represents the total sum of squares of variable $\mathbf{y}_1$:



In the multiple regression of $\mathbf{y}_1$ on $\mathbf{y}_2$ and $\mathbf{y}_3$, $\hat{y}_1 = b_0 + b_2 y_2 + b_3 y_3$ (this is an application of eq. 10.15), the coefficient of multiple determination, which is the square of the coefficient of multiple correlation, is:

$$R^2_{1.23} = \frac{[a+b+c]}{[a+b+c+d]} \qquad \text{with} \quad F = \frac{[a+b+c]/2}{[d]/(n-3)}$$

The partial correlation of $\mathbf{y}_1$ with $\mathbf{y}_2$ while controlling for the effect of $\mathbf{y}_3$ is:

$$r_{12.3} = \sqrt{\frac{[a]}{[a+d]}} \qquad \text{with} \quad F = \frac{[a]/1}{[d]/(n-3)}$$

The semipartial correlation of $\mathbf{y}_1$ with $\mathbf{y}_2$ in the presence of $\mathbf{y}_3$ is:

$$r_{1(2.3)} = \sqrt{\frac{[a]}{[a+b+c+d]}} \qquad \text{with} \quad F = \frac{[a]/1}{[d]/(n-3)}$$

The coefficients of partial and semipartial correlation receive the same sign as the corresponding coefficient of partial regression.

Excerpt from:
Legendre & Legendre 2012, Box 4.1.

4. Variation partitioning

Variation partitioning consists in apportioning the variation[1] of a response variable **y**, or a response data matrix **Y**, among two or more explanatory data sets

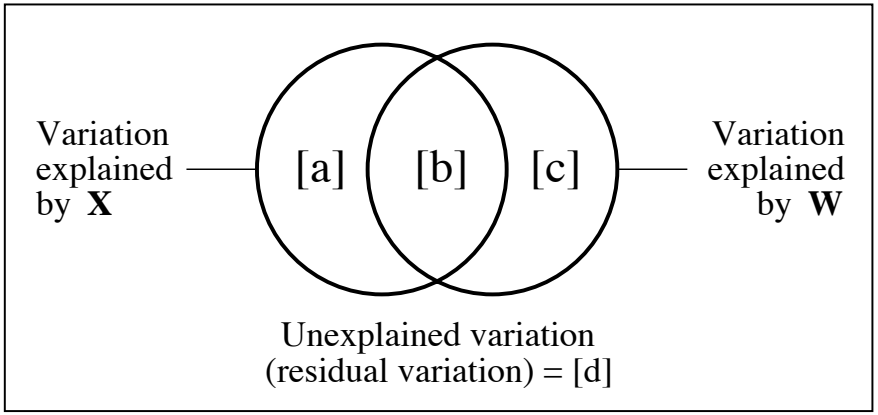
This method was proposed by Borcard et al. (1992).

Peres-Neto et al. (2006) showed that variation partitioning involving matrices of random explanatory variables must be based on adjusted R-squares.

[1] The term *variation*, a looser term than variance, is used because one is partitioning the total sum of squared deviations of **y** from its mean (total SS). There is no need to divide the total sum-of-squares (SS) of **y** by its degrees of freedom to obtain the variance.

Compute the fractions of variation

The objective is to compute the values of fractions [a], [b] and [c].



These values are computed from adjusted *R*-squares (R^2_{adj}, Peres-Neto et al. 2006) of three multiple regressions (Borcard et al. 2018):

lm(**y** ~ cbind(**X**,**W**)) => R2.adj = [a+b+c]
lm(**y** ~ **X**) => R2.adj = [a+b]
lm(**y** ~ **W**) => R2.adj = [b+c]

[b] = [a+b] + [b+c] – [a+b+c] # Fraction explained jointly by **X** and **W**
[a] = [a+b] – [b] *or* [a] = [a+b+c] – [b+c]
[c] = [b+c] – [b] *or* [c] = [a+b+c] – [a+b]
[d] = 1 – [a+b+c]

Data collected at 20 sites of the Thau lagoon, France, in 1988. Response data: Ma = log of bacterial abundances. (Table 10.6 of Legendre & Legendre 2012.)

	MA	NH4	PhaeoA	Prod.	X	Y	X^2
Site1	10.003	0.307	0.184	0.274	-8.75	3.7	76.5625
Site2	9.999	0.207	0.212	0.213	-6.75	2.7	45.5625
Site3	9.636	0.140	0.229	0.134	-5.75	1.7	33.0625
Site4	8.331	1.371	0.287	0.177	-5.75	3.7	33.0625
Site5	8.929	1.447	0.242	0.091	-3.75	2.7	14.0625
Site6	8.839	0.668	0.531	0.272	-2.75	3.7	7.5625
Site7	7.784	0.300	0.948	0.460	-1.75	0.7	3.0625
Site8	8.023	0.329	1.389	0.253	-0.75	-0.3	0.5625
Site9	8.294	0.207	0.765	0.235	0.25	-1.3	0.0625
Site10	7.883	0.223	0.737	0.362	0.25	0.7	0.0625
Site11	9.741	0.788	0.454	0.824	0.25	2.7	0.0625
Site12	8.657	1.112	0.395	0.419	1.25	1.7	1.5625
Site13	8.117	1.273	0.247	0.398	3.25	-4.3	10.5625
Site14	8.117	0.956	0.449	0.172	3.25	-2.3	10.5625
Site15	8.487	0.708	0.457	0.141	3.25	-1.3	10.5625
Site16	7.955	0.637	0.386	0.360	4.25	-5.3	18.0625
Site17	10.545	0.519	0.481	0.261	4.25	-4.3	18.0625
Site18	9.687	0.247	0.468	0.450	4.25	-2.3	18.0625
Site19	8.700	1.664	0.321	0.287	5.25	-0.3	27.5625
Site20	10.240	0.182	0.380	0.510	6.25	-2.3	39.0625

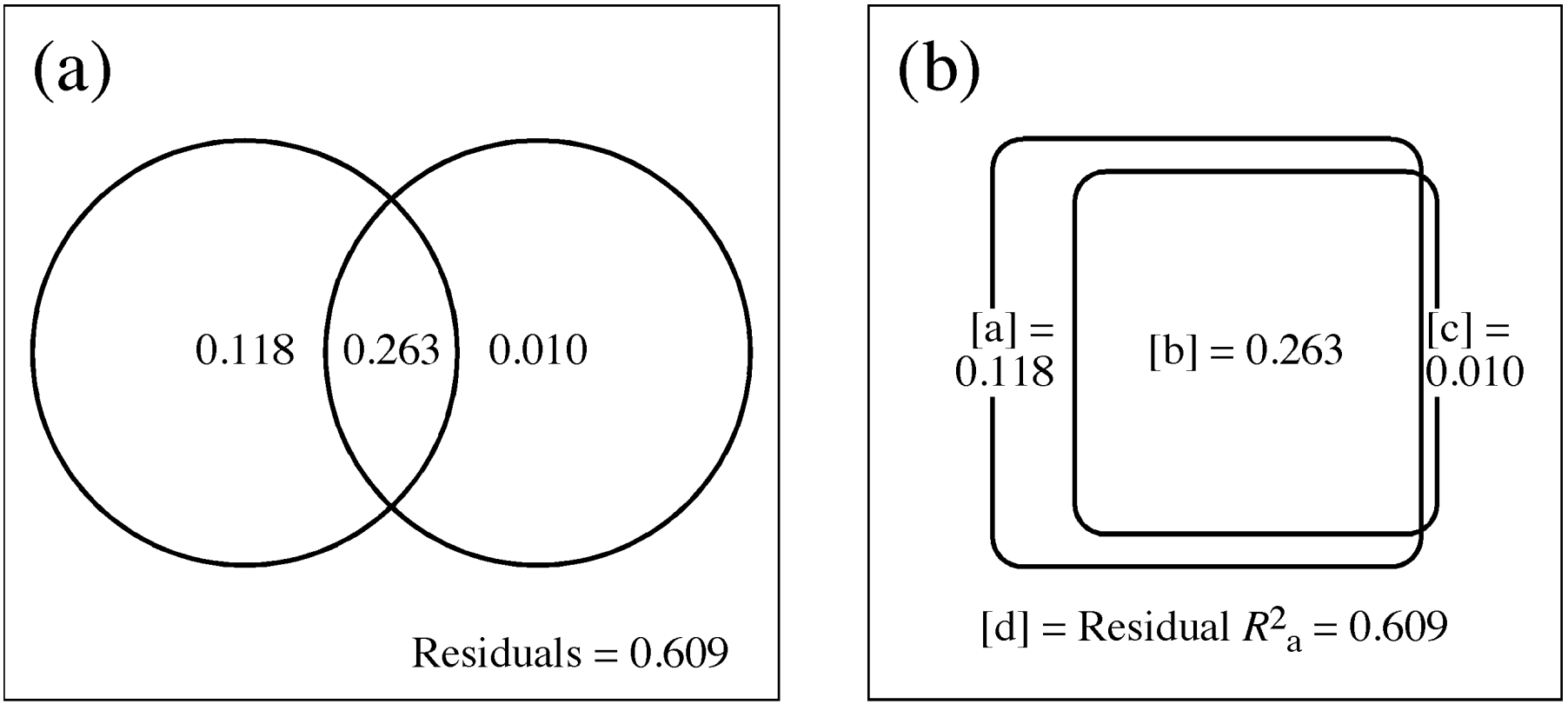
Variation partitioning table – Legendre & Legendre 2012, p. 577.

y = MA (bacterial abundances), **X** = {NH4, PhaeoA, Prod.}, **W** = {X, Y, X^2}.

Fractions of variation	Sums of squares (SS)	Proportions of variation of y (R^2)	Adjusted R^2 (R^2_a)
[a + b]	7.1547	0.4793	0.3817
[b + c]	5.7895	0.3878	0.2731
[a + b + c]	8.7109	0.5835	0.3913
[a]	2.9213	*0.1957*	0.1183
[b]	4.2333	*0.2836*	0.2634
[c]	1.5562	*0.1043*	0.0097
Residuals = [d]	6.2167	*0.4165*	0.6087
[a + b + c + d]	14.9276	1.0000	1.0000

Note

The R^2_{adj} statistic can be negative. This indicates that the proportion of variance of **Y** explained by **X** is worse than a set of *m* random normal deviates would do. A negative R^2_{adj} statistic ("worse than 0") can be ignored during interpretation.



Venn diagrams – Results of variation partitioning of the numerical example. The fractions are adjusted R-squares (R^2_{adj}).
(a) Diagram drawn by function plot.varpart() of the vegan package. Circle sizes are not to scale.
(b) Diagram redrawn, before publication of the results, using rounded rectangles, with fraction sizes proportional to the values. (Legendre & Legendre 2012, Fig. 10.11.)

Run the R code to partition the variation of response variable "Ma" between matrices **X1** of environmental variables and **X2** of spatial variables using function varpart() of the vegan package.

```
# Read the file "Table_10.6.txt"
mat <- read.table(file.choose())
y  <- mat[,2]   # Extract the response variable Ma
X1 <- mat[,3:5] # Matrix X1 of environmental var.
X2 <- mat[,6:8] # Matrix X2 of spatial variables
# Variation partitioning
library(vegan)
res <- varpart(y, X1, X2)
res             # Print out the partitioning table
plot(res, bg=2:3, digits=3)
```

Compare the output table of the function (object "res") to the one shown two slides back and to the Venn diagram produced by function plot.varpart().

Tests of significance of the fractions

The significance of the unique fractions ([a] and [c]) can be tested using the test available in partial redundancy analysis (partial RDA).

The common fraction [b] is not an adjusted component of variance. It cannot be estimated nor tested by ordinary linear modelling methods.

Latest news – A method has recently published by Bauman et al. (2019) to test fraction [b] in the special case of variation partitioning involving environmental variables and Moran's eigenvector maps in spatial analysis.

Note about fraction [b], explained jointly by **X** and **W**

A [b] fraction *is not* a statistical interaction.

In two-way analysis of variance (anova with two crossed factors), an interaction is present when the effect of a factor varies according to the class of the other factor.

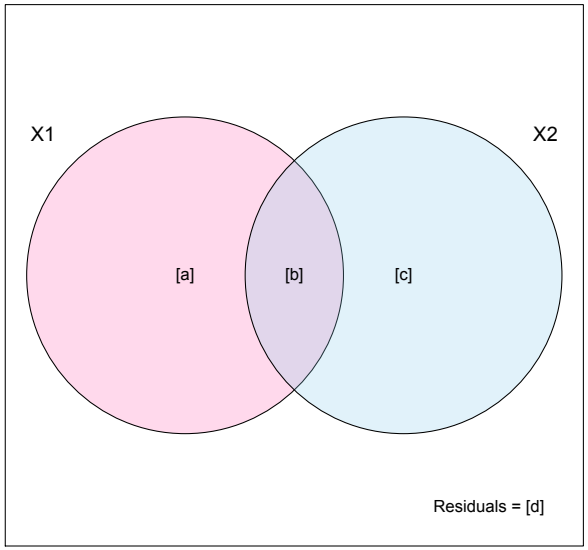
An interaction is most easily measured in balanced crossed-factor anova, where the two factors are independent (uncorrelated, orthogonal). In that case, the [b] fraction is equal to 0.

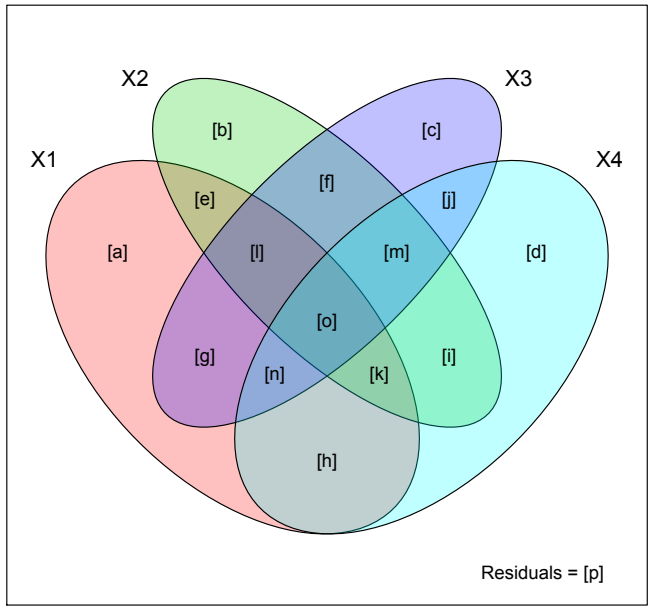
This demonstrates that a [b] fraction is not an interaction.

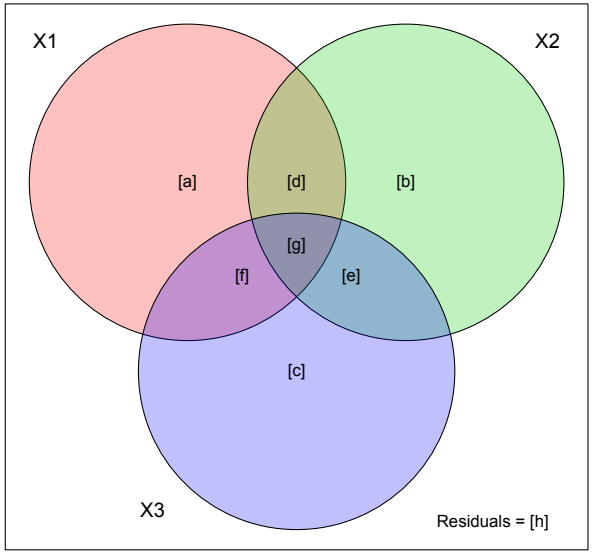
5. R software

Partial regression: use the rda function and compute a partial RDA of **y**.

Variation partitioning: functions varpart() and plot.varpart() in vegan.







6. References

Bauman, D., J. Vleminckx, O. J. Hardy & T. Drouet (2019). Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data. *Oikos* 128: 274–285.

Borcard, D., F. Gillet & P. Legendre. 2018. *Numerical ecology with R, 2nd edition*. Use R! series, Springer International Publishing, New York.

Borcard, D., P. Legendre & P. Drapeau. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73: 1045-1055.

Legendre, P. & L. Legendre. 2012. Interpretation of ecological structures. Chapter 10 in: *Numerical ecology, 3rd English edition*. Elsevier Science BV, Amsterdam.

Peres-Neto, P. R., P. Legendre, S. Dray & D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87: 2614–2625.

End of the presentation