Régression et corrélation multiples et partielles

r^2 partiel, contribution et fraction [a]

Daniel Borcard
Université de Montréal
Département de sciences biologiques
Janvier 2002

Une certaine confusion règne souvent en ce qui a trait aux définitions des notions suivantes:

1. Contribution d'une variable x_j à l'explication de la variation d'une variable dépendante y;
2. Fraction [a] d'un partitionnement de variance.
3. r^2 partiel (coefficient de détermination partielle) entre une variable x_j et une variable y;

Le présent document suppose connues les bases de la régression linéaire simple et multiple, et le principe du partitionnement de la variance en régression (pour ce dernier point, voir en particulier Legendre et Legendre (1998), p. 528 et suivantes).

Définissons d'abord les trois notions.

Définitions

1. **Contribution** d'une variable x_j à l'explication de la variation d'une variable dépendante y

Ce terme est utilisé par Scherrer (1984) dans le cadre du calcul du coefficient de détermination multiple d'une régression linéaire multiple (pp. 699-700). Notons que Scherrer appelle improprement "coefficient de corrélation multiple" le coefficient de détermination multiple (R^2). Le coefficient de corrélation multiple (R) est la racine carrée du coefficient de détermination multiple.

Le coefficient de détermination multiple mesure la proportion de variance d'une variable dépendante y expliquée par un ensemble de k variables explicatives x. Il peut se calculer par

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

où a'_j est le coefficient de régression centré-réduit de la $j_{\text{ième}}$ variable explicative et r_{yx_j} est le coefficient de corrélation linéaire simple (Pearson) entre y et x_j [1].

[1] Scherrer note cette équation $R^2 = \sum_{j=1}^{p-1} a'_j r_{jp}$, où la $p_{\text{ième}}$ variable est la variable dépendante

Dans ce contexte, Scherrer nomme "contribution" de la $j_{\text{ième}}$ variable à l'explication de la variance de y la quantité $a'_j r_{yx_j}$. La somme des contributions de toutes les variables explicatives x_k donne le R^2. Remarquons que chaque contribution peut être positive ou négative.

2. **Fraction [a]** d'un partitionnement de variation

Cette fraction mesure la proportion de variance de y expliquée par la variable explicative x_1 (par exemple) lorsque les autres variables explicatives (x_2, x_3...) sont tenues constantes **par rapport à x_1 seulement** (et non par rapport à y).

On obtient donc la fraction [a] en examinant le r^2 obtenu en régressant y sur le résidu d'une régression de x_1 sur x_2, x_3**…**

Remarque: on peut calculer une fraction [a] pour plusieurs variables explicatives simultanément, mais pour la simplicité de la présentation nous nous en tenons ici au cas où on cherche la fraction d'une seule variable explicative.

3. $\boldsymbol{r^2}$ **partiel** (coefficient de détermination partielle) entre une variable x_j et une variable y

Remarque: lorsqu'il n'y a que deux variables en cause, on utilise généralement le r ou le r^2 minuscule. On utilise les majuscules pour les coefficients de corrélation multiple (R) ou de détermination multiple (R^2).

Tout d'abord, le ***r* partiel** mesure la liaison mutuelle entre deux variables y et x_j lorsque d'autres variables (x_1, x_2, x_3...) sont tenues constantes **par rapport aux deux variables impliquées $\boldsymbol{y}$ et $\boldsymbol{x_j}$** (au contraire du cas précédent).

Le coefficient de corrélation partielle est très utile en régression multiple, où il "...permet d'évaluer directement la proportion de la variation non expliquée de y qui devient expliquée grâce à l'ajout de la variable x_j." (Scherrer, p.702).

Dans le cas où on a une variable explicative x_1 et une variable maintenue constante x_2, le coefficient de corrélation partielle se calcule comme suit (Scherrer, équ. 18-50, p. 704):

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

Le $\boldsymbol{r^2}$ **partiel** est le carré du r partiel ci-dessus, et mesure la proportion de variance du résidu de y par rapport à x_2 expliquée par le résidu de x_1 par rapport à x_2. Il peut donc aussi être obtenu en examinant le r^2 d'une régression du résidu de y par rapport à x_2 sur le résidu de x_1 par rapport à x_2.

Examinons maintenant les propriétés de ces trois éléments dans deux situations différentes impliquant chaque fois une variable dépendante y et deux variables explicatives x_1 et x_2.

Exemple 1. Variables explicatives corrélées entre elles (cas le plus général)

Données:

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

Matrice de corrélations linéaires simples:

	x_1	x_2
y	0.677	-0.824
x_1		-0.612

Matrice de corrélations linéaires partielles:

	x_1	x_2
y	0.385	-0.704
x_1		-0.131

Coefficients de régression:

	Coeff. bruts	Coeff. centrés-réduits
Ord.origine	6.300	0.000
x_1	1.019	0.276
x_2	-0.523	-0.655

Le coefficient de **détermination multiple** de la régression peut être calculé sur la base de l'équation 1 (voir plus haut):

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

Les calculs intermédiaires nous fournissent les informations suivantes:

- **contribution** de x_1 à l'explication de la variance de y = 0.187

- **contribution** de x_2 à l'explication de la variance de y = 0.540

r^2 partiel de y et x_1, en tenant x_2 constant par rapport à y et x_1(équation 2):

$$r_{y,x_1|x_2} = \frac{0.677 - (-0.824 \times -0.612)}{\sqrt{[1-(-0.824)^2][1-(-0.612)^2]}} = \frac{0.1727}{\sqrt{0.3210 \times 0.6255}} = \frac{0.1727}{0.4481} = 0.385$$

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

r^2 partiel de y et x_2, en tenant x_1 constant par rapport à y et x_2:

$$r_{y,x_2|x_1} = \frac{-0.824-(0.677\times-0.612)}{\sqrt{[1-(0.677)^2][1-(-0.612)^2]}} = \frac{-0.4097}{\sqrt{0.5417\times0.6255}} = \frac{-0.4097}{0.5821} = -0.704$$

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

Fraction [a]: proportion de variance de y expliquée par la variable explicative x_1 lorsque x_2 est tenue constante par rapport à x_1 seulement (et non par rapport à y). C'est le r^2 obtenu en régressant y sur le résidu d'une régression de x_1 sur x_2 (détails omis):

[a] = 0.0476

Fraction [c]: proportion de variance de y expliquée par la variable explicative x_2 lorsque x_1 est tenue constante par rapport à x_2 seulement (et non par rapport à y). C'est le r^2 obtenu en régressant y sur le résidu d'une régression de x_2 sur x_1 (détails omis):

[c] = 0.268

Fraction [b]: R^2 de la régression multiple de y sur x_1 et x_2 – [a] – [c] = 0.727 – 0.048 – 0.268 = 0.411

Le partitionnement donne donc (respectivement fractions [a], [b], [c] et[d]):

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

Remarque: avec ces résultats, on peut aussi calculer le r^2 partiel de y sur x_1, en contrôlant pour x_2; ce r^2 partiel peut en effet se définir par [a]/[a]+[d]:

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

Le même calcul pourrait être fait pour le r^2 partiel de y sur x_2.

Exemple 2. Variables explicatives orthogonales entre elles (linéairement indépendantes l'une de l'autre)

Les variables x_1 et x_2 peuvent représenter, par exemple, deux critères de classification décrivant un plan d'expérience à deux facteurs orthogonaux. Dans ce cas, l'analyse de variance peut être calculée par régression multiple, comme dans cet exemple.

Données:

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

Matrice de corrélations linéaires simples:

	x_1	x_2
y	0.677	0.496
x_1		0.000

Matrice de corrélations linéaires partielles:

	x_1	x_2
y	0.780	0.674
x_1		-0.526

Coefficients de régression:

	Coeff. bruts	Coeff. centrés-réduits
Ord.origine	-1.417	0.000
x_1	2.500	0.677
x_2	1.833	0.496

Coefficient de **détermination multiple** (équation 1):

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

Donc:

- **contribution** de x_1 à l'explication de la variance de y = 0.458

- **contribution** de x_2 à l'explication de la variance de y = 0.246

***r*² partiel** de y et x_1, en tenant x_2 constant par rapport à y et x_1 (équation 2):

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

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

r^2 partiel de y et x_2, en tenant x_1 constant par rapport à y et 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 de variance de y expliquée par la variable explicative x_1 lorsque x_2 est tenue constante par rapport à x_1 seulement (et non par rapport à y). Dans cet exemple, x_1 et x_2 sont orthogonaux (linéairement indépendants), de sorte que la fraction [a] est directement égale au r^2 obtenu en régressant y sur x_1 (détails omis):

[a] = 0.458

Fraction [c]: proportion de variance de y expliquée par la variable explicative x_2 lorsque x_1 est tenue constante par rapport à x_2 seulement (et non par rapport à y). Dans cet exemple, x_1 et x_2 sont orthogonaux (linéairement indépendants), de sorte que la fraction [c] est directement égale au r^2 obtenu en régressant y sur x_2 (détails omis):

[c] = 0.246

Fraction [b]: R^2 de la régression multiple de y sur x_1 et x_2 – [a] – [c] =

= 0.704 – 0.458 – 0.246 = 0.000

L'orthogonalité des variables explicatives x_1 et x_2 se manifeste donc par la valeur nulle de la fraction [b].

Le partitionnement donne donc (respectivement fractions [a], [b], [c] et[d]):

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

On voit que, **dans ce cas** (orthogonalité de x_1 et x_2), les fractions [a] et [c] sont égales aux contributions des variables x_1 et x_2 à l'explication de la variance de y.

Ici encore, on peut vérifier que l'équation $r^2_{y,\ x1|x2}$ = [a]/[a]+[d] donne bien le r^2 partiel obtenu plus haut:

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

Commentaires (et résumé des modes de calcul)

1. La contribution (au sens de Scherrer) d'une variable explicative n'est égale à la fraction [a] de variance expliquée (au sens du partitionnement de variance) que dans **un seul cas**: lorsque toutes les variables explicatives sont orthogonales entre elles (linéairement indépendantes entre elles, non corrélées entre elles). Dans ce cas, la fraction [b] du partitionnement de variance est égale à zéro, puisque chaque variable explicative explique une fraction complètement différente de la variance de y.

Le R^2 total de la régression multiple (coefficient de détermination multiple) se calcule alors comme suit:

- soit en faisant la somme des $a'_j r_{yx_j}$

- soit en additionnant les fractions [a] et [c] du partitionnement (puisque [b] est nul!).

2. Dans le cas général, c'est-à-dire lorsque les variables explicatives sont plus ou moins corrélées entre elles, elles expliquent chacune une part de la variation de *y*, mais ces fractions se recouvrent plus ou moins. Chaque variable explicative fait "une partie du travail" de l'autre ou des autres, puisqu'elles sont un peu corrélées entre elles. Le résultat se concrétise dans la fraction [b] du partitionnement, qui n'est plus nulle. En général, cette fraction [b] est positive, et "gruge" donc une partie des fractions [a] et [c]. Ces dernières sont donc plus petites que leurs contributions partielles (les contributions partielles intègrent la fraction [a] ou [c] **plus** une partie de la fraction [b]).

Dans ce cas, le R^2 total de la régression multiple (coefficient de détermination multiple) se calcule comme suit:

- soit en faisant la somme des $a_j' r_{yx_j}$

- soit en additionnant les fractions [a], [c] **et [b]** (puisque cette dernière est maintenant non nulle!).

Remarque: on observe parfois des cas où la fraction [b] est négative! Cela arrive lorsque deux variables explicatives ont des effets forts et opposés sur la variable dépendante, tout en étant corrélées entre elles. Dans ce cas, les fractions [a] et [c] sont plus grandes que leurs contributions partielles! Une explication détaillée de ce phénomène est fournie par Legendre et Legendre (1998) p. 533.

Mode de calcul

Pour obtenir, par exemple, la fraction [a] d'une régression de *y* expliquée par x_1 et x_2, vous pouvez faire comme suit:

1e étape: faites une régression de x_1 expliqué par x_2 et conservez les résidus. Ce faisant, vous avez débarrassé la variable explicative qui vous intéresse (x_1) des effets de l'autre variable explicative (x_2).

2e étape: faites une régression de *y* expliqué par les résidus obtenus ci-dessus. Le r^2 de cette deuxième régression est égal à la fraction [a]. Vous avez alors expliqué *y* avec la partie de x_1 qui n'a aucun rapport avec x_2.

Si vous n'êtes pas intéressé aux valeurs ajustées, mais uniquement aux valeurs des fractions, vous pouvez aussi réaliser tout le partitionnement sans avoir recours à une seule régression partielle. Les étapes sont les suivantes:

1. La régression de *y* sur x_1 donne [a] + [b].

2. La régression de *y* sur x_2 donne [b] + [c].

3. La régression multiple de *y* sur x_1 et x_2 donne [a] + [b] + [c].

4. [a] s'obtient en soustrayant le résultat du point 2 ci-dessus de celui de 3.

5. [b] s'obtient en soustrayant le résultat du point 3 ci-dessus de celui de 1.

6. [d] s'obtient en soustrayant du nombre 1 le résultat obtenu au point 3.

3. Le r^2 partiel, comme dit plus haut, mesure (le carré de la) liaison mutuelle entre deux variables lorsque d'autres variables sont tenues constantes par rapport aux **deux** variables impliquées. Rappelez-vous que la régression de modèle I quantifie l'effet d'une variable explicative sur une variable expliquée, alors que la corrélation mesure leur liaison mutuelle. C'est vrai aussi dans le cas partiel. En ce qui concerne les calculs, ça se traduit de deux manières possibles. La démarche ci-

dessous montre le calcul du r^2 partiel entre y et une variable x_1, en retirant l'effet d'une variable x_2, comme pour la fraction [a], mais en mettant l'accent sur les différences entre les deux méthodes:

• Calcul par les résidus:

1. Faites une régression de y expliqué par x_2 et conservez les résidus. Ainsi, vous débarrassez y des effets de x_2.

2. Faites une régression de x_1 expliqué par x_2 et conservez les résidus. Ainsi, vous débarrassez x_1 des effets de x_2.

3a. Première variante: faites une régression des résidus de l'étape 1 expliqués par les résidus de l'étape 2 ci-dessus.

3b. Seconde variante: calculez la corrélation linéaire entre les résidus de l'étape 1 et les résidus de l'étape 2 ci-dessus. Ceci est une autre façon de calculer le coefficient de corrélation partielle.

Contrairement à l'exemple de la fraction [a] plus haut, cette opération évacue complètement l'influence de x_2, aussi bien de y que de x_1. En conséquence, le r^2 obtenu est le r^2 partiel de y et x_1 en contrôlant l'effet de x_2! Et la pente centrée-réduite (coefficient de régression centré-réduit) de cette régression est le coefficient de corrélation partielle entre y et x_1.

Remarque: si x_1 et x_2 sont orthogonaux, l'étape 2 est inutile, car x_2 n'explique rien de x_1, et donc les résidus sont égaux à x_1 (ils sont centrés, mais c'est sans importance).

• Calcul par les fractions de variation:

Le r^2 partiel se calcule aussi par [a]/[a]+[d]. L'examen de cette équation montre qu'on compare:

- au numérateur, la fraction de variance de y expliquée uniquement par x_1;
- au dénominateur, cette même fraction **plus** la variance non expliquée, mais **pas** les fractions [b] et [c] impliquées dans l'intervention de la variable x_2.

4. Je rappelle enfin que **la fraction [b] n'a rien à voir avec l'interaction d'une ANOVA**. En ANOVA, **l'interaction mesure l'effet d'une variable explicative sur l'influence qu'a l'autre variable explicative sur la variable dépendante**. Or, cette interaction peut avoir une valeur non-nulle lorsque les deux variables explicatives sont orthogonales, ce qui est justement la situation où la fraction [b] vaut zéro.

Bibliographie

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

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