Compléments sur le modèle de régression multiple

JM Galharret [1]

[1]UFR de Psychologie
Université de Nantes

March 22, 2020

Sommaire

Sommaire

Introduction

On s'intéresse à :

- La comparaison d'équations de régression : quelle est celle qui prédit le mieux la variable réponse Y ?
- Le partionnement de la variance étant donnés plusieurs régresseurs comment décomposer R^2.
- La multicolinéarité : certains prédicteurs $X_1, ..., X_p$ dans l'équation de régression sont linéairement liés les uns aux autres ou très fortement corrélés.

Sommaire

Exemples :

Modèle	Equation
$\mathcal{M}_1$	$govact \sim posemot + negemot$
$\mathcal{M}_2$	$govact \sim age + negemot$
$\mathcal{M}_3$	$govact \sim negemot + posemot + ideology$
$\mathcal{M}_4$	$govact \sim negmot + posemot + ideology + age$

On va comparer :

1. **Modèles non emboités** : $\mathcal{M}_1$ et $\mathcal{M}_2$, $\mathcal{M}_2$ et $\mathcal{M}_3$
2. **Modèles emboités** tous les modèles $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ sont des sous-modèles de $\mathcal{M}_4$. $\mathcal{M}_1$ est emboité dans $\mathcal{M}_3$.

Le problème du R^2

Considérons une équation de régression $\mathcal{M}$: $Y = b_0 + b_1X_1 + \ldots + b_pX_p$. On définit le % de la variance expliqué par $\mathcal{M}$:

$$R^2 = \frac{SCE_{\mathcal{M}}}{SCE_Y} = 1 - \frac{SCE_{Res}}{SCE_Y}$$

Propriétés :

- $0 < R^2 < 1$
- R^2 augmente avec le nombre de prédicteurs dans le modèle (jusqu'à valoir 1 si on met $n-1$ variables dans le modèle)
- On ne peut donc utiliser R^2 pour comparer deux modèles que si ils ont le même nombre de variables.

Le R^2 ajusté

On définit un autre coefficient appelé R_a^2 (R^2 ajusté) qui tient compte du nombre de variables du modèle en ajustant les SCE par les ddl:

$$R_a^2 = 1 - \frac{s^2_{Res}}{s^2_Y}$$

Remarque : On a toujours $R_a^2 \leq R^2$

Critère de comparaison

Soient M_1 et M_2 deux équations de régressions de Y ayant respectivement p_1 et p_2 prédicteurs. On calcule les R^2 ajustés de chacun des modèles $R^2_{a,1}$ et $R^2_{a,2}$. Le modèle $\mathcal{M}_1$ sera meilleur que le modèle $\mathcal{M}_2$ si :

$$R^2_{a,1} > R^2_{a,2}$$

Exemples :

Modèle	Equation	R^2	R_a^2
$\mathcal{M}_1$	$govact \sim posemot + negemot$	0.335	0.333
$\mathcal{M}_2$	$govact \sim age + negemot$	0.338	0.336
$\mathcal{M}_3$	$govact \sim negemot + posemot + ideology$	0.388	0.386
$\mathcal{M}_4$	$govact \sim negmot + posemot + ideology + age$	0.388	0.385

- $\mathcal{M}_2$ est meilleur que $\mathcal{M}_1$
- $\mathcal{M}_3$ est meilleur que $\mathcal{M}_2$
- $\mathcal{M}_4$ est meilleur que $\mathcal{M}_1$. $\rightsquigarrow$ significativement ?

Test pour modèles emboités

On considère deux équations de régression emboitées

$$M_0 : Y = b_0 + b_1 X_1 + ... + b_p X_p \tag{1}$$
$$M_1 : Y = b_0 + b_1 X_1 + ... + b_p X_p + b_{p+1} X_{p+1} + ... b_{p+j} X_{p+j} \tag{2}$$

Test de FISHER

Pour tester $H_0 : b_{p+1} = \ldots = b_{p+j} = 0$, on utilise la statistique de test

$$F = \frac{n-p-j-1}{j} \times \frac{SCE_{\mathcal{M}_1} - SCE_{\mathcal{M}_0}}{SCE_{Res_1}}$$

. Lorsque H_0 est vraie on a $F \sim \mathcal{F}(j, n-p-j-1)$

Test de Fisher entre les modèles $\mathcal{M}_1$ et $\mathcal{M}_4$

Table: ANOVA

Model		Sum of Squares	df	Mean Square	F	p
$\mathcal{M}_1$	Regression	504.3	2	252.15	204.306	$< .001$
	Residual	1002.2	812	1.234		
	Total	1506.5	814			
$\mathcal{M}_4$	Regression	585	4	146.250	128.548	$< .001$
	Residual	921.54	810	1.138		
	Total	1506.54	814			

$$F = \frac{810}{4-2} \times \frac{585 - 504.3}{921.5} = 33.35$$

L'équation 1 prédit significativement mieux govact que l'équation 0, $F(2, 810) = 33.35$, $p < .001$

Sommaire

Le partitionnement de la variance avec des facteurs (ANOVA)

Dans le cas de l'ANOVA, c'est-à-dire avec des prédicteurs qualitatifs les facteurs n'ont pas de part de variance commune.
On a :

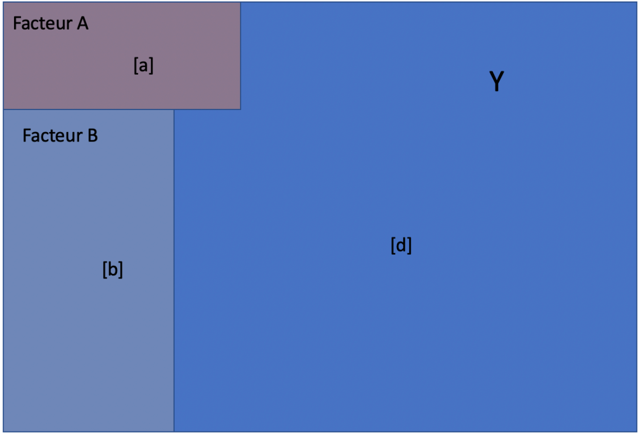
$$SCE_{A+B}=[a]+[b]$$

Remarque : ici dans le modèle on a exclu l'effet d'interaction.

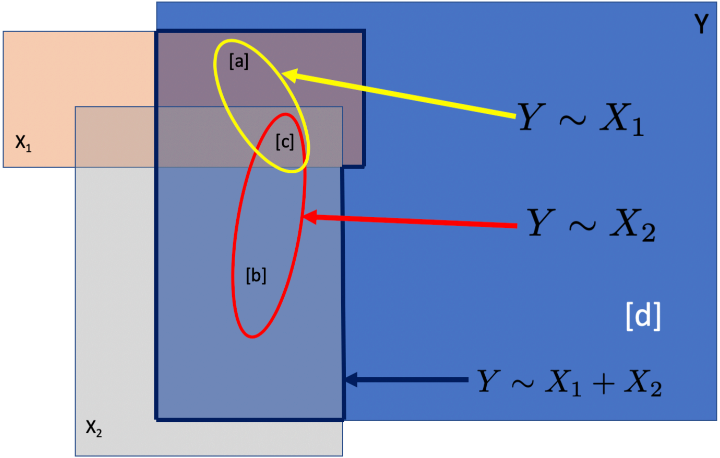
$$\eta^2 = \frac{[a]+[b]}{[a]+[b]+[d]}$$

$$\eta^2_A = \frac{[a]}{[a]+[b]+[d]}$$

$$\eta^2_{p,A} = \frac{[a]}{[a]+[d]}$$



Le partitionnement de la variance avec des facteurs (RLM)



[c] : une partie commune à X_1 et X_2.

[d] : résidu du modèle de régression de Y en fonction de X_1 et X_2

Exemple

Soit le modèle $\mathcal{M}$: $govact \sim negmot + ideology$. Compléter le tableau suivant :

Equation	valeurs	SCE	R^2
$govact \sim negmot$	$[a+c]$	502.87	33.4%
$govact \sim ideology$	$[b+c]$	263.63	17.5 %
$govact \sim negmot + ideology$	$[a+b+c]$	583.50	38.7%
$govact \sim negmot + ideology$	$[d]$	923.05	$-$
$[a+b+c]-[b+c]$	$[a]$	319.87	21.2 %
$[a+b+c]-[a+c]$	$[b]$	80.63	5.4%
$[a+c]-[a]$	$[c]$	183.00	12.1%

On peut vérifier que $319.87+80.63+183=583.5$. Pour le calcul du R^2, on a $SCE_T=[a+b+c]+[d]=583.5+923.05=1506.55$.

Corrélation semi-partielle

On appelle corrélation semi-partielle de Y avec X_1 conditionnellement à $X_2, \ldots, X_p$ le nombre défini par :

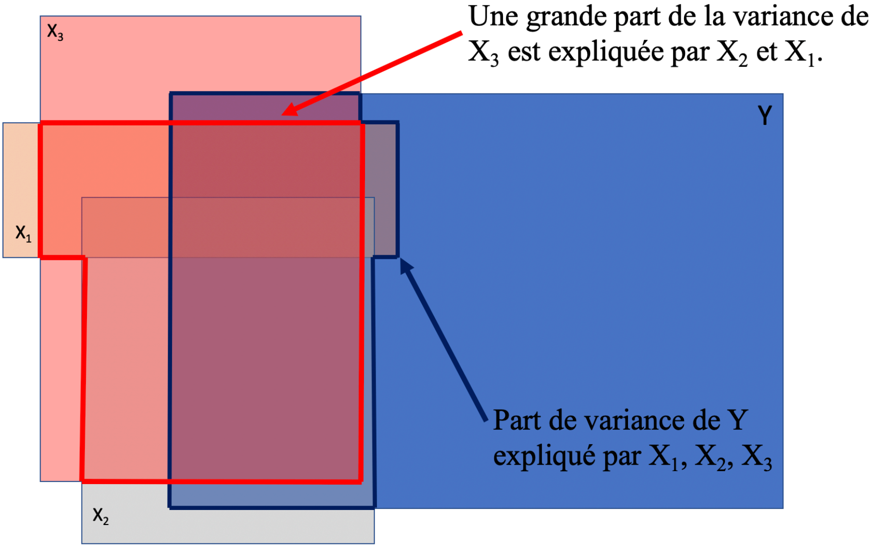
$$r^2_{(Y,X_1)|X_2,\ldots,X_p} = \frac{SCE_{\mathcal{M}_1} - SCE_{\mathcal{M}_0}}{SCE_T}$$

où $\mathcal{M}_1 : Y = b_0 + b_1X_1 + \ldots + b_pX_p$ et $\mathcal{M}_0 : Y = b_0 + b_2X_2 + \ldots + b_pX_p$.

Dans l'exemple précédent on a $r^2_{(Y,X_1)|X_2} = 21.2\%$ et $r^2_{(Y,X_2)|X_1} = 5.4\%$

Sommaire

Le problème de la multicolinéarité



VIF (Variance Inflation Factor)

Soit une équation de régression $Y = b_0 + b_1 X_1 + \ldots + b_p X_p$.
Pour quantifier le lien entre X_1 et les autres prédicteurs $X_2, \ldots X_p$ on écrit l'équation de régression de X_1 en fonction de $X_2, \ldots X_p$ et on calcule le pourcentage de variance de X_1 prédit par $X_2, \ldots X_p$ noté R_1^2 et on définit $VIF(X_1)$ par

$$VIF(X_1) = \frac{1}{1 - R_1^2}$$

Règle

On considère que X_1 est linéairement liée à $X_2, \ldots, X_p$ si $VIF(X_1) > 5$. (ce qui revient à $R_1^2 > .80$)

Multicolinéarité

On calcule les VIF de tous les prédicteurs du modèle et ensuite si une ou plusieurs prédicteurs ont un $VIF > 5$ alors :

1. On choisit le prédicteur ayant le plus grand VIF (disons qu'il s'agit de X_1).
2. On estime l'équation de Y en fonction $X_2, ..., X_p$ et on recalcule tous les VIF.
3. Si aucun des $VIF > 5$ on conserve l'équation sinon on reprend à l'étape 1.

Retour sur l'exemple : $\mathcal{M}_1$: *govact* $\sim$ *negmot* $+$ *ideology*
On avait d'après la matrice de corrélation $r = -.349, p < .001$ entre negemot et ideology. $R_1^2 = r^2 = .1218$ et donc $VIF(X_1) = \frac{1}{1 - .1218} = 1.14$ donc pas de problème de colinéarité entre negemot et ideology.