Les effets d'interaction

Jean-François Bickel

Statistique II – SP08

1. Qu'est-ce qu'une interaction?

- Soit le modèle de régression

  $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k$

- Jusqu'à maintenant, nous avons considéré l'effet de chaque variable indépendante x_1, $x_2 \ldots x_k$ comme constant quelque soit la valeur prise par les autres variables indépendantes

- La possibilité existe pourtant que l'effet de x_1, ou de x_2, ou… de x_k ne soit pas constant, mais *varie* en fonction des valeurs prises par une des autres variables indépendantes introduite dans le modèle
- Par exemple, que l'effet de x_1 diffère selon la valeur prise par x_2
- On dit dans ce cas qu'il y a *interaction* entre x_1 et x_2

- ***Nota Bene***

  On peut étendre ce principe et s'intéresser aux cas où l'effet d'une variable x_1 ou x_2 ou…x_k dépend de 2, 3… autres variables du modèle; par souci de simplification, on en restera au cas le plus commun d'interaction entre 2 variables

- Nous allons examiner tour à tour trois formes d'interaction, selon le type de variables indépendantes qu'elles impliquent

a) Interaction entre 2 variables quantitatives (intervalles)

b) Interaction entre 1 variable quantitative et 1 variable catégorielle

c) Interaction entre 2 variables catégorielles

2. Interaction entre 2 variables quantitatives (intervalles)

- Considérons à titre d'illustration les deux variables âge et niveau d'éducation, comme facteurs conditionnant le revenu du travail

- Faisons l'hypothèse que l'effet positif de l'âge sur le revenu est plus fort pour les personnes avec un niveau de formation plus élevé, car celles-ci, au fur et à mesure qu'elles avancent en âge, peuvent mieux tirer parti des opportunités de promotion et bénéficient davantage de la règle d'ancienneté
- Si cette hypothèse est correcte, alors il y a interaction entre âge et éducation

- Comment tester une telle hypothèse et introduire une interaction dans le modèle de régression tel que nous connaissons?
- Partons du modèle de base

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

avec
x_1 = âge
x_2 = éducation

- Ce que stipule notre hypothèse, est que l'effet de l'âge (le coefficient β_1) est fonction de l'éducation (x_2)
- Ceci peut être représenté sous la forme suivante

$$\beta_1 = C + Dx_2$$

où C et D sont des nombres à estimer

- C peut être interprété comme la valeur de β_1 quand éducation égale à zéro (i.e. quand $x_2=0$)
- D est un coefficient qui nous dit de combien l'effet de l'âge (β_1) change quand le niveau d'éducation s'élève d'une unité
- En remplaçant dans l'équation initiale β_1 par son équivalent $C+Dx_{2;}$ on obtient

$$E(y) = \alpha + (C+Dx_2)x_1 + \beta_2 x_2$$

- En multipliant les termes entre parenthèses par x_1, on obtient

  $E(y) = \alpha + Cx_1 + Dx_2x_1 + \beta_2x_2$

- Si on change l'ordre des éléments et revient à la notation usuelle, on a de manière équivalente

  $E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$

- La nouvelle équation de régression contient donc les deux variables indépendantes x_1 et x_2 mais aussi une nouvelle variable, définie comme étant le produit de x_1 et x_2
- A chacun de ces trois termes est associé, comme usuellement, un coefficient de régression, dénotés β_1, β_2 et β_3
- Ces derniers sont estimés par les coefficients b_1, b_2 et b_3 calculés sur la base des données observées

Syntaxe

1) Création du terme d'interaction

```
compute ageXeduc=age05*educat05.
exe.
```

2) Modèle de régression

```
regression
  /missing listwise
  /statistics defaults ci change
  /noorigin
  /dependent i05wy
  /method=enter age05 educat05
  /method=enter ageXeduc.
```

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Changement dans les statistiques				
					Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
1	.450[a]	.202	.202	51393.316	.202	530.579	2	4183	.000
2	.458[b]	.210	.209	51159.985	.007	39.243	1	4182	.000

- La variation du R^2 (.007; p<.001) nous donne une première indication de l'existence d'un effet d'interaction

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	-14839.2	2795.291		-5.309	.000	-20319.48	-9358.963
	AGE05 Age durant l'année de l'interview	889.857	64.143	.198	13.873	.000	764.104	1015.611
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	7109.914	282.245	.359	25.191	.000	6556.564	7663.264
2	(constante)	6160.728	4356.672		1.414	.157	-2380.663	14702.120
	AGE05 Age durant l'année de l'interview	311.827	112.210	.069	2.779	.005	91.835	531.819
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	2084.078	850.061	.105	2.452	.014	417.507	3750.649
	ageXeduc	129.652	20.697	.323	6.264	.000	89.076	170.229

Interprétation

1) Examiner la significativité statistique du coefficient du terme d'interaction
   - Ici, elle est inférieure à .05, il y a donc interaction
   - Si le test donne un résultat supérieur à p=.05, il est préférable de supprimer le terme d'interaction de l'équation et de traiter les effets des variables en question comme indépendants l'un de l'autre

2) Examiner le signe du coefficient du terme d'interaction

- Ici, il est de signe positif, ce qui indique que l'effet de l'âge sur le revenu s'accroît en même temps que s'accroît le niveau d'éducation (et réciproquement)

3) Interpréter le coefficient du terme d'interaction

- Pour cette interprétation, on peut calculer l'effet de l'âge pour différentes valeurs d'éducation
- Il suffit pour cela de reprendre la formule vue plus haut posant l'effet de l'âge comme étant fonction linéaire de l'éducation
  effet de l'âge = $C+Dx_2$
  ou alternativement
  effet de l'âge = $\beta_1 + \beta_3 x_2$

- Ici, cela donne

  Effet de l'âge = 312 + (130 x éducation)

  (N.B. pour faciliter les calculs, les coefficients sont arrondis)

- En choisissant certains niveaux « typiques » d'éducation, et en appliquant la formule, on obtient l'effet estimé de l'âge pour ces différentes situations
- Par exemple, pour educat05=0 (école obligatoire inachevée), on obtient

  312 + (130 x 0) = 312

- Ce qui s'interprète comme suit:
  pour les personnes qui n'ont pas achevé l'école obligatoire, le revenu augmente en moyenne de 312 Frs par année d'âge
- Autre exemple, pour educat05=10 (université), on obtient
  312 + (130 x 10) = 1612
- Ce qui s'interprète comme suit:
  pour les personnes avec une formation universitaire, le revenu augmente en moyenne de 1612 Frs par année d'âge

- Le tableau suivant indique la valeur de l'effet de l'âge pour quelques valeurs de niveaux d'éducation
- Il met en évidence que plus ce niveau augmente, plus l'effet de l'âge sur le revenu est grand

Effet de l'âge sur le revenu pour différents niveaux d'éducation

Niveau d'éducation	*Effet de l'âge*
0 (=école obligatoire inachevée)	312
4 (=apprentissage)	832
6 (=maturité)	1092
10 (=université)	1612

4) Quand, dans l'équation de régression, il y a un terme d'interaction, les coefficients pour les variables incluses dans l'interaction prennent un sens particulier

- Le coefficient pour âge (312) réfère à la situation où éducation=0
- Le coefficient pour éducation (2084) réfère à la situation où âge=0
  Lorsque, comme ici, la situation à laquelle se réfère le coefficient ne fait pas sens, celui-ci n'est pas interprété

5) Mais, il y a une autre façon d'interpréter les résultats de notre équation de régression qui découle du fait que le produit des deux variables formant l'interaction est symétrique

- On peut donc aussi regarder comment l'effet de l'éducation varie avec l'âge

- Selon la formule, on a
  effet de l'éducation = $C + Dx_1$
  ou alternativement
  effet de l'éducation = $\beta_2 + \beta_3 x_2$
- Dans notre cas, l'effet de l'éducation est donné par
  2084 + (130 x âge)
- Si on applique cette formule à différentes valeurs d'âge, on obtient le tableau suivant

Effet de l'éducation sur le revenu à différents âges

Âge	*Effet de l'éducation*
20	4684
30	5984
40	7284
50	8584
60	9884

- Ainsi, à 30 ans, chaque degré d'éducation supplémentaire équivaut à un revenu plus élevé en moyenne de 5'984 Frs
- Alors qu'à 60 ans, chaque degré d'éducation supplémentaire rapporte en moyenne 9'884 Frs de revenu supplémentaire
- Ce tableau indique que plus l'âge augmente, plus l'effet de l'éducation sur le revenu est grand

- Comment interpréter un tel phénomène?
- Une première hypothèse pourrait être que plus on avance en âge et progresse dans sa carrière professionnelle, plus le « profit » que l'on peut tirer d'un niveau de formation plus élevé est grand
- Une seconde hypothèse ferait intervenir l'idée de cohorte:
  pour les cohortes plus récentes, un niveau d'éducation plus élevé apporte un bénéfice moindre en termes de revenu

3. Interaction entre 1 variable quantitative (intervalle) et 1 variable catégorielle

- Considérons à titre d'illustration les deux variables âge et sexe, comme facteurs conditionnant le revenu du travail

- Faisons l'hypothèse que les femmes, au fur et à mesure qu'elles avancent dans leurs parcours professionnels bénéficient moins que les hommes d'opportunités de promotion et d'avantages au titre de l'ancienneté
- Il en résulte que l'effet de l'âge sur le revenu sera plus faible parmi les femmes que parmi les hommes
- Si l'hypothèse est correcte, il y a effet d'interaction entre âge et genre

- Pour tester l'hypothèse, suivons le même principe que précédemment
- Introduisons dans un modèle de régression, en plus des deux variables âge et sexe, un terme d'interaction constitué du produit âge x sexe

- Le modèle a dès lors la forme suivante

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Avec

x_1 = âge

x_2 = sexe

$x_1 x_2$ = âge x sexe

➢Sexe étant codé en une variable dummy

Syntaxe

1) Création du terme d'interaction

```
compute ageXsex=age05*femmes.
exe.
```

2) Modèle de régression

```
regression
   /missing listwise
   /statistics defaults ci change
   /noorigin
   /dependent i05wy
   /method=enter age05 femmes
   /method=enter ageXsex.
```

Coefficients

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	29794.678	2743.521		10.860	.000	24415.921	35173.436
	AGE05 Age durant l'année de l'interview	1285.900	61.667	.285	20.852	.000	1164.999	1406.800
	femmes	-42225.3	1575.922	-.367	-26.794	.000	-45314.97	-39135.684
2	(constante)	-3179.426	3618.987		-.879	.380	-10274.56	3915.711
	AGE05 Age durant l'année de l'interview	2099.569	85.014	.466	24.697	.000	1932.897	2266.242
	femmes	24275.721	5130.567	.211	4.732	.000	14217.084	34334.358
	ageXsex	-1640.588	120.716	-.632	-13.590	.000	-1877.256	-1403.920

Interprétation

1) Examiner la significativité statistique du terme d'interaction
   - Ici, elle est inférieure à .05, i.e. il y a interaction
2) Interpréter les coefficients
   - Quand dans le terme d'interaction figure une variable dummy, les différents coefficients (celui du terme d'interaction et ceux des deux variables formant l'interaction) prennent un sens précis

- Le coefficient pour la variable âge (2100 en arrondissant) indique l'effet de l'âge sur le revenu pour la catégorie de référence de la variable femmes, c'est-à-dire pour les hommes
- Donc, chez les hommes, chaque année d'âge supplémentaire est associée à une augmentation moyenne du revenu de 2100 Frs
- Une variation qui est statistiquement significative ($p<.001$)

- Le coefficient pour le terme d'interaction (-1641) représente la différence de l'effet de l'âge sur le revenu entre les hommes et les femmes
- Pour les femmes, l'effet de l'âge sur le revenu est donc de

  2100 + (-1641) = 460

- Pour les femmes, chaque année d'âge supplémentaire est associée avec un accroissement moyen du revenu de 460 Frs
- Autrement dit, l'effet de l'âge est environ 4.5 fois plus faible chez les femmes que chez les hommes

3) A l'inverse, on peut aussi interpréter l'interaction en référence à la manière dont l'effet de genre varie en fonction de l'âge

- Le coefficient pour femmes (24276) indique que les femmes (variable dummy=1) ayant 0 ans d'âge ont un revenu supérieur de 24'276 Frs que les hommes (catégorie de référence) ayant le même âge
- Comme personne dans l'échantillon n'est âgé de 0 ans (et pour cause!), ce coefficient n'est ici pas interprété

- Mais regardons ce qui se passe pour les personnes âgées de 30 ans
- L'écart de revenu des femmes par rapport aux hommes est de

  24276 + (-1641 x 30) = -24954

- Donc, à l'âge de 30 ans, les femmes ont en moyenne un revenu inférieur de 24'954 Frs par rapport aux hommes du même âge

- Que se passe-t-il à 60 ans?
- L'écart de revenu est de

  24276 + (-1641 x 60) = -72184 Frs

- Autrement dit, à l'âge de 60 ans, les femmes ont en moyenne un revenu inférieur de 72'184 Frs par rapport aux hommes du même âge
- Ainsi, l'écart de revenu entre genres augmente avec l'âge

- Une explication pourrait être que les femmes bénéficient moins des possibilités de promotion liée à l'ancienneté (en raison notamment de carrières professionnelles interrompues ou du fait qu'elles exercent beaucoup plus fréquemment un emploi à temps partiel)

– Autre explication possible, cette fois en termes de cohortes:
  les cohortes plus récentes de femmes ont un différentiel de revenu d'avec les hommes plus faible que les cohortes plus anciennes, par exemple parce qu'elles sont plus formées que leurs devancières

4. Interaction entre deux variables catégorielles

- Considérons à titre d'illustration les deux variables sexe et nationalité
- Ces deux variables étant conçues comme des facteurs conditionnant le revenu du travail
- Par souci de simplification, on ne distingue que deux catégories pour la nationalité: suisse versus étrangère

- Faisons l'hypothèse que les femmes étrangères exercent des emplois dans des professions particulièrement peu valorisées et offrant des salaires particulièrement modestes
- L'écart de revenu du travail entre genres est dès lors plus fort parmi la population étrangère que parmi la population suisse
- Si l'hypothèse est correcte, il y a effet d'interaction entre genre et nationalité

- Pour tester cette hypothèse, suivons le même principe que précédemment
- Introduisons dans un modèle de régression, en plus des deux variables indépendantes sexe et nationalité – toutes les deux sous la forme de variables dummies -, un terme d'interaction constitué du produit sexe x nationalité

- Le modèle a dès lors la forme suivante

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Avec

x_1 = sexe

x_2 = nationalité

$x_1 x_2$ = sexe x nationalité

  - Sexe et nationalité étant codés en variables dummies

Syntaxe

1) Création de la variable dummy pour nationalité, puis du terme d'interaction

```
recode nat3 (1=0) (2,3=1) into
   etranger.
exe.

compute sexXnat=femmes*etranger.
   exe.
```

2) Modèle de régression

```
regression
   /missing listwise
   /statistics defaults ci change
   /noorigin
   /dependent i05wy
   /method=enter femmes etranger
   /method=enter sexXnat.
```

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	81623.584	1230.407		66.339	.000	79211.332	84035.837
	femmes	-42144.6	1657.140	-.366	-25.432	.000	-45393.44	-38895.689
	etranger	2424.358	2702.645	.013	.897	.370	-2874.263	7722.978
2	(constante)	81972.615	1265.374		64.781	.000	79491.809	84453.420
	femmes	-42813.0	1751.130	-.372	-24.449	.000	-46246.15	-39379.858
	etranger	-571.979	3707.513	-.003	-.154	.877	-7840.674	6696.716
	sexXnat	6393.402	5415.684	.024	1.181	.238	-4224.218	17011.022

Interprétation

1) Regarder la significativité statistique du terme d'interaction
   - Ici, il est supérieur à .05; il n'y a donc pas d'interaction
   - Pour montrer comment on interprète les coefficients, poursuivons néanmoins la démarche

2) Interpréter les coefficients

- Quand dans le terme d'interaction figure deux variable dummies, les différents coefficients (celui du terme d'interaction et ceux des variables formant l'interaction) prennent un sens précis

- Le coefficient pour la variable femmes (-42813) mesure l'écart des femmes par rapport aux hommes pour les observations appartenant à la catégorie de référence pour la variable nationalité (code 0), i.e. pour les personnes de nationalité suisse
- Autrement dit, les femmes suisses gagnent en moyenne 42'813 Frs de moins que les hommes suisses
- Un écart qui est très significatif ($p<.001$)

- Le coefficient pour le terme d'interaction (6393) mesure la différence dans les écarts entre genres selon que l'on considère la population de nationalité suisse ou la population de nationalité étrangère
- L'écart de revenu des femmes de nationalité étrangère par rapport aux hommes de nationalité étrangère est de

  -42813 + 6393 = -36'420 Frs

- L'écart de revenu entre les genres ne varie donc guère selon que l'on considère la population suisse ou étrangère
  Dans le premier cas, l'écart est de 42'813 Frs en faveur des hommes
  Dans le second, il est de 36'420 Frs en faveur des hommes
- La différence entre ces écarts, indiquée par le coefficient d'interaction (6393), est de faible ampleur d'un point de vue substantiel et elle est statistiquement non significative

3) Prenons la relation inverse, c'est-à-dire les différences de revenu entre nationalité en fonction du sexe

- Le coefficient pour la variable etranger (-572) mesure l'écart des personnes de nationalité étrangère par rapport aux personnes de nationalité suisse pour les observations appartenant à la catégorie de référence pour la variable femmes (code 0), i.e. pour les hommes

- Autrement dit, les hommes étrangers gagnent en moyenne 572 Frs de moins que les hommes suisses
- Un écart qui n'est pas statistiquement significatif (p=.877)

- Le coefficient pour le terme d'interaction (6393) mesure la différence dans les écarts entre suisses et étrangers selon que l'on considère les hommes ou les femmes
- Les femmes de nationalité étrangère gagnent en moyenne

  -572 + 6393 = 5'821 Frs

  de plus que les femmes de nationalité suisse

- L'écart de revenu entre Suisses et Etrangers ne varie donc guère selon que l'on considère les hommes ou les femmes
  Il est dans le premier cas de 572 Frs en défaveur des personnes de nationalité étrangère
  Il est dans le second de 5'821 Frs en faveur des personnes de nationalité étrangère
- La différence entre ces écarts, indiquée par le coefficient d'interaction (6393), est de faible ampleur d'un point de vue substantiel et elle est statistiquement non significative

4) En synthèse

- Il existe un important différentiel de revenu entre hommes et femmes, au détriment de ces dernières
- Et ce différentiel ne varie pas de manière substantielle en fonction de la nationalité suisse ou étrangère