

Corrigé de l'exercice 6

Statistique II – SP08

Enseignant: Jean-François Bickel

Première étape

- Construire les variables dummies nécessaires pour l'analyse
- Créer le terme d'interaction (une variable qui est le produit du niveau d'éducation et de femmes)

Syntaxe

```
recode sex (1=0) (2=1) into femmes.  
recode nat2 (2=1) (1=0) into etranger.  
recode p05w39 (1=1) (2=0) into emploiTP.  
exe.
```

```
compute  educXfem=educat05*femmes.  
exe.
```

Deuxième étape

- S'assurer que l'ensemble de l'analyse s'effectue sur un nombre constant d'observations
 - cf. le document « Questions pratique 3: maintenir constant le nombre d'observations »
- Ici, on procède via la construction d'une variable filtre, qui sélectionne les observations n'ayant aucune donnée manquante pour l'ensemble des variables considérées

Syntaxe

```
compute nbmanque=nmiss(i05wy, femmes,  
educat05, educXfem, age05, etranger,  
emploiTP).  
exe.
```

```
compute nomanque=0.  
if (nbmanque=0) nomanque=1.  
exe.
```

```
filter by nomanque.
```

- Alternativement, il est possible de procéder à une unique analyse de régression (une unique commande **regression**) dans laquelle les variables indépendantes sont successivement introduites
- Cf. syntaxe page suivante

Syntaxe

```
regression  
  /missing listwise  
  /statistics defaults ci change  
  /noorigin  
  /dependent i05wy  
  /method=enter femmes educat05  
  /method=enter educXfem  
  /method=enter age05 etranger emploiTP.
```

Troisième étape

- Analyse de régression avec éducation et sexe sans effet d'interaction

Syntaxe

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

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.515 ^a	.265	.265	49811.830

Changement dans les statistiques				
Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
.265	703.765	2	3902	.000

Anova

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	3.49E+012	2	1.746E+012	703.765	.000 ^a
	Résidu	9.68E+012	3902	2481218427		
	Total	1.32E+013	3904			

Commentaires

- La variation de R^2 est la différence de la Réduction Proportionnelle de l'Erreur (i.e. de R^2) entre le modèle complet (avec âge et éducation) et le modèle restreint (sans ces variables)
- Cette différence fait l'objet d'un test de F
- Observons ici que la valeur de F (703.765) est aussi la valeur indiquée sous F dans le tableau ANOVA

- Rappelons que le test qui figure dans le tableau ANOVA vise à déterminer la probabilité que l'ensemble des coefficients β du modèle considéré soient égal à zéro
 - compte tenu des coefficients (b) observés
- Dans le cas présent, la valeur de F pour ce dernier test est égal à la valeur de F dans le test de la différence de R^2 entre modèle complet et modèle restreint, car ce dernier équivaut à un modèle vide (=sans variable indépendante)

- Globalement, l'introduction de l'âge et de l'éducation améliore le modèle, le rend plus proche des valeurs observées
 - autrement dit, la somme des erreurs de prédiction est moindre
 - on pourrait aussi dire que la prise en compte de l'âge et de l'éducation apporte une information supplémentaire effective pour décrire le phénomène « revenu du travail »

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)	42162.397	2009.953		20.977	.000	38221.739	46103.054
	femmes	-38082.3	1607.449	-.328	-23.691	.000	-41233.77	-34930.732
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	7219.054	279.149	.358	25.861	.000	6671.763	7766.345

Commentaires

- Ce premier modèle n'apporte pas en soi d'informations nouvelles par rapport à ce qu'avaient mis en évidence les exercices précédents
- Mais donne (rappelle) l'information générale de base sur la manière dont les deux facteurs (sexe et éducation) sur lesquels se centre l'analyse conditionnent le revenu

- Les coefficients pour les deux variables sont statistiquement significatifs et indiquent que
 - i. les femmes, à niveau d'éducation égal, gagnent moins que les hommes
 - globalement, l'écart est de quelque 38'000 Frs
 - ii. net de l'effet du genre, le revenu du travail augmente avec le niveau d'éducation
 - chaque degré supplémentaire rapportent en moyenne 7'219 Frs de plus

Quatrième étape

- On introduit le terme d'interaction
 - i.e. le produit sexe x éducation (variable **educXfem**)

Syntaxe

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

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.515 ^a	.265	.265	49811.830
2	.529 ^b	.280	.279	49327.404

Changement dans les statistiques				
Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
.265	703.765	2	3902	.000
.014	78.017	1	3901	.000

Commentaires

- La mesure de Réduction Proportionnelle de l'Erreur ($=R^2$) augmente de .265 à .280, soit une différence de .014 (arrondis!!!)
- Cette différence est significative ($p < .001$)
- L'introduction du terme d'interaction améliore donc le modèle, le rend plus proche des valeurs observées
 - i.e. la somme des erreurs de prédiction est moindre

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)	42162.397	2009.953		20.977	.000	38221.739	46103.054
	femmes	-38082.3	1607.449	-.328	-23.691	.000	-41233.77	-34930.732
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	7219.054	279.149	.358	25.861	.000	6671.763	7766.345
2	(constante)	28552.285	2517.145		11.343	.000	23617.240	33487.331
	femmes	-10942.5	3460.493	-.094	-3.162	.002	-17727.08	-4157.988
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	9512.772	379.278	.471	25.081	.000	8769.169	10256.374
	educXfem	-4892.854	553.948	-.274	-8.833	.000	-5978.909	-3806.799

Commentaires

- L'interaction éducation x femmes combine une variable intervalle (éducation) et une variable catégorielle codée sous la forme d'une variable dummy
 - i.e. une variable binaire codée 0 et 1
- Dans ce cas, le coefficient des variables composant l'interaction ont une signification précise

- Le **coefficient pour éducation** (9613 en arrondissant à l'unité) mesure l'effet de ce facteur pour la catégorie codée 0 (=catégorie de référence) de la variable dummy
- Donc, le coefficient de 9613 signifie que **pour les hommes**, chaque degré supplémentaire d'éducation rapporte en moyenne 9'613 Frs de plus. Il s'agit d'un effet très significatif ($p < .001$)

- Le **coefficient pour femmes** (-10943 en arrondissant à l'unité) mesure l'effet du genre pour les personnes qui, sur la variable quantitative, ont une valeur de 0
- Donc, le coefficient de -10943 indique que **parmi les personnes n'ayant pas achevé l'école obligatoire**, les femmes gagnent en moyenne 10'943 Frs de moins que les hommes
- Cet effet est très significatif ($p=.002$).

- Considérons maintenant le **coefficient d'interaction** (-4893 en arrondissant à l'unité)
- L'interaction étant un produit, **deux modes de lecture** sont possibles

1. Dans un **point de vue centré sur l'éducation** et son effet, le coefficient d'interaction mesure l'effet additionnel de l'éducation pour les femmes par rapport à ce qu'est cet effet (de l'éducation) pour les hommes
 - L'effet de l'éducation **pour les femmes** est donc de
$$9613 + (-4893) = 4720$$

- Autrement dit, pour les femmes, chaque degré d'éducation supplémentaire rapporte en moyenne 4'720 Frs
- C'est donc moins que les 9'613 Frs des hommes
- Et cette différence (entre hommes et femmes) est statistiquement significative
 - c'est précisément ce qu'indique le résultat du test pour le coefficient d'interaction

2. Dans un **point de vue centré sur le genre** et son effet, le coefficient d'interaction mesure de combien cet effet croît avec chaque degré d'éducation supplémentaire
 - Dans le cas présent, il s'agit d'une croissance négative (le coefficient est négatif) : **pour chaque degré supplémentaire d'éducation**, l'écart de genre « croît » donc en moyenne de -4'893 Frs

- Si on considère les personnes ayant fait un apprentissage (educat05=4), la différence de genre est de
 $-10943 + (-4893 \times 4) = -30515$
- Autrement dit, les femmes ayant ce niveau d'éducation gagnent en moyenne 30'515 Frs de moins que leurs équivalents hommes

- Si on considère les universitaires (educat05=10), l'écart de revenu est de $-10943 + (-4893 \times 10) = -59873$
- Autrement dit, les femmes universitaires gagnent en moyenne 59'873 Frs de moins que les hommes universitaires

Raison statistique et raison sociologique

- On peut donc sur la base des résultats de la régression, aussi bien dire
 - a) que l'effet de l'éducation est plus fort pour les hommes que pour les femmes
 - b) que l'écart entre genres augmente quand le niveau d'éducation croît

- Ce sont là deux interprétations possibles des mêmes « données », de la même analyse
- Comment choisir une des interprétations ?
Peut-on choisir une des interprétations ?

1. Statistiquement parlant,
 - les deux interprétations sont correctes
 - il n'y a rien qui permette de décider pour l'une ou pour l'autre
 - le fait que la première interprétation (a) semble plus simple et directe (elle implique moins de calcul) n'est bien sûr pas une raison suffisante!

2. Chaque interprétation renvoie à un point de vue différent sur la réalité
 - Chacune pose une assertion différente sur le processus qui a produit ou qui a été susceptible de produire le résultat observé

- La première assertion (a) concerne le processus par lequel le niveau d'éducation contribue à déterminer le niveau de revenu du travail
- Et pose que ce processus « se passe » différemment pour les hommes et pour les femmes, qu'il est plus « profitable » pour les premiers que pour les secondes

- Autrement dit, le revenu du travail est considéré comme
- étant fonction du niveau d'éducation
 - qui est la variable qualifiée de principale
- et cette relation varie en fonction du genre
 - qui est la variable qualifiée de médiatrice

- La seconde assertion (b) concerne le processus par lequel le statut d'homme ou de femme contribue à déterminer le niveau de revenu du travail
- Et pose que ce processus « se passe » différemment selon le niveau d'éducation, plus précisément que l'écart de revenu entre les genres se creuse avec l'élévation du niveau d'éducation

- Autrement dit, le revenu du travail est considéré comme
- étant fonction du genre
 - variable principale
- et cette relation est différente selon le niveau d'éducation
 - variable médiatrice

- C'est la teneur logique de ces deux types d'assertions, leur plausibilité, leur adéquation avec d'autres éléments du « stock » de connaissances, leur cohérence avec la ou les questions qui guident la recherche qui doivent être soumis à examen
- C'est cet examen qui **peut** conduire à privilégier l'une ou l'autre, ou à reconnaître l'indécidabilité (au moins en l'état) de la question

- L'interprétation privilégiée peut ainsi dépendre de l'intérêt ou question de départ
 - par exemple, si on s'intéresse avant tout à mesurer l'influence de l'éducation sur le revenu, le choix se portera sur (a), l'apport étant de montrer que cet effet varie pour les hommes et pour les femmes
- Elle peut dépendre de connaissances existant sur les processus en jeu
 - par exemple issues de travaux portant sur le fonctionnement du marché du travail, sur les mécanismes de discrimination, etc.

- Parfois, le choix peut reposer sur des arguments logiques ou de plausibilité
 - par exemple, on peut être en mesure de formuler des hypothèses plausibles, à défaut d'être démontrées, sur le fait que les femmes ont plus de difficultés à faire valoir l'éducation acquise que les hommes (interprétation a), et à l'inverse peiner à trouver des raisons pouvant expliquer pourquoi l'écart entre genres croît avec le niveau d'éducation (interprétation b)

- Parfois, certains éléments du raisonnement peuvent être corroborés, ou à l'inverse réfutés, par un « retour » aux données et des analyses statistiques supplémentaires
- D'autres éléments (ou les mêmes) peuvent reposer sur l'acquis d'autres chercheurs, ou du même chercheur dans des recherches antérieures ou parallèles, que ces recherches soient de nature statistique ou aient recours à d'autres méthodologies

- D'autres enfin peuvent être de pures conjectures
- Mais quoiqu'il en soit, dans la recherche sur le monde social, **le raisonnement statistique ne se suffit jamais à lui-même**

Cinquième étape

- On introduit les variables de contrôle (âge, nationalité, type d'emploi)

Syntaxe

```
regression  
  /missing listwise  
  /statistics defaults ci change  
  /noorigin  
  /dependent i05wy  
  /method=enter femmes educat05  
  /method=enter educXfem  
  /method=enter age05 etranger emploiTP.
```

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.515 ^a	.265	.265	49811.830
2	.529 ^b	.280	.279	49327.404
3	.611 ^c	.374	.373	46007.695

Changement dans les statistiques				
Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
.265	703.765	2	3902	.000
.014	78.017	1	3901	.000
.094	195.423	3	3898	.000

Commentaires

- La mesure de Réduction Proportionnelle de l'Erreur ($=R^2$) augmente de .280 à .374, soit une différence de .094
- Cette différence est significative ($p<.001$)
- L'introduction des variables âge, nationalité et type d'emploi améliore donc le modèle, le rend plus proche des valeurs observées
 - i.e. la somme des erreurs de prédiction est moindre

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)	42162.397	2009.953		20.977	.000	38221.739	46103.054
	femmes	-38082.3	1607.449	-.328	-23.691	.000	-41233.77	-34930.732
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	7219.054	279.149	.358	25.861	.000	6671.763	7766.345
2	(constante)	28552.285	2517.145		11.343	.000	23617.240	33487.331
	femmes	-10942.5	3460.493	-.094	-3.162	.002	-17727.08	-4157.988
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	9512.772	379.278	.471	25.081	.000	8769.169	10256.374
	educXfem	-4892.854	553.948	-.274	-8.833	.000	-5978.909	-3806.799
3	(constante)	-2899.967	3134.603		-.925	.355	-9045.584	3245.649
	femmes	3066.447	3375.834	.026	.908	.364	-3552.121	9685.015
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	8502.733	363.333	.421	23.402	.000	7790.392	9215.075
	educXfem	-4229.474	518.740	-.237	-8.153	.000	-5246.501	-3212.447
	AGE05 Age durant l'année de l'interview	1041.559	61.838	.222	16.843	.000	920.322	1162.796
	etranger	2947.400	2395.654	.016	1.230	.219	-1749.454	7644.254
	emploiTP	-34007.9	1780.277	-.290	-19.103	.000	-37498.25	-30517.520

Commentaires

- Ces variables de contrôle doivent être regardée
 - i. pour elles-mêmes : exerce-t-elle un effet sur la variable dépendante et si oui lequel ?
 - ii. pour leurs effets éventuels sur les relations déjà observées et mises en évidence

- L'âge exerce un effet positif sur le revenu du travail, qui augmente en moyenne de 1'042 Frs pour chaque année d'âge supplémentaire, toute chose égale par ailleurs
- Compte tenu des autres variables du modèle, la nationalité n'exerce pas d'effet sur le revenu
 - Le coefficient de 2947 est suffisamment petit pour qu'on ne puisse raisonnablement exclure que le « vrai » coefficient soit égal à 0 (effet nul)

- Le type d'emploi exerce un fort impact ($p < .001$)
- La valeur du coefficient qui est de 34008 (arrondi à l'unité) indique que les personnes à plein temps gagnent en moyenne 34'008 Frs de plus que les personnes à temps partiel, les autres variables étant maintenues constantes

- L'introduction des variables de contrôle ne modifie pas substantiellement le sens et la force des effets pour éducation d'une part, pour l'interaction sexe x éducation d'autre part
- Il n'en va pas de même pour le coefficient de la variable femmes, dont la valeur passe -10943 à 3066 et qui est maintenant non significatif

- Ce coefficient, compte tenu de la présence de l'interaction, réfère à la différence des femmes par rapport aux hommes pour les personnes ayant une valeur de 0 sur la variable éducation
- Pour ces personnes là, prendre en compte le type d'emploi modifie l'ampleur des différences, qui deviennent non significatives

- A moins d'avoir à disposition des informations supplémentaire, ce fait est trop spécifique pour qu'on en propose une interprétation