Corrigé de l’exercice 4

Jean-François Bickel

Statistique II – SP08

Préambule: création des variables dummies à partir des variables catégorielles originelles

Syntaxe

```
recode sex (1=0) (2=1) into femmes.
recode sex (1=1) (2=0) into hommes.
exe.

recode nat3 (1=1) (2,3=0) into natch.
recode nat3 (2=1) (1,3=0) into nateurop.
recode nat3 (3=1) (1,2=0) into natautre.
exe.
```

Etape 1: introduction du genre parmi les variables indépendantes

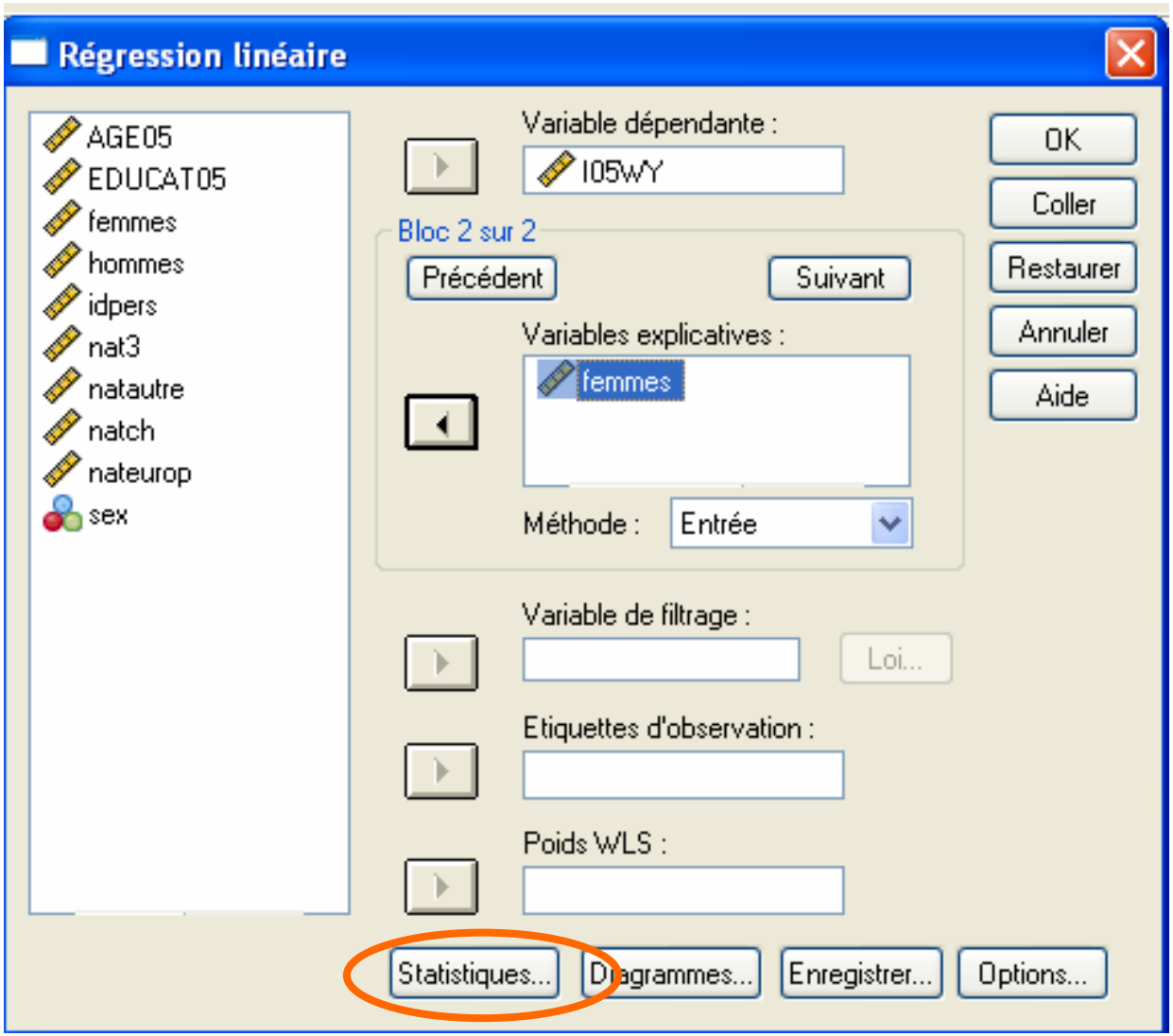
Syntaxe

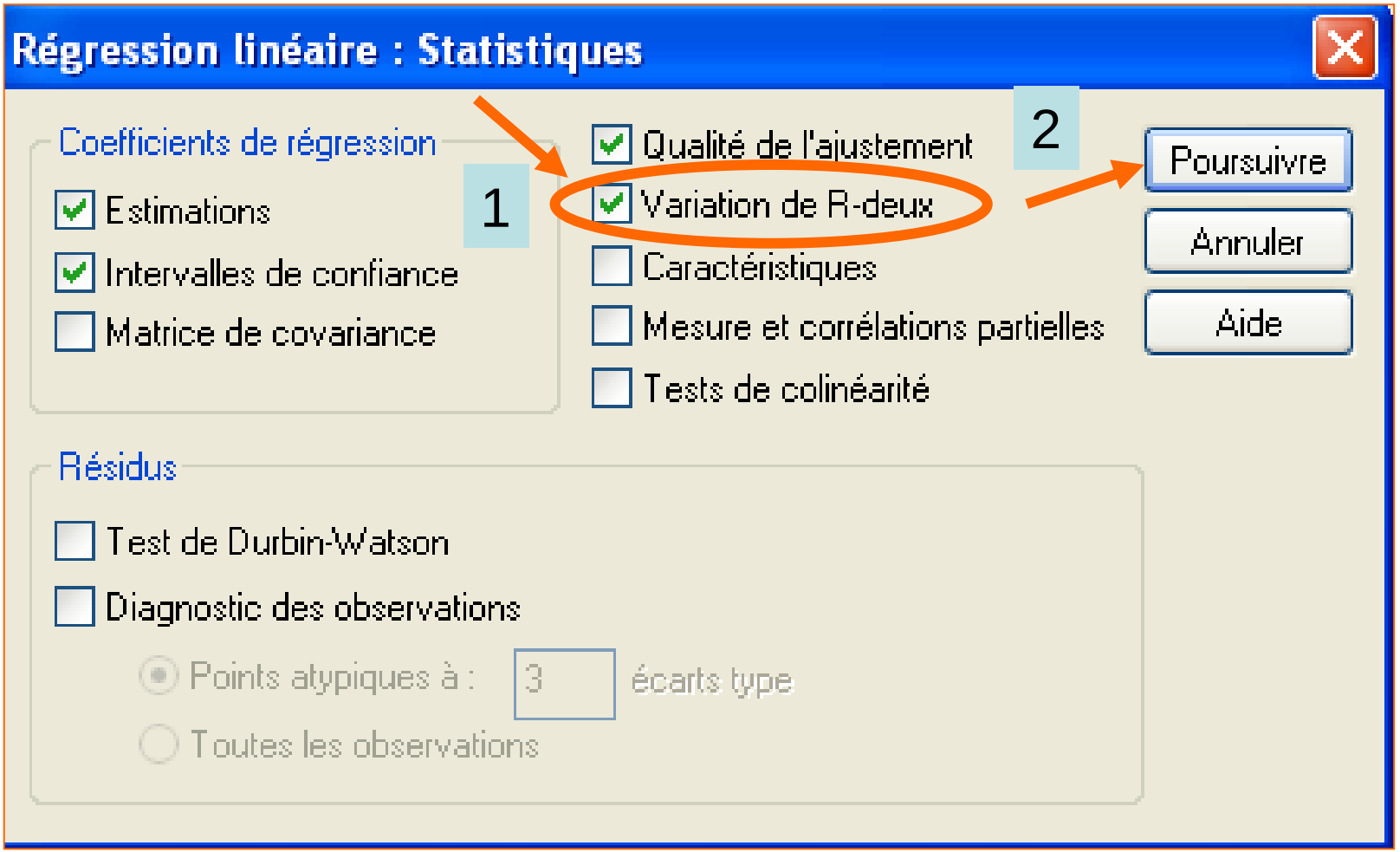
```
regression
 /missing listwise
 /statistics defaults ci change
 /noorigin
 /dependent i05wy
 /method=enter age05 educat05
 /method=enter femmes.
```

- Observons dans la syntaxe la séquence
  `/method=enter age05 educat05`
  `/method=enter femmes.`
- Cette séquence, combinée avec mot-clé **`change`**
  permet que SPSS calcule l'amélioration du modèle résultant de l'introduction de la variable femmes (modèle complet) par rapport au modèle sans cette variable (modèle réduit)

- Rappelons que ce calcul se base sur la différence dans la réduction proportionnelle de l'erreur (R^2) entre le modèle sans la variable en question (=modèle réduit) et le modèle avec cette variable (=modèle complet)

Avec l'interface graphique…




Régression linéaire : Statistiques
Coefficients de régression
Estimations
Intervalles de confiance
Matrice de covariance
1
Qualité de l'ajustement
Variation de R-deux
Caractéristiques
Mesure et corrélations partielles
Tests de colinéarité
2
Poursuivre
Annuler
Aide
Résidus
Test de Durbin-Watson
Diagnostic des observations
Points atypiques à :
3
écarts type
Toutes les observations

- Avec l'interface graphique, on demande à SPSS de calculer la différence de R^2 et le test qui lui est associé de la manière suivante
- Dans la fenêtre principale du menu régression, sélectionner l'option **Statistiques**
- Dans la fenêtre **Statistiques**, cocher la case **Variation de R-deux**

Ce tableau indique qu'il s'agit d'une analyse (=1 commande régression), composée de deux modèles successifs, les variables indépendantes incluses étant rappelées

Variables introduites/éliminées[b]

Modèle	Variables introduites	Variables éliminées	Méthode
1	EDUCAT05, AGE05[a]	.	Introduire
2	femmes[a]	.	Introduire

a. Toutes variables requises introduites

b. Variable dépendante : I05WY

Résumé du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.450[a]	.202	.202	51393.316
2	.555[b]	.308	.308	47862.918

a. Valeurs prédites : (constantes), EDUCAT05, AGE05

b. Valeurs prédites : (constantes), EDUCAT05, AGE05, femmes

Changement dans les statistiques				
Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
.202	530.579	2	4183	.000
.106	640.839	1	4182	.000

- Dans la première partie du tableau, on s'intéresse uniquement au R^2; on observe qu'il augmente substantiellement d'un modèle à l'autre, passant de .202 à .308
- Dans la seconde partie du tableau (bas), le premier chiffre de la seconde ligne représente la différence entre les 2 modèles (.308-.202=.106)

- Cette différence, qui est aussi exprimée par la variation de la valeur du F (640.839) est significative
- Autrement dit, l'introduction de la variable femmes a significativement amélioré la qualité du modèle

ANOVA[c]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	2.80E+012	2	1.401E+012	530.579	.000[a]
	Résidu	1.10E+013	4183	2641272901		
	Total	1.39E+013	4185			
2	Régression	4.27E+012	3	1.424E+012	621.438	.000[b]
	Résidu	9.58E+012	4182	2290858912		
	Total	1.39E+013	4185			

a. Valeurs prédites : (constantes), EDUCAT05, AGE05

b. Valeurs prédites : (constantes), EDUCAT05, AGE05, femmes

c. Variable dépendante : I05WY

- Le tableau ANOVA nous indique le résultat du test selon lequel tous les coefficients des variables indépendantes du modèle considéré valent zéro (versus au moins un de ceux-ci est différent de zéro)
- Pour les deux modèles considérés, ce test est très significatif ($p<.001$). On est assuré qu'au moins un des coefficients est différent de zéro

- Ceci est, à dire vrai, peu informatif!
- Ce qui devrait nous interroger est si d'aventure le test indiquait que l'hypothèse H_0 ne peut être rejetée (= si p>.05), ce qui impliquerait qu'il n'y a rien dans le modèle qui explique quoi que ce soit…

Coefficients

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Signification
1	(constante)	-14839.2	2795.291		-5.309	.000
	AGE05	889.857	64.143	.198	13.873	.000
	EDUCAT05	7109.914	282.245	.359	25.191	.000
2	(constante)	7332.833	2746.661		2.670	.008
	AGE05	937.957	59.767	.208	15.694	.000
	EDUCAT05	6263.755	264.973	.316	23.639	.000
	femmes	-37775.5	1492.229	-.328	-25.315	.000

a. Variable dépendante : I05WY

Modèle		Intervalle de confiance à 95% de B	
		Borne inférieure	Borne supérieure
1	(constante)	-20319.48	-9358.963
	AGE05	764.104	1015.611
	EDUCAT05	6556.564	7663.264
2	(constante)	1947.918	12717.748
	AGE05	820.783	1055.132
	EDUCAT05	5744.266	6783.243
	femmes	-40701.06	-34849.931

- Le coefficient pour les femmes (-37776 en arrondissant) veut dire que celles-ci ont un revenu moyen inférieur de 37776 Frs par rapport au revenu moyen des hommes, et ce en contrôlant par âge et niveau d'éducation
- Cet effet du genre est très significatif ($p<.001$)

- Il y a 95% de chances que la différence de revenu entre hommes et femmes ici estimée à 37776 soit dans la réalité comprise entre 40701 et 34850 Frs

- La comparaison des modèles 1 et 2 montre aussi une réduction du coefficient pour l'éducation qui passe 7110 à 6262, soit une réduction de plus de 10% (12% exactement)
- Comment expliquer cette différence?
  - Les femmes ont un revenu plus faible que les hommes
  - Une partie de cet écart vient du niveau d'éducation plus faible des femmes

- Cette composante de l'effet du genre est incluse dans la mesure de l'effet de l'éducation du modèle 1 (i.e. quand la variable genre n'est pas elle-même incluse)
- L'introduction du genre dans le modèle 2 supprime de l'effet de l'éducation la composante en question, liée à l'association entre genre et éducation
- L'effet de l'éducation qui est mesuré est ici contrôlé par le genre, c'est-à-dire « net » de l'association entre genre et éducation

Etape 2: introduction de la nationalité parmi les variables indépendantes

Syntaxe

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

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.450[a]	.202	.202	51397.872
2	.555[b]	.308	.308	47868.508
3	.561[c]	.314	.313	47674.830

Changement dans les statistiques				
Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
.202	530.507	2	4182	.000
.106	640.415	1	4181	.000
.006	18.020	2	4179	.000

- L’introduction de la variable nationalité (sous la forme de 2 variables dummies) améliorent un peu le R^2 (différence de .006). Cette amélioration est significative ($p<.001$)
- Le tableau suivant « ANOVA » n’apporte pas d’information substantielle

ANOVA[d]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	2.80E+012	2	1.401E+012	530.507	.000[a]
	Résidu	1.10E+013	4182	2641741241		
	Total	1.39E+013	4184			
2	Régression	4.27E+012	3	1.423E+012	621.218	.000[b]
	Résidu	9.58E+012	4181	2291394011		
	Total	1.39E+013	4184			
3	Régression	4.35E+012	5	8.705E+011	382.973	.000[c]
	Résidu	9.50E+012	4179	2272889428		
	Total	1.39E+013	4184			

a. Valeurs prédites : (constantes), EDUCAT05, AGE05

b. Valeurs prédites : (constantes), EDUCAT05, AGE05, femmes

c. Valeurs prédites : (constantes), EDUCAT05, AGE05, femmes, natautre, nateurop

d. Variable dépendante : I05WY

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Signification
1	(constante)	-14832.2	2795.574		-5.306	.000
	AGE05	889.772	64.149	.198	13.870	.000
	EDUCAT05	7110.398	282.272	.359	25.190	.000
2	(constante)	7332.734	2746.982		2.669	.008
	AGE05	937.929	59.774	.208	15.691	.000
	EDUCAT05	6263.970	265.008	.316	23.637	.000
	femmes	-37771.9	1492.583	-.328	-25.306	.000
3	(constante)	6475.398	2759.603		2.346	.019
	AGE05	939.924	59.539	.209	15.787	.000
	EDUCAT05	6305.510	264.075	.318	23.878	.000
	femmes	-37774.1	1487.908	-.328	-25.387	.000
	nateurop	1556.693	2496.930	.008	.623	.533
	natautre	49190.281	8216.360	.077	5.987	.000

Modèle		Intervalle de confiance à 95% de B	
		Borne inférieure	Borne supérieure
1	(constante)	-20312.96	-9351.344
	AGE05	764.007	1015.537
	EDUCAT05	6556.996	7663.800
2	(constante)	1947.189	12718.278
	AGE05	820.740	1055.117
	EDUCAT05	5744.413	6783.526
	femmes	-40698.21	-34845.693
3	(constante)	1065.109	11885.687
	AGE05	823.195	1056.653
	EDUCAT05	5787.781	6823.238
	femmes	-40691.16	-34856.977
	nateurop	-3338.618	6452.004
	natautre	33081.846	65298.716

- Un rapide coup d’œil aux coefficients pour âge, éducation et genre montrent que l’introduction de la variable de nationalité n’apporte pas de modifications substantielles

- Si on examine dans le détail les dummies pour la nationalité, on remarque que les ressortissants européens ne se différencient pas significativement de ceux suisses
- l'écart de 1557 a une valeur de t de .623, p=.533
  - Voir aussi l'intervalle de confiance qui contient le zéro

- Ce qui frappe est le coefficient pour les Non-européens
- En contrôlant par âge, éducation et genre, le revenu de ceux-ci est en moyenne supérieur de 49’190 Frs à celui des Suisses
- On peine a priori à voir ce qui pourrait expliquer un tel écart

- Egalement frappant est le fait que malgré cette valeur de coefficient très élevée, la valeur du t (quoique significative) est relativement modeste (5.987)
- Ceci est dû à l'erreur standard, qui est elle aussi très grande
- Une telle configuration peut signaler un problème: d'effectif et/ou de présence de valeurs extrêmes

- Que se passe-t-il avec la nationalité « autre » ?
- Pour aller y voir, on peut sélectionner les seuls individus « nationalité autres », voir combien ils sont et quel est leur niveau de revenu

Procédure via la syntaxe

```
compute filtre=0.
if (nat3=3) filtre=1.
exe.

filter by filtre.

freq var=i05wy.
```

I05WY

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	500	1	1.9	2.9	2.9
	3460	1	1.9	2.9	5.9
	5000	1	1.9	2.9	8.8
	7200	1	1.9	2.9	11.8
	14000	1	1.9	2.9	14.7
	15940	1	1.9	2.9	17.6
	17880	1	1.9	2.9	20.6
	...				
	61600	1	1.9	2.9	82.4
	65000	1	1.9	2.9	85.3
	68900	1	1.9	2.9	88.2
	84000	1	1.9	2.9	91.2
	84500	1	1.9	2.9	94.1
	120000	1	1.9	2.9	97.1
	2000000	1	1.9	2.9	100.0
	Total	34	65.4	100.0	
Manquante	-3	17	32.7		
	-1	1	1.9		
	Total	18	34.6		
Total		52	100.0		

- Le tableau de fréquence indique que dans cette catégorie, ils sont 34 à déclarer un revenu du travail, ce qui est très peu compte tenu de la taille de l'échantillon
- On voit aussi qu'il y a parmi eux un revenu déclaré de 2'000'000 de Frs
  A l'évidence, une valeur extrême!

- Nous reprendrons cette question plus loin dans le cours, au chapitre « Diagnostics »