UNIVERSITE GASTON BERGER
L'excellence au service du développement

UFR des Lettres et Sciences humaines
Département de sociologie
Semestre 2

SOCIO532.2
STATISTIQUES & INFORMATIQUE APPLIQUÉES AUX SCIENCES SOCIALES
© El Hadj Touré, Ph.D.

LABO SPSS # 6
Régression linéaire multiple

Strictement parlant, la régression linéaire simple est appropriée lorsque les deux variables sont quantitatives. Cependant, il est possible d'inclure une VI qualitative dans un modèle de régression. La procédure est la même que pour la régression linéaire simple. La seule différence est qu'il faut vous assurer que la VI qualitative dichotomique ou dichotomisée est codée 0/1 de façon à faciliter l'interprétation des résultats de l'analyse de régression. Nous verrons donc comment approfondir l'analyse de régression linéaire simple.

Par ailleurs, que la VI soit quantitative ou dichotomique, la régression simple fournit une explication incomplète de la variation dans la VD. Pour augmenter la proportion de variance expliquée, on inclut dans le modèle d'autres VIs. On parle alors de régression linéaire multiple. Elle est utilisée lorsque l'on veut établir l'effet de deux prédicteurs ou plus (VIs) sur une VD quantitative. Précisément, la régression linéaire multiple permet de prédire, à l'aide d'une équation, un score d'une VD quantitative (Y) à partir des scores d'autres VIs (X) qui peuvent être quantitatives ou dichotomiques.

La convergence est le modèle de causalité par excellence de la régression linéaire multiple. Selon la façon dont l'analyste fait entrer les variables dans le modèle de régression multiple, on distingue deux procédures parmi les plus fréquemment utilisées : la **régression standard** et la **régression hiérarchique**. La procédure standard consiste, pour l'analyste, à entrer simultanément les variables dans l'analyse, alors que la procédure hiérarchique consiste à entrer les variables, par ordre, les unes après les autres ou par bloc. La régression standard est appropriée pour le modèle de causalité convergente. Toutefois, on peut toujours utiliser la procédure hiérarchique pour savoir de combien l'ajout d'une VI additionnelle ou d'un bloc de VIs supplémentaire augmente la proportion expliquée[1].

Dans ce labo SPSS, nous apprendrons à :

- ✓ Inclure une VI qualitative dans un modèle de régression linéaire simple,
- ✓ Procéder à l'analyse de régression linéaire multiple afin de mettre en pratique le modèle de convergence,
- ✓ Tester la multi-colinéarité et diagnostiquer un modèle de régression multiple.

[1] Une autre procédure de régression consiste, pour le logiciel, à entrer automatiquement les variables selon leur contribution statistiquement significative à l'amélioration du modèle : on parle de régression avec entrée progressive (méthode ascendante, méthode pas à pas, méthode descendante).

La démonstration sera suivie d'un exercice pratique. Auparavant, en guise de rappel et d'approfondissement, revenons sur l'analyse de régression linéaire simple.

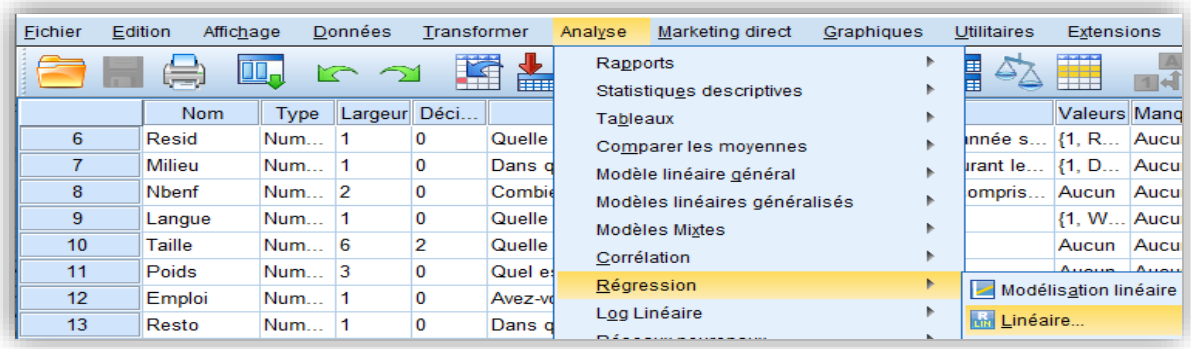
1. Régression linéaire simple : quelques éléments de rappel

1.1. Analyse des données

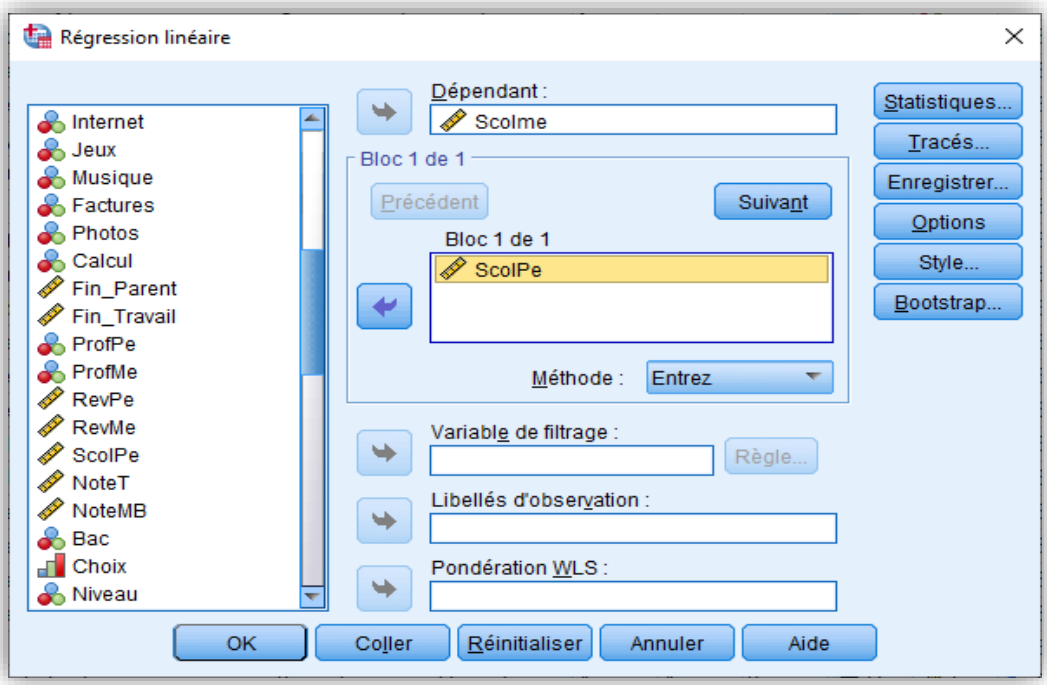
Ouvrons la base de données « Sondage_EtudiantsSocioL2_2021(labo9&10) » et revenons sur la relation entre le nombre d'années de scolarité du père (ScolPe) et le nombre d'années de scolarité de la mère (ScolMe) chez les étudiants inscrits en L2 de sociologie à l'UGB, soit deux variables quantitatives.

Scolarité du père (ScolPe) ⟶ Scolarité de la mère (ScolMe)

Analyse
Régression
Linéaire



Une fenêtre de dialogue apparaît ! Cliquez sur la variable dépendante (scolme), puis indépendante (scolpe). Contrairement à la corrélation, la régression accorde une importance à l'identification de la VD et de la VI.



Validez pour obtenir la page des résultats !

Récapitulatif des modèles[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.597[a]	.357	.350	4.770

a. Prédicteurs : (Constante), ScolPe Combien d'années de scolarité, approximativement, votre père a-t-il complétées ?

b. Variable dépendante : Scolme Combien d'années de scolarité, approximativement, votre mère a-t-elle complétées ?

Le premier tableau, « Récapitulatif des modèles », résume le modèle de la relation entre la scolarité du père et celle de la mère. Le coefficient de corrélation R est de 0,597, traduisant l'existence d'une relation positive et modérée entre ces deux variables. Le R-deux de 0,357 signifie que la scolarité du père explique une proportion de 35,7% de la variation dans la scolarité de la mère. De même, nous réduisons de 35,7% nos erreurs de prédiction de la scolarité de la mère (VD) quand nous connaissons celle du père (VI). Finalement, 64,3 % de la variation reste à expliquer.

ANOVA[a]

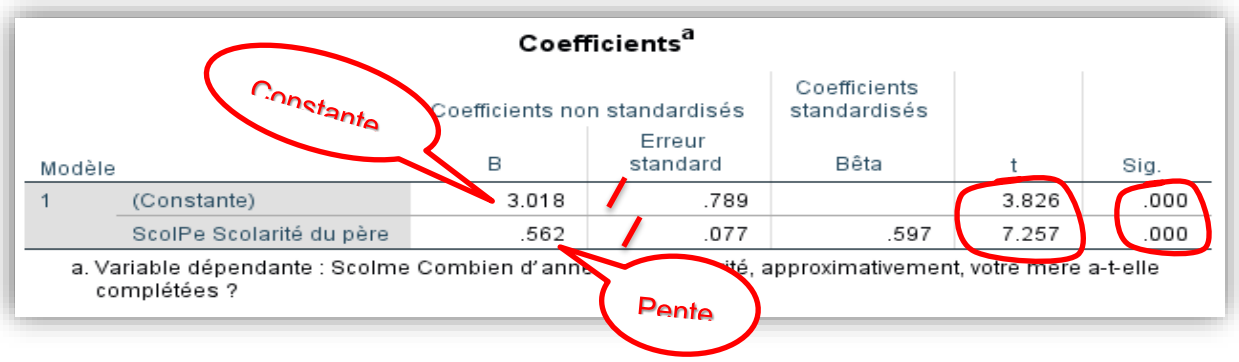
Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1198.438	1	1198.438	52.668	.000[b]
	Résidu	2161.686	95	22.755		
	Total	3360.124	96			

a. Variable dépendante : Scolme Combien d'années de scol… …ximativement, votre mère a-t-elle complétées ?

b. Prédicteurs : (Constante), ScolPe Combien d'années de scola… approximativement, votre père a-t-il complétées ?

n-2

Dans le tableau « ANOVA », SPSS présente le résultat du test de signification statistique à l'aide de l'ANOVA. Autrement dit, la valeur du coefficient R (0,597), calculée sur des données d'échantillon, est-elle significative au niveau de la population étudiée ? On constate que la valeur calculée du F est de 52,67. Or, pour 1 (k-1=2-1) degré de liberté au numérateur et 95 (n-k=97-2) degrés de liberté au dénominateur et au seuil 0,05, la valeur critique du ***F*** est de 3,94. La valeur calculée du F est donc de loin plus grande que la valeur critique. Par conséquent, nous rejetons l'hypothèse nulle voulant qu'il n'y ait pas de relation entre la scolarité du père et celle de la mère : le coefficient de corrélation n'est donc pas égal à 0. Dit autrement, la probabilité de commettre une erreur en rejetant l'hypothèse nulle est très faible puisqu'elle est inférieure à 0,05 : elle est précisément de 0,000, et donc inférieure au seuil exigeant de 0,001. Il existe une corrélation statistiquement significative entre la scolarité du père et celle de la mère dans la population étudiante étudiée.



Finalement, le troisième tableau, « Coefficients », contient les termes clés de l'**équation de la droite de régression** et leurs **degrés de signification**. En réalité, la régression consiste à calculer l'équation d'une droite, laquelle résume le mieux possible le nuage de points du diagramme. L'**intersection** ou constante (**a**) est de 3,018 alors que la **pente** ou coefficient de régression (**b**) est de 0,562.

D'une part, la constante signifie que la droite de régression coupe l'axe vertical (Y) à une scolarité (de la mère) de 3,018 années. Autrement dit, la valeur estimée de la scolarité de la mère est de 3,018 années lorsque celle du père est nulle (**ce qui n'est pas pertinent pour rendre compte de l'effet de la VI sur la VD!**).

D'autre part, la pente signifie que lorsque la scolarité du père augmente d'une unité (1 année), la scolarité de la mère augmente de 0,562 années, indiquant ainsi une relation positive. Concrètement, si on passe d'un homme qui a 0 année de scolarité à un homme qui a 1 année de scolarité, on doit s'attendre à obtenir une scolarité de la femme plus élevée de 0,562 année. Si on passe d'un homme qui a 10 années de scolarité à un homme qui en a 20 (la différence est de 10), on doit s'attendre à obtenir chez la femme une scolarité plus élevée de 10*0,562, soit 5,62 années.

Voici l'équation de la droite de régression :

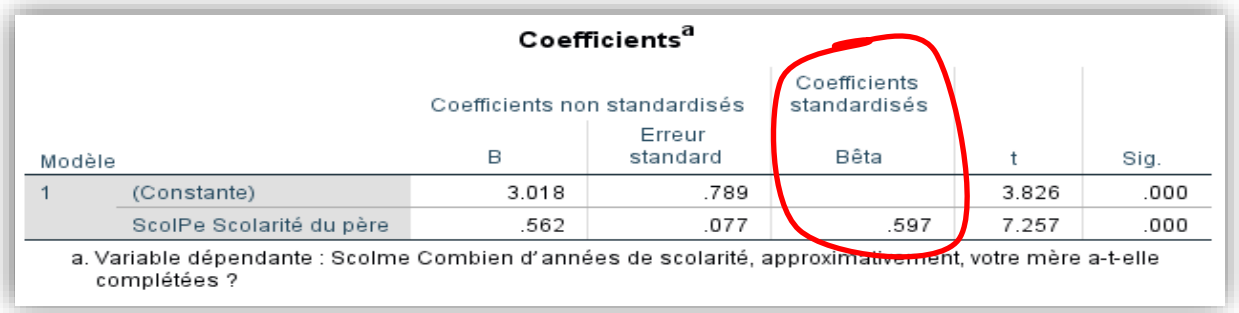
Y = a + b (X)

Scolarité **de la mère** = 3,018 + 0,562 (Scolarité **du père**)

D'une certaine manière, on peut dire que les valeurs de Y sont prédites par la combinaison d'une constante et d'une variable X. Ainsi, connaissant la scolarité du père d'un étudiant qui ne figure pas dans l'échantillon, on peut estimer ou prédire la scolarité de la mère. Par exemple, pour un père qui a une scolarité de 21 années (X=1), on peut prédire que la mère aura 14,8 années.

$$\hat{Y} = 3{,}018 + 0{,}562\ (21) = 3{,}018 + 11{,}80 = 14{,}8$$

1.2. Le coefficient de régression bêta



Le coefficient de régression **b** (0,562) est un coefficient non standardisé. Dans le cas de régressions où les prédicteurs analysés ne sont pas mesurés de la même façon, le coefficient bêta permet de déterminer lesquels ont un impact plus important sur la variable dépendante. Le coefficient bêta β est un coefficient de régression standardisé. Il est de **0,597** d'après le tableau ci-dessous. Pour l'obtenir, il suffit de transformer le coefficient de régression b par l'équation :

$$\beta = b\ (Sx / Sy)$$

Autrement dit, il faut multiplier le coefficient de régression b par le rapport entre l'écart-type de la variable X (VI) et l'écart-type de la variable Y (VD).

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
ScolPe	97	0	21	8,04	6,288
Scolme	97	0	18	7,54	5,916

Sachant que le coefficient de régression b = 0,562, et que l'écart-type de ScolPe = 6,288 et l'écart-type de ScolMe = 5,916, on peut retrouver ainsi le bêta :

$$\beta = 0{,}562*(6{,}288/5{,}916) = \mathbf{0{,}597}.$$

Interprétation statistique : Le coefficient bêta est de 0,597. Il suggère que lorsque la scolarité du père augmente d'un écart-type (soit 6,288), la scolarité de la mère augmente de 0,597 écart-type (soit 0,597*5,916 = 3,32). Ce qui signifie précisément que chaque fois que la scolarité du père augmente de 6,288 années, celle de la mère augmente proportionnellement de 3,32 années.

En régression simple, le coefficient Bêta est l'équivalent du coefficient de corrélation : donc le coefficient de corrélation simple **R** = β. Précisément : **β = R** = 0,597.

Coefficient standardisé « Bêta : Il (β) donne la même information que le coefficient de régression (pente **b**), mais sur une base standardisée selon l'écart-type. Le coefficient « Bêta » représente le changement en écart-type de Y pour une augmentation d'un écart-type de X. Il est approprié pour comparer la force de l'effet respectif de plusieurs prédicteurs dont l'échelle de mesure est différente.

1.3. Présentation et interprétation des résultats

Dans le cadre d'un article scientifique, d'un mémoire ou d'une thèse, les résultats de l'analyse de régression et de corrélation sont souvent présentés dans un tableau unique avant de faire l'objet d'une interprétation statistique ou sociologique dans le texte. Voici un exemple d'une façon de faire.

Tableau 1. *Modèle de régression de la scolarité de la Mère selon la scolarité du père chez les étudiants*

	b (erreur-type)
Scolarité du père	0.56***
	(0.08)
Constante	3.02***
	(0.79)
n	97
R-deux	0.36

Notes. Les entrées correspondent à des coefficients de régression non standardisés (avec les erreurs-types entre parenthèses). ***p<.001.

Interprétation statistique (analyse succincte)[2]: L'analyse de régression montre que la scolarité du père a un effet positif significatif sur la scolarité de la mère chez la population étudiante étudiée (b=0.56; β=0.60; p<0,001). Le nombre d'années de scolarité complétées par la mère est d'autant plus important que celui du père augmente. Le coefficient bêta suggère précisément que chaque fois que la scolarité du père augmente de 6,288 années (un écart-type), celle de la mère augmente proportionnellement de 3,32 années (0,60 écart-type). Pour un père qui a complété 21 ans de scolarité, on peut prédire le nombre d'années de scolarité de la mère à 14,8 ans. Le modèle linéaire explique 36% de la variation dans la variable dépendante (R^2=0.36). L'analyse subséquente des résidus (non montrés ici) confirme la validité de ce modèle, puisque le diagnostic révèle le respect des postulats de normalité, de linéarité et d'homogénéité des variances. Aucun cas extrêmement déviant n'est détecté, les résidus standardisés variant de -2.16 à 2.30.

Interprétation théorique/sociologique (discussion): Les résultats suggèrent que l'hypothèse de l'homogamie est confirmée. Les hommes instruits épousent des femmes instruites. Cela peut s'expliquer par le fait que la société valorise l'éducation comme un marqueur déterminant du statut social.

[2] On suppose que l'analyse descriptive univariée a été effectuée auparavant en ces termes : Au total, 97 étudiants ont répondu à propos de la scolarité en années du père (M=8.04±6.29) et celle de la mère (M=7.54±5.92)

2. Régression linéaire multiple : le modèle de convergence

Au départ, la scolarité du père explique 36% de la variation de la scolarité de la mère chez les étudiants. Il est possible d'augmenter cette proportion expliquée en impliquant d'autres prédicteurs comme la profession du père. *Toutes choses étant égales, y a-t-il un effet convergent de la scolarité et de la profession du père sur la scolarité de la mère chez les étudiants inscrits en L2 de sociologie à l'UGB ? Précisément, la scolarité du père a-t-elle un effet sur la scolarité de la mère, en contrôlant l'effet de la profession du père ? De même, la profession du père a-t-elle un effet sur la scolarité de la mère, en contrôlant l'effet de la scolarité du père ?*

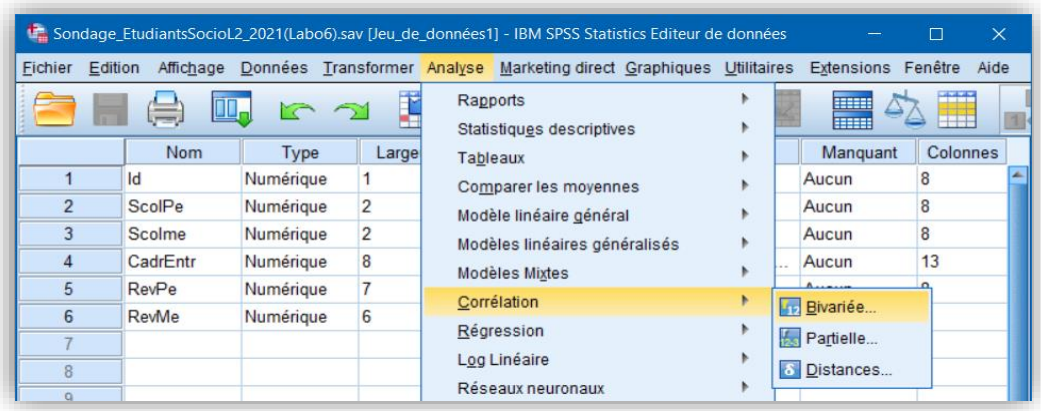
Scolarité du père (ScolPe) → **Scolarité de la mère (ScolMe)**

Profession du père (CadrEntr) →

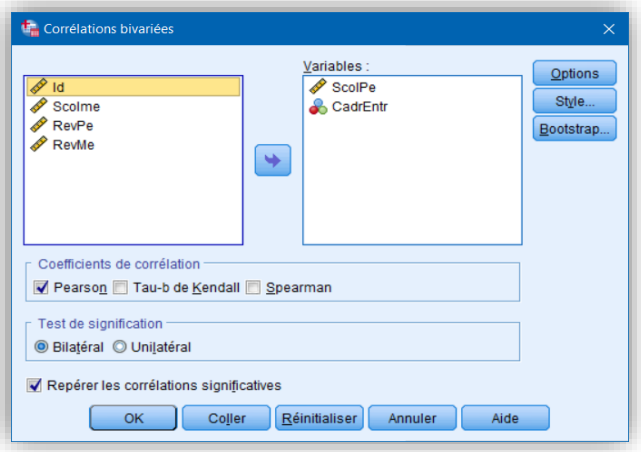
Toutes choses étant égales, on s'attend à ce que la scolarité du père et la profession du père ait ensemble et en présence l'une de l'autre un effet sur la scolarité de la mère. La plausibilité de cette hypothèse d'homogamie peut s'expliquer par l'importance du statut socioéconomique chez les hommes dans le choix d'une épouse.

Avant de soumettre ce modèle de convergence à l'épreuve de la régression linéaire multiple, assurons-nous qu'il n'y a pas un problème de colinéarité entre les deux VIs.

Sortons la matrice selon la procédure ci-dessous :



Renseignez les deux VIs : ScolPe et CadrEntr. Validez !



On obtient la matrice de corrélations ci-dessous.

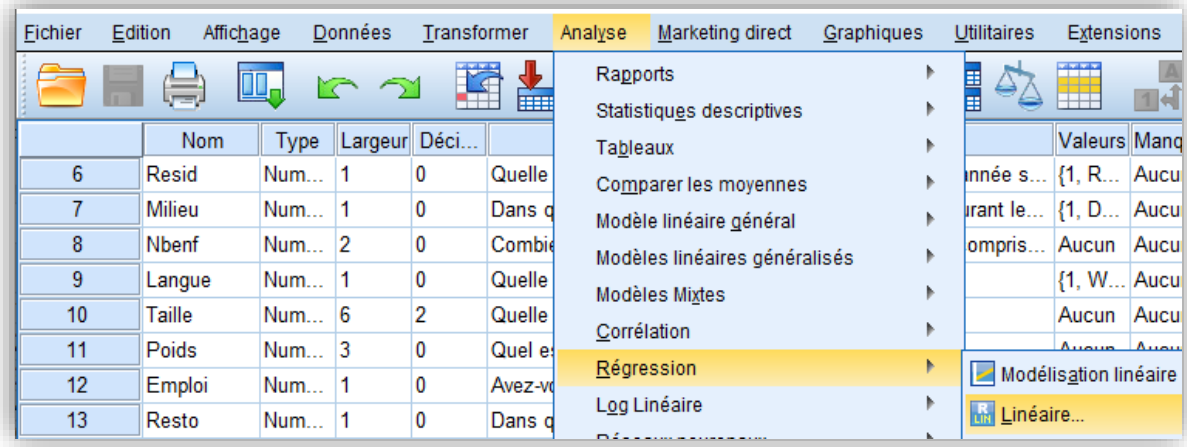
Corrélations

		ScolPe Combien d' années de scolarité, approximativement, votre père a-t-il complétées ?	CadrEntr Profession du père
ScolPe Combien d' années de scolarité, approximativement, votre père a-t-il complétées ?	Corrélation de Pearson	1	,132
	Sig. (bilatérale)		,216
	N	97	90
CadrEntr Profession du père	Corrélation de Pearson	,132	1
	Sig. (bilatérale)	,216	
	N	90	96

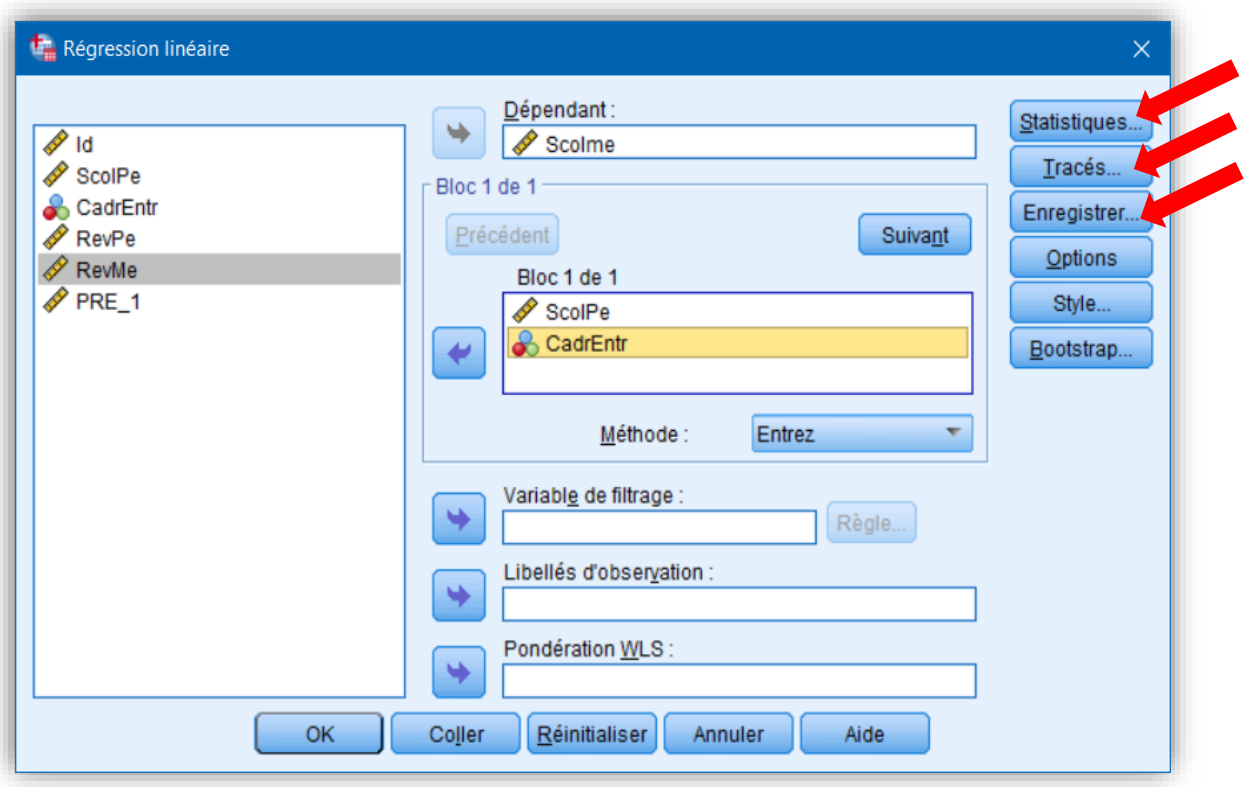
<u>Interprétation statistique :</u> Il n'y a aucun problème de colinéarité, puisque le coefficient de corrélation entre la scolarité et la profession du père est de seulement 0,13 (< à 0,70) et est non significatif. Les deux prédicteurs ne sont pas redondants et peuvent être inclus dans un même modèle.

Pour procéder à l'analyse de régression linéaire multiple, suivons la même procédure :

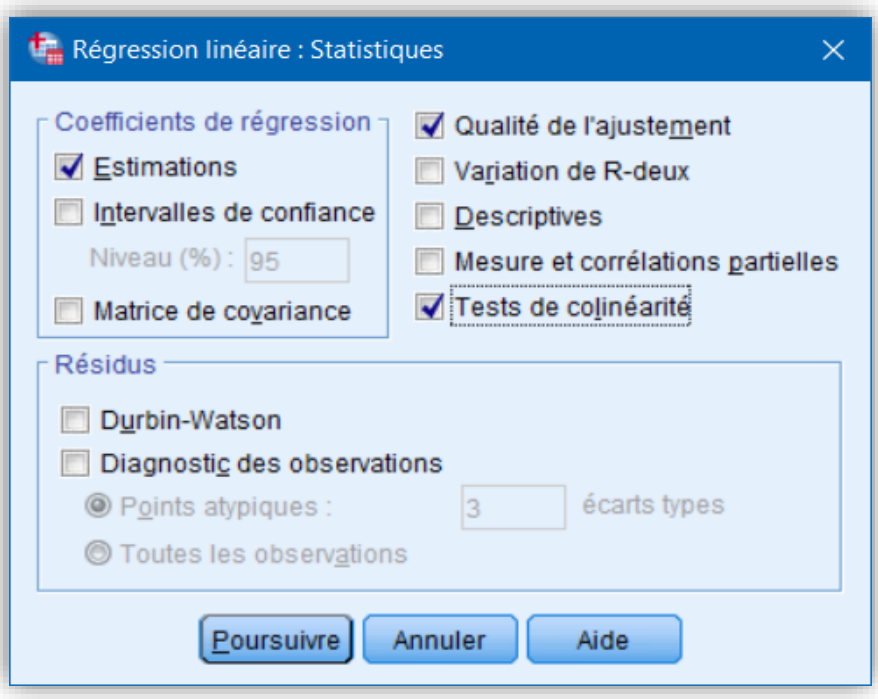
Analyse
Régression
Linéaire



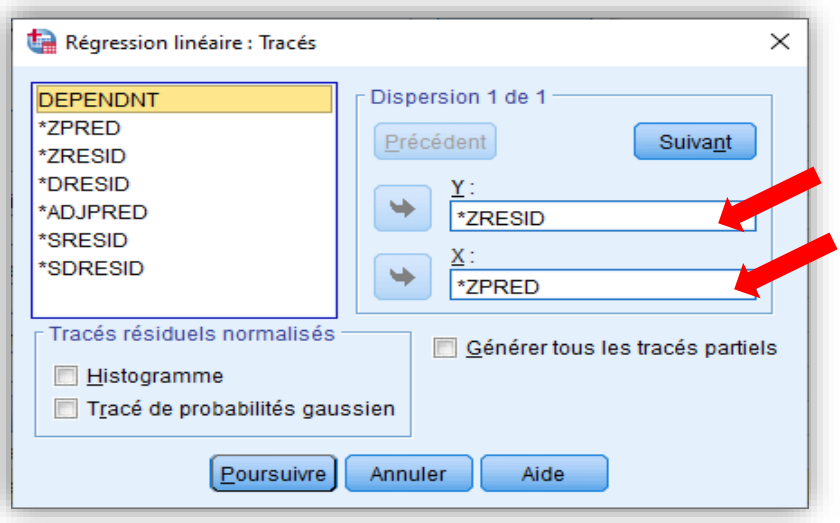
Une fenêtre de dialogue apparaît ! Cliquez sur la variable dépendante (Scolme), puis sur les variables indépendantes (Scolpe et CadrEntr) puisqu'il s'agit d'une analyse de régression multiple. Statistiquement, le modèle de convergence n'accorde aucune différence entre la VI (ScolPe) et la VC (CadrEntr).



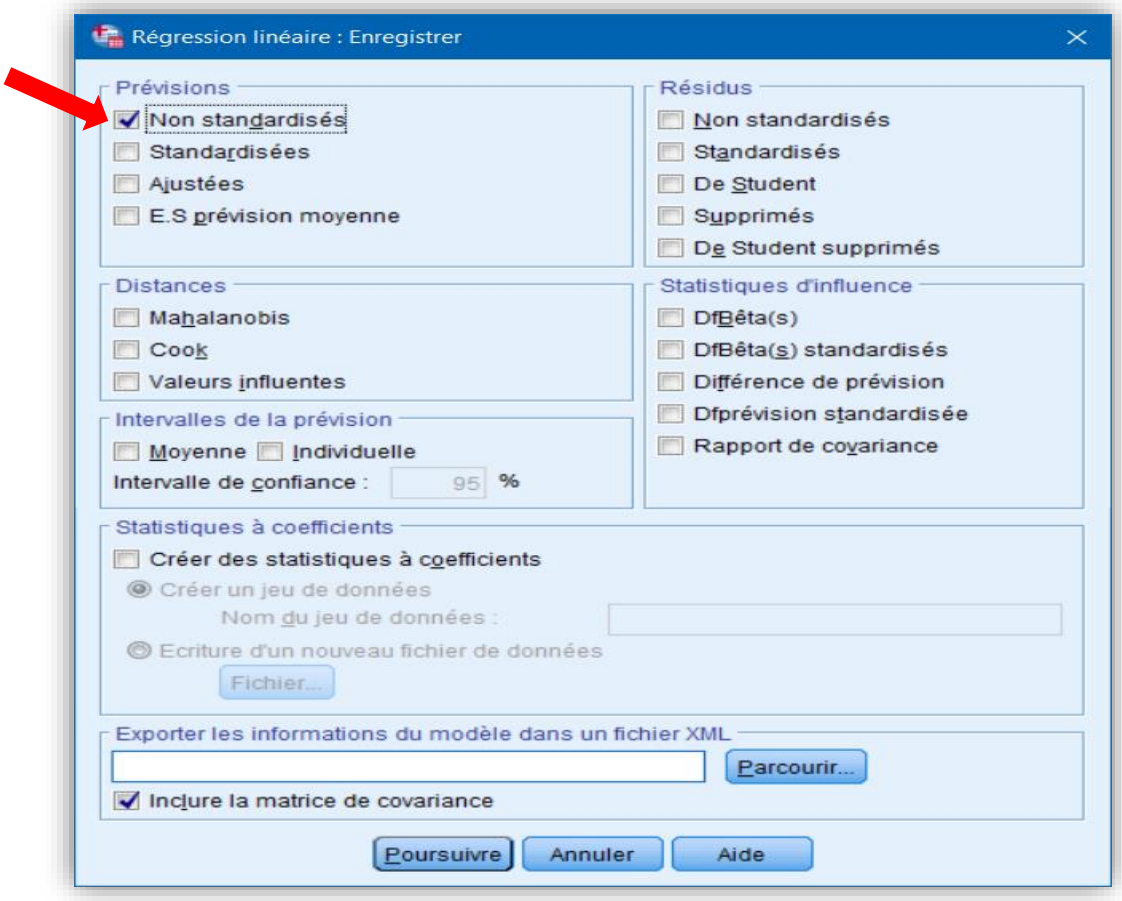
Pour procéder au diagnostic de la colinéarité de façon convaincante, cliquez sur la palette STATISTIQUES, et cochez Tests de colinéarité. Poursuivez !



Pour procéder au diagnostic du modèle de régression, cliquez sur la palette TRACÉS, et mettez les valeurs prédites standardisées (ZPRED) sur l'axe X et les valeurs résiduelles standardisées (ZRESID) sur l'axe Y. Poursuivez !



Sortez les valeurs prédites de la scolarité de la mère (Y), selon les scores des deux VIs (X1=Scolpe; X2=CadrEntr). Pour ce faire, cliquez sur la palette ENREGISTRER pour voir apparaître la fenêtre ci-dessous ! Puis, cochez la cage **Non standardisés** sous **Prévisions**!



Poursuivez et validez le tout pour voir apparaître les résultats.

3.1. Les coefficients de corrélation et de détermination multiples pour déterminer la force et la proportion expliquée du modèle multivarié

Récapitulatif des modèles[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,721[a]	,520	,509	4,170

a. Prédicteurs : (Constante), CadrEntr Profession du père, ScolPe Combien d'années de scolarité, approximativement, votre père a-t-il complétées ?

b. Variable dépendante : Scolme Combien d'années de scolarité, approximativement, votre mère a-t-elle complétées ?

Ce tableau, « Récapitulatif des modèles », précise le coefficient de corrélation multiple (0,721), et le **R-deux multiple**, qui est de 0,52. Le coefficient de corrélation (0.72) indique l'existence d'une corrélation modérée entre la scolarité et la profession du père d'une part, et la scolarité de la mère d'autre part. Pris ensemble, la scolarité et la profession du père explique 52% de la variation constatée dans la scolarité de la mère chez les étudiants, le coefficient de détermination étant de 0,52. La variance expliquée a augmenté de 19 points en pourcentage (de 33 % elle passe à 52 %) à la suite de l'ajout de la profession du père dans le modèle initial.

3.2. La signification statistique de la corrélation multiple au niveau de la population étudiée

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1639,256	2	819,628	47,131	,000[b]
	Résidu	1512,966	87	17,390		
	Total	3152,222	89			

a. Variable dépendante : Scolme Combien d'années de scolarité, approximativement, votre mère a-t-elle complétées ?

b. Prédicteurs : (Constante), CadrEntr Profession du père, ScolPe Combien d'années de scolarité, approximativement, votre père a-t-il complétées ?

Le tableau « ANOVA » présente le résultat du **test de signification statistique du coefficient de corrélation multiple** à l'aide du F d'ANOVA.

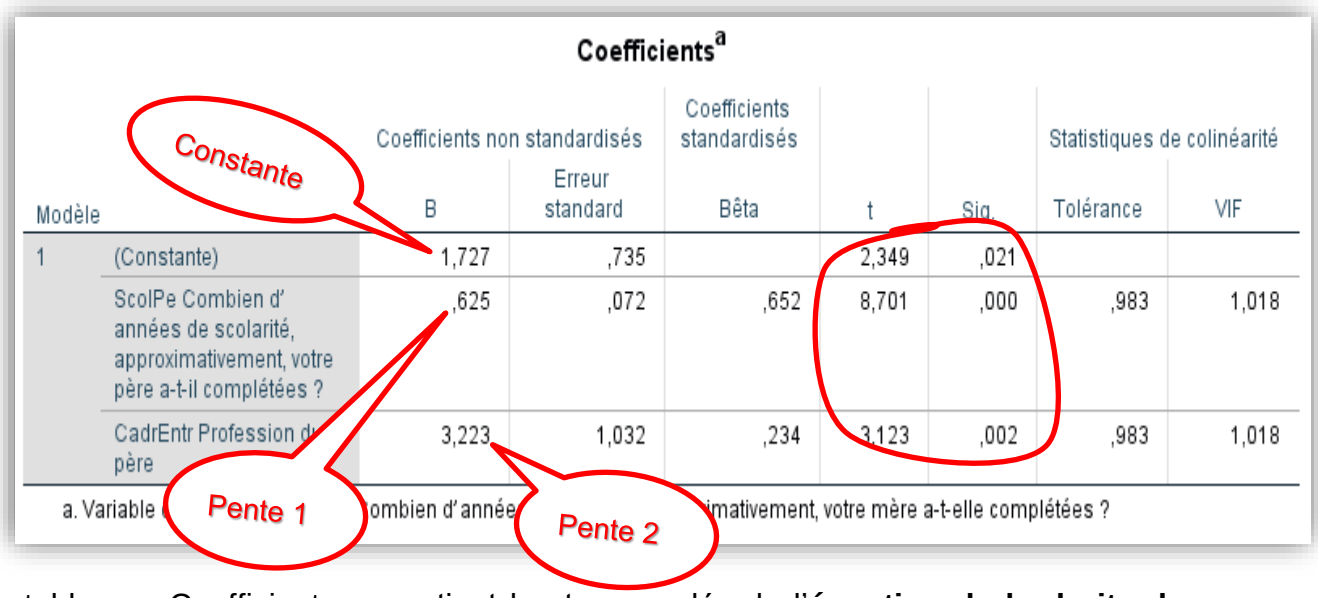
On peut aussi à partir du résultat de l'ANOVA retrouver la valeur du R^2 multiple :

SC Régression / SC Total = 1639,256/3152,222 = 0,52.

Le résultat du test d'ANOVA de la corrélation multiple est significatif au moins à 99,9 % (F=47,13; dl1= 2; dl2= 87; p<0.001). Ce qui indique que l'effet de la scolarité et de

la profession du père sur la scolarité de la mère est statistiquement significatif dans l'ensemble de la population étudiante étudiée.

3.3. L'équation de régression linéaire multiple pour procéder à la prédiction



Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur standard	Bêta	t	Sig.	Tolérance	VIF
1	(Constante)	1,727	,735		2,349	,021		
	ScolPe Combien d' années de scolarité, approximativement, votre père a-t-il complétées ?	,625	,072	,652	8,701	,000	,983	1,018
	CadrEntr Profession du père	3,223	1,032	,234	3,123	,002	,983	1,018

a. Variable ombien d'année imativement, votre mère a-t-elle complétées ?

Le tableau « Coefficients », contient les termes clés de l'**équation de la droite de régression multiple** et leurs **degrés de signification**.

La constante (1,727) signifie que lorsque la scolarité du père est nulle (0) et que la profession du père est nulle (Autre profession), on s'attend à ce que la mère complète 1,727 année de scolarité. Elle est significative (a=1,727 ; p<0.05).

Le modèle contient deux coefficients de régression. La scolarité du père a un effet positif significatif sur la scolarité de la mère chez les étudiants, même après avoir contrôlé l'effet de la profession du père (b=0.625; β=0,65 ; p<0.001). Toutes choses étant égales, une augmentation d'une unité (année) de la scolarité du père augmente la scolarité de la mère de 0.625 année. De même, la profession du père a un effet positif significatif sur la scolarité de la mère chez les étudiants, même après avoir contrôlé l'effet de la scolarité du père (b=3.22; β=0.23 ; p<0.01). Toutes choses étant égales, lorsqu'on passe d'une autre profession à une profession de cadre/entrepreneur chez le père, la scolarité de la mère augmente de 3,22 années. L'effet de la scolarité du père (β=0.65) est près de trois fois plus important sociologiquement que l'effet de la profession du père (β=0.23). En dernière analyse, on peut avoir confiance au pouvoir explicatif et prédictif de l'équation de régression multiple puisque les deux coefficients de régression sont significatifs au moins à 95%.

Voici l'équation du plan de régression : $\mathbf{Y = a + b_1X_1 + b_2X_2}$

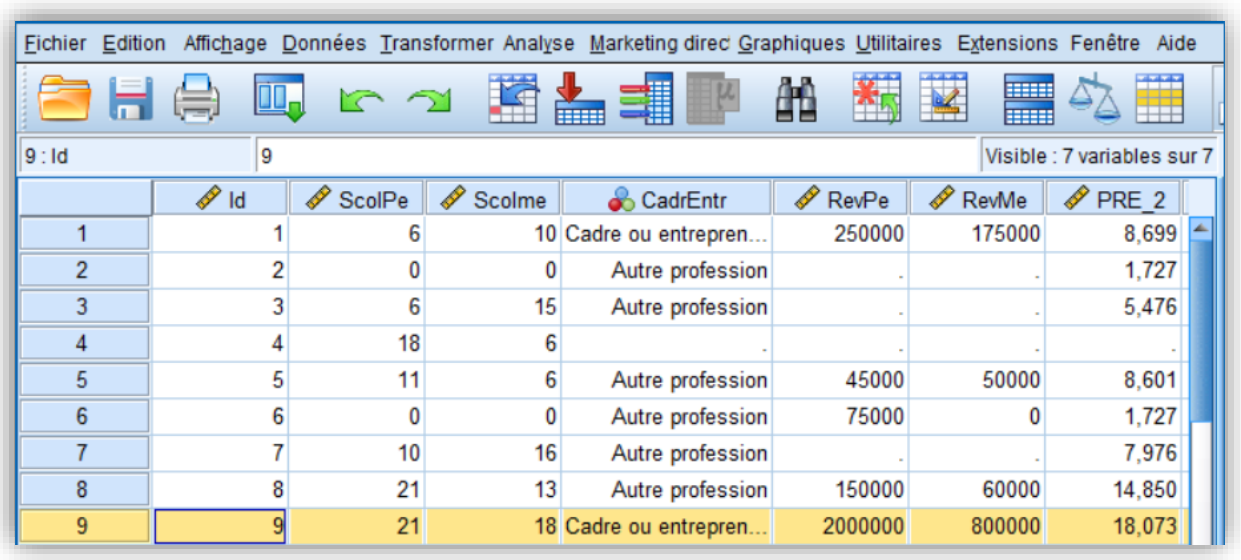
$Y = 1,73 + 0,625(X_1) + 3,22(X_2)$

Scolarité mère = 1,73 + 0,625(**Scolarité père**) + 3,22(**Profession père**)

D'une certaine manière, les valeurs de Y sont prédites par la combinaison d'une constante, de l'effet d'une variable X1 et de l'effet d'une variable X2. Ainsi, connaissant la scolarité et la profession du père, on peut estimer ou prédire la scolarité de la mère ***Par exemple, quelle est la scolarité prédire de la mère, si le père possède 21 années de scolarité (X=21) et est de profession cadre/entrepreneur (X=1) ?***

$\hat{Y}$ = 1,73 + 0,625 (21) + 3,22(1) = **18,07**

D'après l'équation de régression linéaire multiple, on s'attend à ce que tout père possédant 21 années de scolarité et cadre/entrepreneur, soit associé à une mère ayant complété 18 années de scolarité. On retrouve les valeurs prédites que nous venons d'enregistrer sous forme de variable (PRE_2) dans la base de données :



	Id	ScolPe	Scolme	CadrEntr	RevPe	RevMe	PRE_2
1	1	6	10	Cadre ou entrepren...	250000	175000	8,699
2	2	0	0	Autre profession	.	.	1,727
3	3	6	15	Autre profession	.	.	5,476
4	4	18	6	.	.	.	.
5	5	11	6	Autre profession	45000	50000	8,601
6	6	0	0	Autre profession	75000	0	1,727
7	7	10	16	Autre profession	.	.	7,976
8	8	21	13	Autre profession	150000	60000	14,850
9	9	21	18	Cadre ou entrepren...	2000000	800000	18,073

3.4. Le diagnostic du modèle de régression

Tout d'abord, vérifions si le postulat d'absence de mutlicolinéarité en analysant la tolérance et le facteur d'inflation de la variance (VIF).

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur standard	Bêta	t	Sig.	Tolérance	VIF
1	(Constante)	1,727	,735		2,349	,021		
	ScolPe Combien d' années de scolarité, approximativement, votre père a-t-il complétées ?	,625	,072	,652	8,701	,000	,983	1,018
	CadrEntr Profession du père	3,223	1,032	,234	3,123	,002	,983	1,018

a. Variable dépendante : Scolme Combien d'années de scolarité, approximativement, votre mère a-t-elle complétées ?

<u>Interprétation statistique :</u> Sans surprise, il n'y a aucun problème de colinéarité, puisque la tolérance est très proche de 1 (Tolérance=0,983, compris entre 0,5 et 1) et le VIF dépasse très peu 1 (VIF=1,018 < 2). Les deux VIs, scolarité et profession du père, comporte chacune une variance spécifique suffisamment importante pour qu'elles ne soient pas redondantes.

Ensuite, vérifions les trois postulats de normalité, de linéarité et d'homogénéité en analysant les résidus.

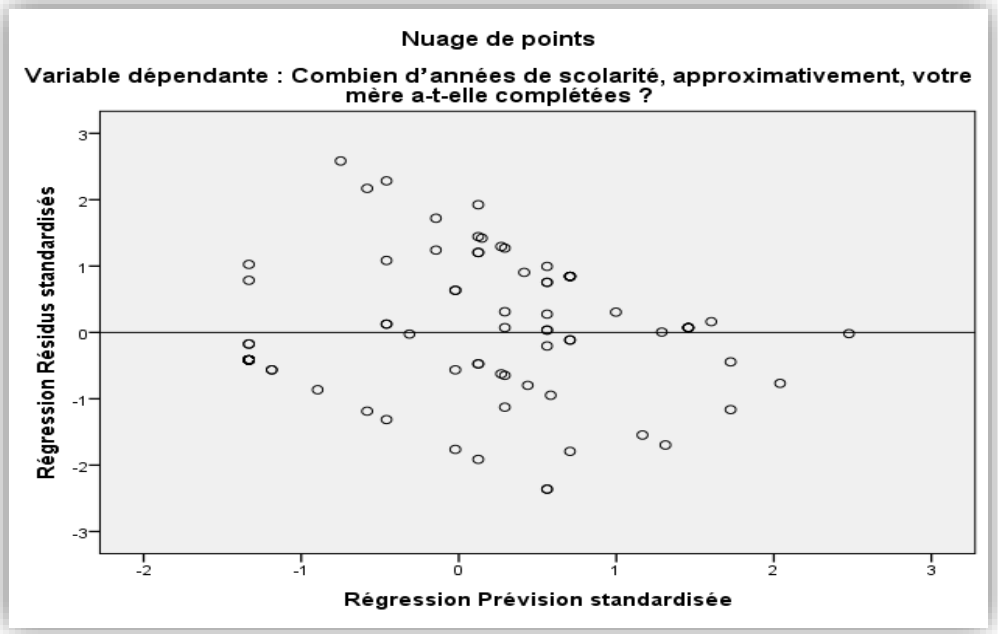
Statistiques des résidus[a]

	Minimum	Maximum	Moyenne	Ecart type	N
Prévision	1,73	18,07	7,44	4,292	90
Résidu	-9,851	10,774	,000	4,123	90
Prévision standardisée	-1,332	2,477	,000	1,000	90
Résidu standardisé	-2,362	2,584	,000	,989	90

a. Variable dépendante : Scolme Combien d'années de scolarité, approximativement, votre mère a-t-elle complétées ?

Analyse ou interprétation statistique des résidus :

1. **Linéarité** de la relation entre la VD et la VI :
   Les points du diagramme des résidus ci-dessous ne suivent pas une quelconque forme curvilinéaire. Ils indiquent donc l'existence d'une relation linéaire entre la scolarité du père (ScolPe=VI1), la profession du père (CadrEntr=VI2) et la scolarité de la mère (ScolMe=VD)
2. **Normalité** de la VD à l'intérieur des niveaux de la VI :
   Les points du diagramme des résidus sont répartis de façon assez similaire de part et d'autre de la ligne horizontale autour de 0. Comme le montrent les statistiques des résidus, les résidus standardisés varient de -2.362 à 2.584. Il n'y a donc aucun cas extrêmement déviant, les scores résiduels étant inférieurs à 3 en valeurs absolues. Il semble y avoir une normalité de la distribution de **ScolMe** (VD) à l'intérieur des niveaux de **ScolPe** (VI1) et **CadrEntr** (VI2).
3. **Homogénéité des variances** de la VD à l'intérieur des niveaux de la VI :
   Dans l'ensemble, les résidus sont répartis de façon assez homogène verticalement. Le nuage de points ne forme pas de groupes qui indiqueraient la présence d'une hétéroscédasticité. Il y a donc homogénéité des variances de **ScolMe** (VD) à l'intérieur des niveaux de **ScolPe** (VI1) et **CadrEntr** (VI2)



3.5. Présentation des résultats aux fins de diagnostic causal

Dans le cadre d'un article scientifique, d'un mémoire ou d'une thèse, les résultats de l'analyse de régression linéaire simple et multiple sont souvent présentés dans un tableau unique avant de faire l'objet d'une interprétation statistique ou sociologique dans le texte. Voici un exemple d'une façon de faire.

Tableau 3. *Modèle de régression de la scolarité de la mère selon la scolarité et la profession du père chez les étudiants*

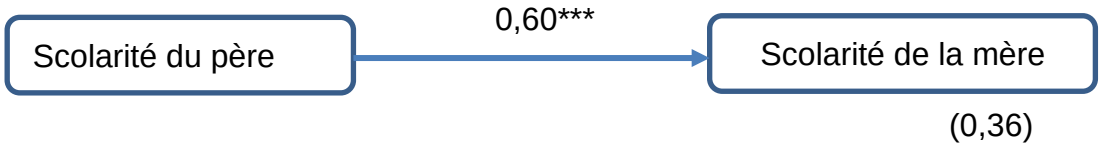
	Modèle 1	Modèle 2
Scolarité du père	0.56***	0.63***
	(0.08)	(0.07)
Profession du père (cadre/entrepreneur)	--	3.22**
		(1.03)
Constante	3.02***	1,73*
	(0.79)	(0.74)
n	90	90
R-deux	0.36	0.52

Notes. Les entrées correspondent à des coefficients de régression non standardisés (avec les erreurs-types entre parenthèses). ***p<.01; **p<.01; *p<.05.

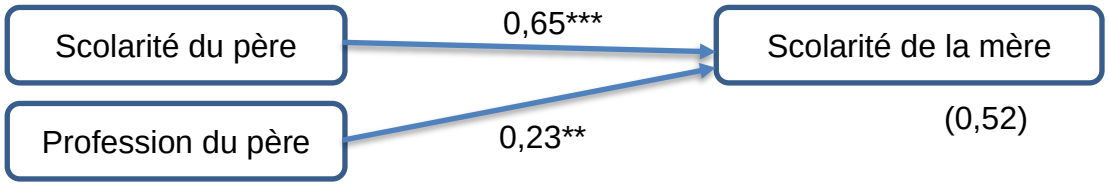
NB : Comme il est question du modèle de convergence, on aurait pu également présenter seulement les résultats du modèle 2. Mais l'avantage de considérer le modèle 1, et donc d'y aller en régression hiérarchique, c'est de pouvoir connaître la proportion expliquée (r-deux) qui s'ajoute à la proportion initiale.

Figure 1. *Schématisation de l'effet convergent de la scolarité et profession du père sur la scolarité de la mère chez les étudiants*

a) Modèle initial



b) Modèle causal complet (convergence)



Notes : Les entrées correspondent à des coefficients standardisés bêta, simple et multiple. Le chiffre sous la variable dépendante correspond au r-deux simple ou multiple. ***p<0,001.

NB : Comme il est question du modèle de convergence, on aurait pu également se passer de la schématisation du modèle initial. Le plus important, c'est le modèle causal complet incluant tous les prédicteurs. On aurait pu également présenter les coefficients de régression standardisés, qui toutefois sont affectés par les différences de mesure des variables.

Interprétation statistique (analyse succincte)[3]:

L'analyse de régression linéaire bivariée (modèle 1 du tableau) montre que la scolarité du père a un effet positif significatif sur la scolarité de la mère chez la population étudiante étudiée (b=0.56; β=0.60; p<0,001). Le nombre d'années de scolarité complétées par la mère est d'autant plus important que celui du père augmente. Le coefficient bêta suggère précisément que chaque fois que la scolarité du père augmente de 6,288 années (un écart-type), celle de la mère augmente proportionnellement de 3,32 années (0,60 écart-type). Pour un père qui a complété 21 ans de scolarité, on peut prédire le nombre d'années de scolarité de la mère à 14,8 ans. Le modèle linéaire explique 36% de la variation dans la variable dépendante (R^2=0.36). L'analyse subséquente des résidus (non montrés ici) confirme la validité de ce modèle, puisque le diagnostic révèle le respect des postulats de normalité, de linéarité et d'homogénéité des variances. Aucun cas extrêmement déviant n'est détecté, les résidus standardisés variant de -2.16 à 2.30.

Que se passe-t-il maintenant lorsqu'on ajoute la profession du père dans le modèle initial ? L'analyse de régression linéaire multiple (modèle 2 du tableau) montre que la scolarité du père a un effet positif significatif sur la scolarité de la mère chez les étudiants, même après avoir contrôlé l'effet de la profession du père (b=0.625; β=0,65 ; p<0.001). Toutes choses étant égales, une augmentation d'une unité (année) de la scolarité du père augmente la scolarité de la mère de 0.625 année. De même, la profession du père a un effet positif significatif sur la scolarité de la mère chez les étudiants, même après avoir contrôlé l'effet de la scolarité du père (b=3.22; β=0.23 ; p<0.01). Toutes choses étant égales, lorsqu'on passe d'une autre profession à une profession de cadre/entrepreneur chez le père, la scolarité de la mère augmente de 3,22 années. L'effet de la scolarité du père (β=0.65) est près de trois fois plus important sociologiquement que l'effet de la profession du père (β=0.23). Pris ensemble, la scolarité et la profession du père explique 52% de la variation constatée dans la scolarité de la mère chez les étudiants (R-deux=0,52). La variance expliquée a augmenté de 19 points en pourcentage (de 33 % elle passe à 52 %) à la suite de l'ajout de la profession du père dans le modèle initial.

En conséquence, tel qu'illustré à travers la figure 1, il y a convergence. Les résultats attestent l'effet convergent de la scolarité et profession du père sur la scolarité de la mère. Chacun de ces prédicteurs présente un effet explicatif et prédictif convaincant, leur effet étant additif.

Par ailleurs, le postulat d'absence de colinéarité est respecté puisque la tolérance et le VIF ne sont pas éloignés de 1 (Tolérance= 0,983 ; VIF= 1,018). L'analyse subséquente des résidus confirme la validité du modèle multivarié, puisque le diagnostic révèle le respect des postulats de normalité, de linéarité et d'homogénéité des variances, et les résidus standardisés variant de -2.16 à 2.30.

Interprétation théorique/sociologique (discussion): Les résultats suggèrent que l'hypothèse de l'homogamie est confirmée par le modèle de convergence. Les hommes instruits et qui sont surtout des cadres/entrepreneurs épousent des femmes

[3] On suppose que l'analyse descriptive univariée a été effectuée auparavant en ces termes : Au total, 97 étudiants ont répondu à propos de la scolarité en années du père (M=8.04±6.29) et celle de la mère (M=7.54±5.92). Alors que parmi les 96 répondants valides, 24% sont des cadres ou entrepreneurs, 76% d'entre eux relevant d'une autre profession.

instruites. Cela peut s'expliquer par le fait que la société valorise l'éducation et la profession comme un marqueur déterminant du statut socioéconomique.

3. Exercice pratique

En tant que chercheur, vous étudiez les facteurs associés au revenu de la mère chez les étudiants inscrits en L2 de sociologie à l'UGB. La question de recherche est la suivante : *Le revenu du père (RevPe) influe-t-il sur le revenu de la mère (RevMe), en contrôlant l'effet convergent de la profession du père (CadrEntr) ?* Vous souhaitez donc étudier l'homogamie, en termes de revenu, chez les parents des étudiants en tenant compte de la profession du père.

Pour élucider cette question de recherche, répondez aux interrogations ci-dessous :

Étape 1 : Cadre opératoire et théorique

a) Quelles sont la variable dépendante, la variable indépendante, et la variable-contrôle additive dans ce contexte ?

b) Étant donné que la question de recherche suggère la convergence comme modèle causal, proposez une explication théorique de l'élaboration. Autrement dit, est-il plausible que le revenu et la profession du père agissent simultanément sur le revenu de la mère ?

Étape 2 : Description des variables

a) Faites sortir les statistiques descriptives des variables **Revme et Revme**;

b) Faites sortir les fréquences et pourcentages de la variable : **CadrEntr**;

c) Selon la façon dont chacune des trois variables est distribué, peut-on élucider leur relation à l'aide de l'analyse de régression linéaire multiple? Justifiez.

Étape 3 : Analyse bivariée de la relation initiale (X-Y)

a) Sortez les résultats de l'analyse de régression linéaire simple concernant la relation entre **Revpe** et **Revme**;

b) Interprétez statistiquement le coefficient de corrélation R, le R-deux, le test de la corrélation, la constante, le coefficient de régression et le bêta.

c) Supposons qu'on passe d'un père qui gagne 100 000 f cfa par mois à un père qui a comme revenu mensuel 200 000 f cfa, de combien en F cfa augmentera ou diminuera-t-il le revenu de la mère?

- Rédiger l'équation de régression de la relation entre le revenu du père et celui de la mère. Puis prédisez, pour un étudiant, le revenu mensuel de la mère (**Y**) si le père gagne 100 000 f cfa par mois (**X)**.
- Peut-on avoir confiance en la capacité de prédiction de l'équation pour l'ensemble de la population étudiée? Autrement dit, ses paramètres sont-ils statistiquement significatifs?

Étape 4 : Analyse multivariée (X-Y, contrôlant Z)

a) Sortez les résultats de l'analyse de régression linéaire multiple concernant la relation entre Revpe et Revme, contrôlant pour CadrEntr.

b) Interprétez statistiquement le coefficient de corrélation multiple R, le R-deux, le test de la corrélation, la constante, les coefficients de régression.

c) Rédiger l'équation de régression linéaire multiple de la relation entre le revenu du père et celui de la mère, contrôlant pour la profession du père. Puis prédisez, pour un étudiant, le revenu mensuel de la mère (Y) si le père gagne 500 000 f cfa par mois (X=500 000) et s'il est un cadre/entrepreneur (X=1).

d) Peut-on avoir confiance en la capacité de prédiction de l'équation pour l'ensemble de la population étudiée? Autrement dit, ses paramètres sont-ils statistiquement significatifs?

e) Schématiser les résultats du modèle initial et du modèle causal complet, en utilisant les bêtas et les p-values des tests des coefficients de régression;

f) Quel est le diagnostic causal de l'élaboration (convergence ou non) ?

g) Le modèle de régression est-il valide? Procéder à l'analyse de la tolérance, du VIF et des résidus.