Régression linéaire multiple

Régression linéaire multiple

- Il est fort possible que la variabilité de la variable dépendante **Y** soit expliquée non pas par une seule variable indépendante X mais plutôt par une combinaison linéaire de plusieurs variables indépendantes $\mathbf{X_1, X_2, \ldots, X_p}$.
- Dans ce cas le modèle de régression multiple est donné par:

$$\mathbf{Y = a_1X_1 + a_2X_2 + \ldots + a_pX_p + \beta + \varepsilon}$$

- Aussi, à l'aide des données de l'échantillon nous estimerons les paramètres $\mathbf{\beta, a_1, \ldots, a_p}$ du modèle de régression de façon à minimiser la somme des carrés des erreurs.

- Le coefficient de corrélation multiple $\mathbf{R^2}$, aussi appelé coefficient de détermination, nous indique le pourcentage de la variabilité de $\mathbf{Y}$ expliquée par les variables indépendantes $\mathbf{X_1, X_2, \ldots, X_p}$.
- Lorsqu'on ajoute une ou plusieurs variables indépendantes dans le modèle, le coefficient $\mathbf{R^2}$ augmente.
- La question est de savoir si le coefficient $\mathbf{R^2}$ augmente de façon significative.
- Notons qu'on ne peut avoir plus de variables indépendantes dans le modèle qu'il y a d 'observations dans l'échantillon (règle générale: $n \geq 5p$).

Calcul du coefficients de corrélation linéaire multiple

- $R_{Yx1x2} = \sqrt{\dfrac{r_{Yx1} + r_{Yx2} - 2\,(r_{Yx1}\, r_{Yx2}\, r_{x1x2})}{(1 - r^2_{x1x2})}}$

Sachant que : r_{Yx1} , r_{Yx2} et r_{x1x2} sont des coefficients de corrélation linéaire simple des variables $y\ x_1$, $y\ x_2$ et $x_1\ x_2$

Les coefficients de corrélation partielle

- 1) le coe:fficient de corrélation partielle entre y et x_1

$$R_{Yx1x2} = \frac{r_{Yx1} - r_{Yx2}\ r_{x1x2}}{\sqrt{(1 - r^2_{yx2})\ (1 - r^2_{x1x2})}}$$

- 2) le coe:fficient de corrélation partielle entre y et x_2

$$R_{Yx2x1} = \frac{r_{Yx2} - r_{Yx1}\ r_{x1x2}}{\sqrt{(1 - r^2_{yx1})\ (1 - r^2_{x2x1})}}$$

3) le coe:fficient de corrélation partielle entre x_1 et x_2

$$R_{x1x2Y} = \frac{r_{x1x2} - r_{Yx1}\ r_{Yx2}}{\sqrt{(1 - r^2_{yx1})\ (1 - r^2_{Yx2})}}$$

Calcul des coefficients de l'équation de la droite régression multiple

$$a_1 = \frac{\operatorname{cov}(x_1, x_2)\,\operatorname{cov}(x_2, y) - \operatorname{cov}(x_1, y)\, v(x_2)}{\operatorname{cov}^2(x_1, x_2) - v(x_1)\, v(x_2)}$$

$$a_2 = \frac{\operatorname{cov}(x_1, x_2)\,\operatorname{cov}(x_1, y) - \operatorname{cov}(x_2, y)\, v(x_1)}{\operatorname{cov}^2(x_1, x_2) - v(x_2)\, v(x_1)}$$

- $B = \overrightarrow{y} - a_1 \overrightarrow{x}_1 - a_2 \overrightarrow{x}_2$

Exemple:

MODÈLE 1.

The regression equation is

Totale = 3,05 Terrain - 20730 Acre + 43,3 Pied2 - 4352 Pièces + 10049 Chambre + 7606 SbainsC + 18725 Sbains + 882 Foyers - 89131

Predictor	Coef	StDev	T	P
Constant	**-89131**	18302	-4,87	0,000
Terrain	**3,0518**	0,5260	5,80	0,000
Acre	**-20730**	7907	-2,62	0,011
Pied2	**43,336**	7,670	5,65	0,000
Pièces	**-4352**	3036	-1,43	0,156
Chambre	**10049**	5307	1,89	0,062
SbainsC	**7606**	3610	2,11	0,039
Sbains	**18725**	6585	2,84	0,006
Foyers	**882**	3184	0,28	0,783

S = 29704 **R-Sq = 88,9%** **R-Sq(adj) = 87,6%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	8	4,93877E+11	61734659810	69,97	0,000
Residual Error	70	61763515565	882335937		
Total	78	5,55641E+11			

MODÈLE 2

Regression Analysis

The regression equation is

Totale = 3,11 Terrain - 21880 Acre + 40,2 Pied2 + 4411 Chambre + 8466 SbainsC + 14328 Sbains- 97512

Predictor	Coef	StDev	T	P
Constant	**-97512**	17466	-5,58	0,000
Terrain	**3,1103**	0,5236	5,94	0,000
Acre	**-21880**	7884	-2,78	0,007
Pied2	**40,195**	7,384	5,44	0,000
Chambre	**4411**	3469	1,27	0,208
SbainsC	8466	3488	2,43	0,018
Sbains	**14328**	5266	2,72	0,008

S = 29763 **R-Sq = 88,5%** **R-Sq(adj) = 87,6%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	4,91859E+11	81976430646	92,54	0,000
Residual Error	72	63782210167	885864030		
Total	78	5,55641E+11			

MODÈLE 3

Regression Analysis

The regression equation is

Totale = 3,20 Terrain - 22534 Acre + 41,1 Pied2 + 10234 SbainsC + 14183 Sbains - 90408

Predictor	Coef	StDev	T	P
Constant	**-90408**	16618	-5,44	0,000
Terrain	3,2045	0,5205	6,16	0,000
Acre	-22534	7901	-2,85	0,006
Pied2	41,060	7,383	5,56	0,000
SbainsC	10234	3213	3,19	0,002
Sbains	14183	5287	2,68	0,009

S = 29889 **R-Sq = 88,3%** **R-Sq(adj) = 87,5%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	4,90426E+11	98085283380	109,80	0,000
Residual Error	73	65214377146	893347632		
Total	78	5,55641E+11			

Modèle sans la superficie du terrain (# d ’acres) à cause de la multi colinéarité avec la valeur du terrain.

MODÈLE 4

The regression equation is

Totale = - 55533 + 1,82 Terrain + 49,8 Pied2 + 11696 SbainsC + 18430 Sbains

Predictor	Coef	StDev	T	P
Constant	**-55533**	11783	-4,71	0,000
Terrain	1,8159	0,1929	9,42	0,000
Pied2	49,833	7,028	7,09	0,000
SbainsC	11696	3321	3,52	0,001
Sbains	18430	5312	3,47	0,001

S = 31297 **R-Sq = 87,0%** **R-Sq(adj) = 86,3%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	4,83160E+11	1,20790E+11	123,32	0,000
Residual Error	74	72481137708	979474834		
Total	78	5,55641E+11			

Parmi les 4 modèles précédents, lequel choisiriez vous et pourquoi?

- Probablement le modèle 4 car toutes les variables indépendantes sont significatives au niveau 5% (c.-à-d. p-value < 5% pour chaque **a** dans le modèle) et bien que le $\mathbf{R}^2$ soit plus petit, il n 'est que marginalement plus petit.

- Dans le modèle 1 les variables ' # de pièces ' et ' # de foyers ' ne sont pas statistiquement significatives au niveau 5% (p-value > 5%). La variable ' # de chambres ' est à la limite avec un p-value = 0,0624.

- Dans le modèle 2 la variable ‘ # de chambres ’ n ’est pas statistiquement significative au niveau 5%.

- Dans le modèle 3 (et les modèles précédents), le coefficient de la variable ‘ # d ’acres ’ est négatif ce qui est à l ’encontre du « bon sens » et de ce qu ’on a observé sur le diagramme de dispersion et le coefficient de corrélation de Pearson positif (r = 0,608).

- Le coefficient négatif pour la variable ‘ # d ’acres ’ dans les modèles 1 à 3 est causé par le fait qu’il y a une forte relation linéaire entre la valeur du terrain et la superficie du terrain (r = 0,918); problème de multicolinéarité.

Comment choisir un modèle de régression linéaire parmi tous les modèles possibles?

Il existe plusieurs techniques:

- sélection pas à pas en ajoutant une variable à la fois et en commençant par la plus significative (stepwise, forward).
- sélection à partir du modèle incluant toutes les variables et en enlevant une variable à la fois en commençant par la moins significative (backward).
- faire tous les modèles possibles et choisir le meilleur sous-ensemble de variables (best subset) selon certains critères spécifiques (ex: **R^2 ajusté**, **C_p** de Mallow.)

Le choix du meilleur modèle se fait selon la combinaison:

- **La plus grande valeur de R^2 ajusté pour le nombre de variables dans le modèle.**
- **La plus petite valeur de C_p .**
- **Pour les modèles avec R^2 ajusté et C_p comparables, on choisira le modèle qui a le plus de « sens » selon les experts dans le domaine.**
- **Pour les modèles avec R^2 ajusté et C_p comparables, le modèle avec les variables indépendantes les plus faciles et moins coûteuses à mesurer.**
- **La validité du modèle.**

Intervalle de confiance au niveau 1-α pour la moyenne de Y et une nouvelle valeur de Y (prévision) étant donné une combinaison de valeurs spécifiques pour X_1, X_2, …, X_p .

- Pour le modèle 4 et une propriété avec terrain= 65 000DA, pi^2 = 1500, 2 salles de bain complète et 1 non-complète, on obtient l'estimation ponctuelle suivante:
  - **est. valeur totale = 1,816*65 000 + 49,833*1 500 + 11 696*2 + 18 430*1 -55 533 = 179 074DA**
  - **intervalle de confiance à 95% pour la moyenne de la valeur totale:**

    [170 842, 187 306]
  - **intervalle de confiance à 95% pour une valeur totale prédite :**

    [116 173, 241 974]

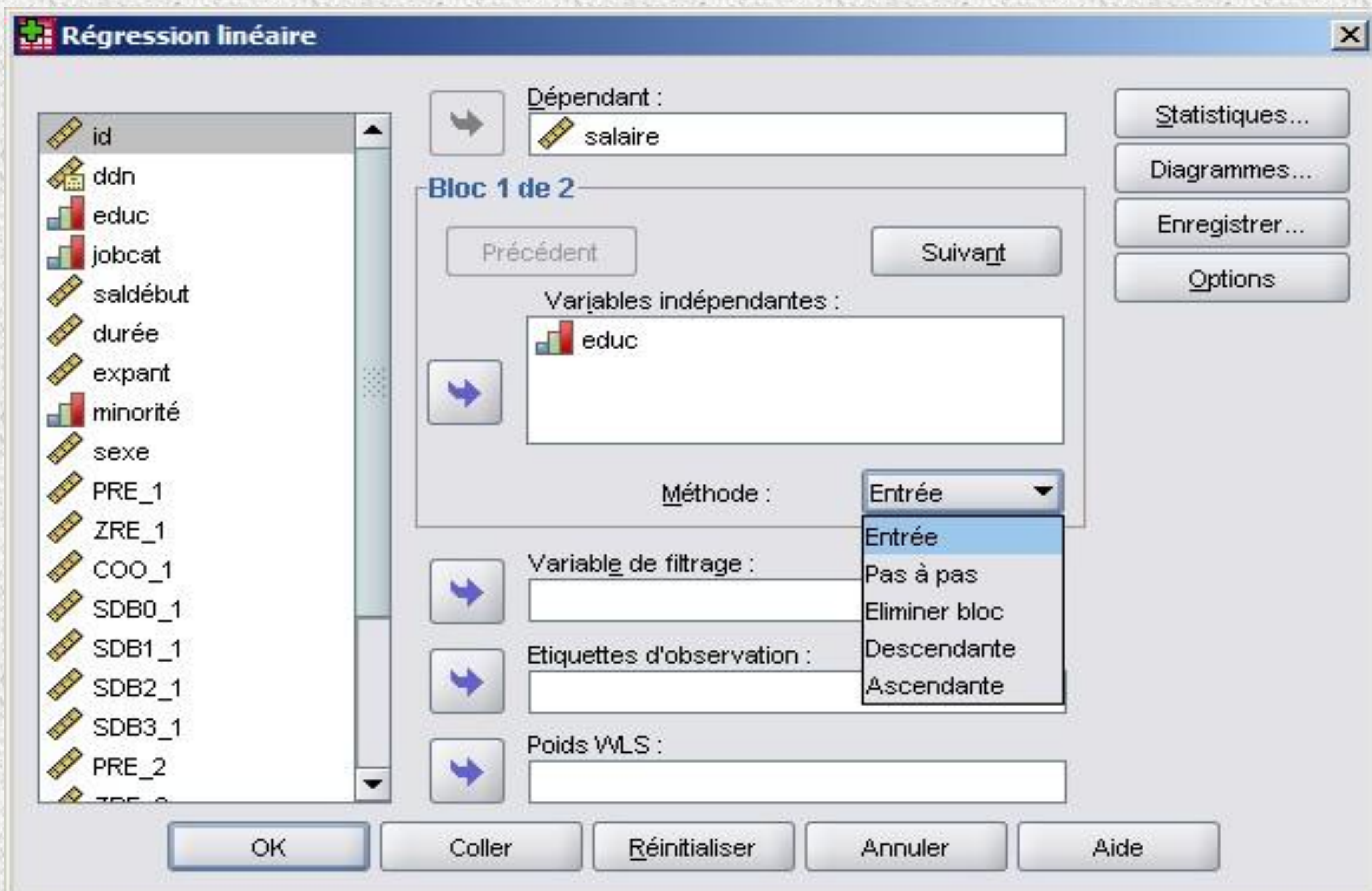
Remarques:

- **Les longueurs des intervalles de confiance au niveau 95% du modèle de régression multiple pour une propriété de 1500 pi^2 sont plus petites que pour le modèle de régression simple.**
- **Donc l'addition de plusieurs autres variables dans le modèle a aidé à expliquer encore plus la variabilité de la valeur totale et à améliorer nos estimations.**
- **Si deux ou plusieurs variables indépendantes sont corrélées on dira qu'il y a multicolinéarité. Ceci peut influencer les valeurs des paramètres dans le modèle.**
- **Aussi, si deux variables indépendantes sont fortement corrélées, alors seulement une des deux variables sera incluse dans le modèle, l 'autre n 'apportant que très peu d 'information supplémentaire.**
- **Certaines conditions sont nécessaires à la validité du modèle et de l 'inférence correspondante (similaire à la régression linéaire simple).**

Application de la régression multiple sous SPSS

1. Pour réaliser l'analyse, cliquez sur **Analyse**, **Régression**, puis **Linéaire**.

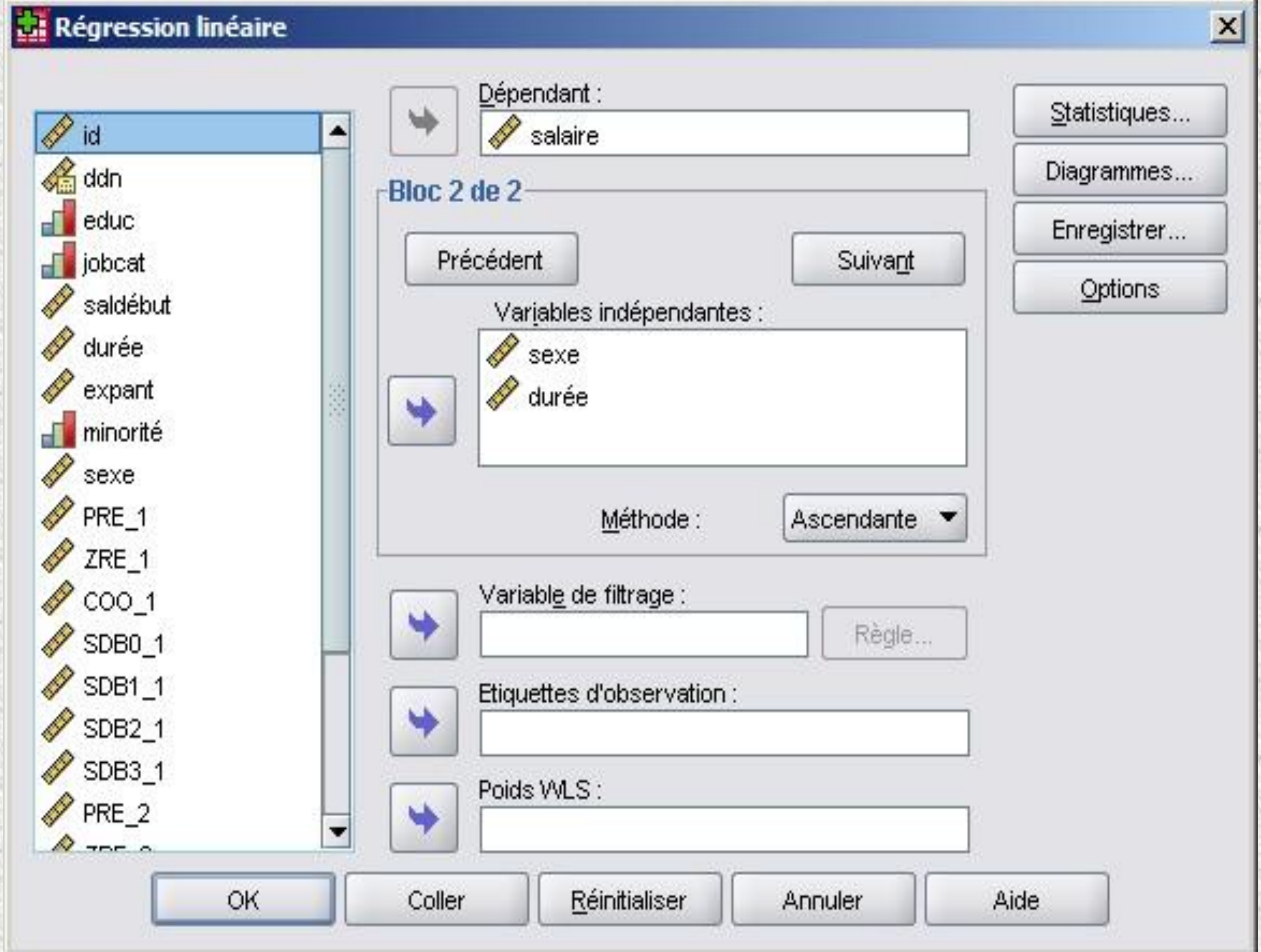
2. En cliquant sur , insérez la variable dépendante et la ou les variable(s) indépendante(s) dans les boîtes appropriées.



3. Si vous désirez absolument que la première variable indépendante soit incluse, privilégiez la méthode **Entrée**.

4. Pour créer des blocs (groupes) de variable(s) indépendante(s) dans le cadre d'une régression hiérarchique, cliquez sur 

lorsque le premier bloc est construit, puis insérez les variables indépendantes des autres blocs en répétant cette procédure. La méthode de régression (Entrée, Pas à pas, etc.) peut être déterminée pour chaque bloc. Habituellement, la méthode **Entrée** est utilisée à moins d'a priori théoriques particuliers.

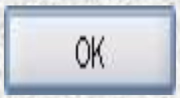

Régression linéaire
Dépendant :
salaire
Bloc 2 de 2
Précédent
Suivant
Variables indépendantes :
sexe
durée
Méthode :
Ascendante
Variable de filtrage :
Règle...
Etiquettes d'observation :
Poids WLS :
Statistiques...
Diagrammes...
Enregistrer...
Options
id
ddn
educ
jobcat
saldébut
durée
expant
minorité
sexe
PRE_1
ZRE_1
COO_1
SDB0_1
SDB1_1
SDB2_1
SDB3_1
PRE_2
OK
Coller
Réinitialiser
Annuler
Aide

5. Vous pouvez choisir une variable de filtrage pour limiter l'analyse à un sous-échantillon formé par les participants ayant obtenu une ou des valeur(s) particulière(s) à cette même variable.

6. Vous pouvez aussi spécifier une variable qui permettra d'identifier les coordonnées sur le graphique (**Étiquettes d'observation**).

7. Enfin, vous pouvez choisir une variable numérique pondérée (**Poids WLS**) pour effectuer l'analyse des moindres carrés. Par cette analyse, les valeurs sont pondérées en fonction de leurs variances réciproques, ce qui implique que les observations avec de larges variances ont un impact moins important sur l'analyse que les observations associées à de petites variances.

8. Assurez-vous d'avoir sélectionné les options nécessaires (par exemple, sous le bouton Statistiques).

9. Pour procéder à l'analyse, cliquez sur 

Une présentation détaillée de toutes les options est disponible dans le procédurier de la régression simple.

Le bouton 

Pour la régression multiple, nous suggérons de cochez les options suivantes :

L’encadré Coefficients

Estimations : valeurs *b* pour chaque VI et son test de signification

Intervalles de confiance : intervalle pour chaque coefficient dans la population

L’encadré Résidus

Durbin-Watson : évaluation de l’indépendance des erreurs

Diagnostic des observations : valeur de la VD observée, prédite, du résiduel et du résiduel standardisé pour chaque observation. Indique quelles observations ont un résiduel standardisé de plus de 2 ou 3 é.-t. (au choix de l'utilisateur)

Les autres statistiques

Qualité de l'ajustement : fournit le test pour évaluer l'ensemble du modèle (F), le R multiple, le R^2 correspondant et le R^2 ajusté

Variation de R-deux : changement du R^2 après l'ajout d'un nouveau bloc de VI

Caractéristiques : moyenne, é.-t. et N pour toutes les variables du modèle

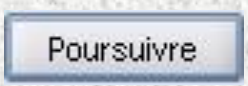
Mesure et corrélations partielles :

Corrélation entre chaque VI et la VD

Corrélation partielle entre chaque VI et VD en contrôlant pour les autres VI

Corrélation « partie » ou semi-partielle entre chaque VI et la variance non expliquée de la VD par les autres VI

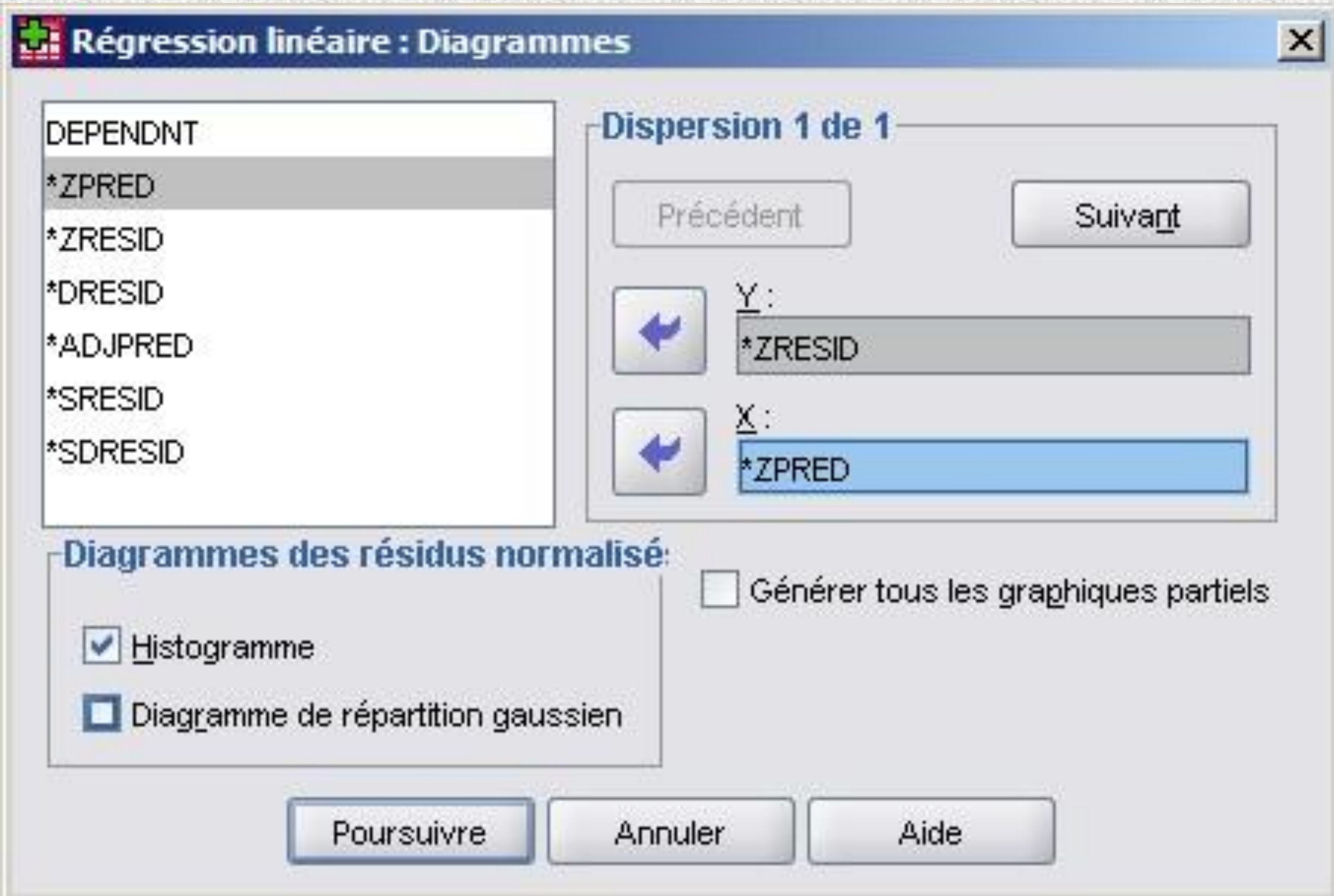
Test de colinéarité : évaluation de la multicolinéarité dans le modèle (VIF).

Cliquez sur 

pour revenir à la boite de dialogue principale.

Le bouton 

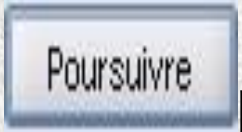
Les graphiques offerts permettent de vérifier par un examen visuel les prémisses de la régression linéaire multiple. Celui croisant les valeurs prédites (*ZPRED) et résiduelles (*ZRESID) standardisées illustre le respect (ou le non respect) de la prémisse d’homogénéité (répartition aléatoire des points autour de 0) et de linéarité (tendance des points à se concentrer autour d’une ligne).


Régression linéaire : Diagrammes
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Dispersion 1 de 1
Précédent
Suivant
Y :
*ZRESID
X :
*ZPRED
Diagrammes des résidus normalisé
Histogramme
Diagramme de répartition gaussien
Générer tous les graphiques partiels
Poursuivre
Annuler
Aide

Pour faire plus d'un graphique, utilisez le bouton 

.

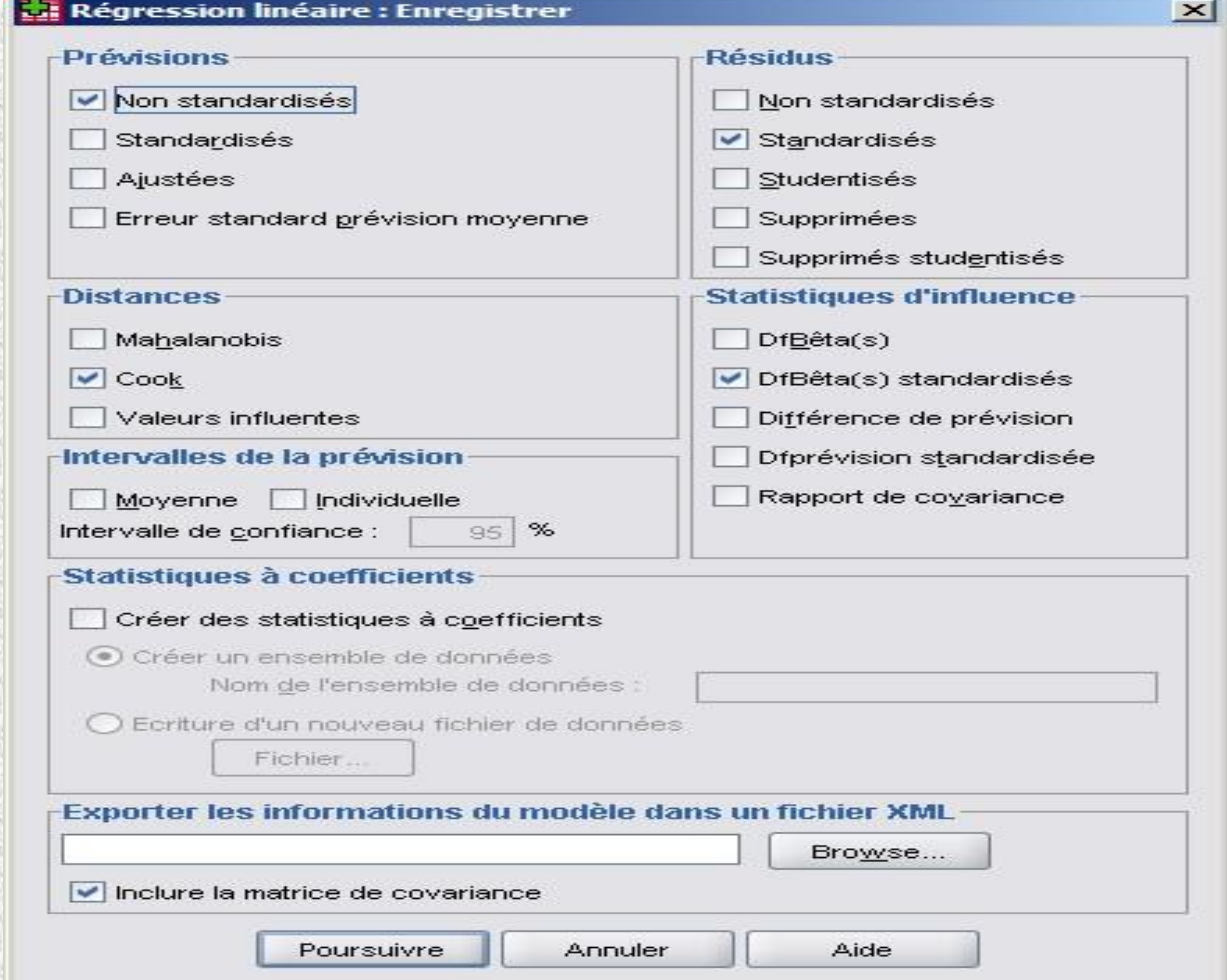
L’encadré des diagrammes des résidus normalisés permet d’illustrer la distribution des résiduels (histogramme et diagrammes de répartition gaussiens), ce qui vous permet de faire un examen visuel du respect de la prémisse de normalité de la distribution des erreurs.

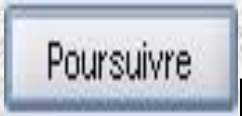
Cliquez sur 

pour revenir à la boîte de dialogue principale.

Le bouton 

Toutes les options disponibles dans ce menu permettent de créer des nouvelles variables ayant les valeurs calculées par le modèle. Il s'agit donc de choisir les variables diagnostiques permettant d'évaluer la qualité du modèle et celles qui permettent de détecter les variables ayant une importante influence sur le modèle. On choisira donc minimalement les résidus standardisés, mais on peut également ajouter les valeurs prédites non standardisées et standardisées (valeur de la VD calculée pour chaque observation) ainsi que la distance de Cook et les DfBêta(s) standardisés. Notez qu'en cochant des options dans la boîte de dialogue **Enregistrer**, vous allez obtenir un tableau de résultats de plus portant sur les statistiques des résidus et comprenant minimalement la moyenne, l'écart-type, les valeurs minimales et maximales ainsi que le N.

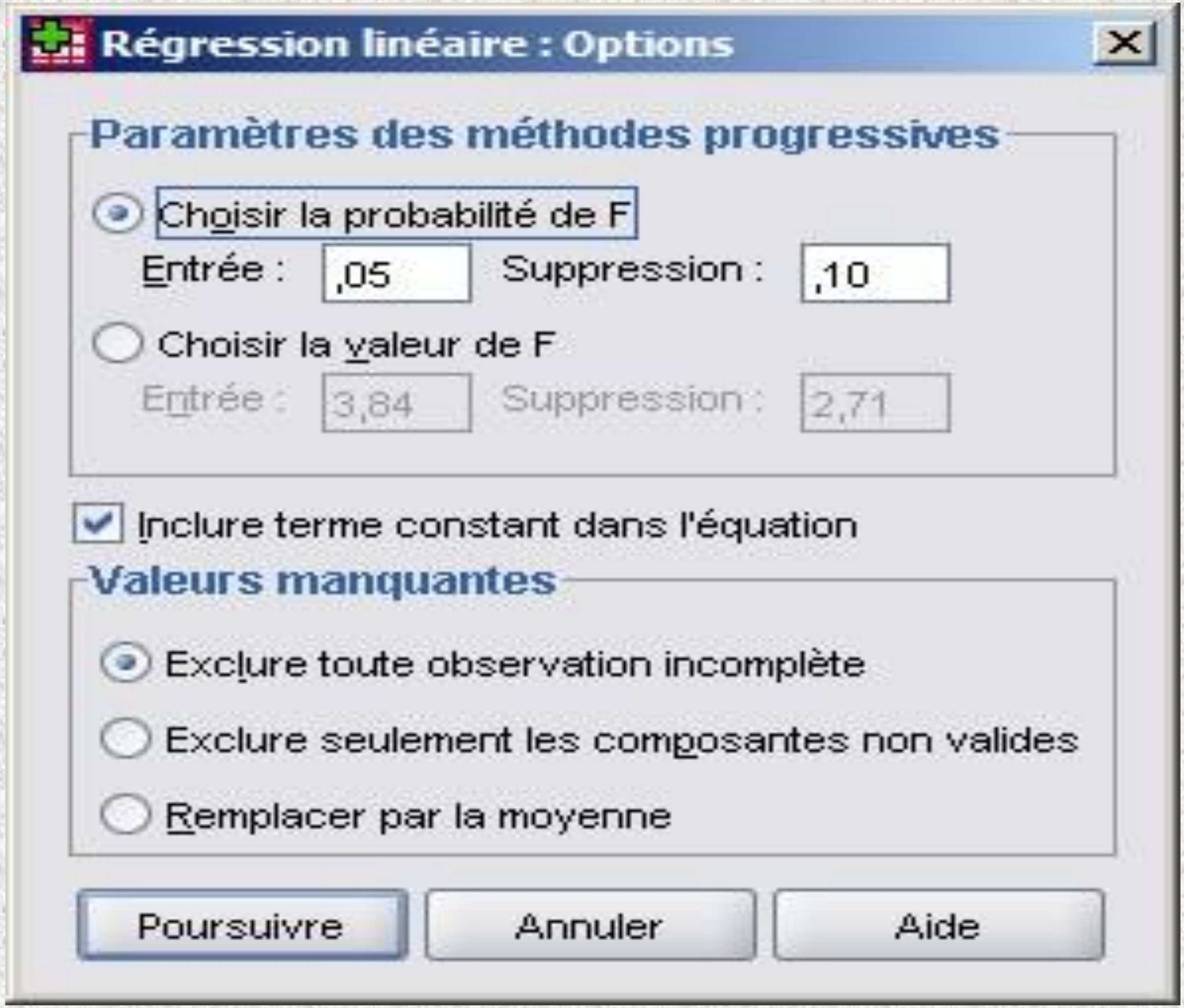

Régression linéaire : Enregistrer
Prévisions
Non standardisés
Standardisés
Ajustées
Erreur standard prévision moyenne
Résidus
Non standardisés
Standardisés
Studentisés
Supprimées
Supprimés studentisés
Distances
Mahalanobis
Cook
Valeurs influentes
Statistiques d'influence
DfBêta(s)
DfBêta(s) standardisés
Différence de prévision
Dfprévision standardisée
Rapport de covariance
Intervalles de la prévision
Moyenne
Individuelle
Intervalle de confiance : 95 %
Statistiques à coefficients
Créer des statistiques à coefficients
Créer un ensemble de données
Nom de l'ensemble de données :
Ecriture d'un nouveau fichier de données
Fichier...
Exporter les informations du modèle dans un fichier XML
Browse...
Inclure la matrice de covariance
Poursuivre
Annuler
Aide

Cliquez sur 

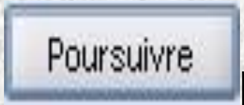
pour revenir à la boîte de dialogue principale.

Le bouton 

La dernière fenêtre vous permet de déterminer les paramètres de sélection des méthodes d'entrée progressives (Ascendante ou descendante - *stepwise*). Vous pouvez utiliser la valeur de la probabilité associée à la valeur F (soit la valeur de *p*) ou encore la valeur de la statistique F pour introduire ou retirer des variables. Idéalement, vous conservez les valeurs par défaut à moins que vous ne vouliez que les critères d'entrée ou de retrait des variables de votre modèle soient plus sévères ou plus inclusifs.


Régression linéaire : Options
Paramètres des méthodes progressives
Choisir la probabilité de F
Entrée : ,05
Suppression : ,10
Choisir la valeur de F
Entrée : 3,84
Suppression : 2,71
Inclure terme constant dans l'équation
Valeurs manquantes
Exclure toute observation incomplète
Exclure seulement les composantes non valides
Remplacer par la moyenne
Poursuivre
Annuler
Aide

Évidemment, vous laissez aussi la constante dans l’équation. Vous pouvez finalement spécifier ce que vous désirez faire avec les valeurs manquantes. Encore une fois, l’option par défaut est à privilégier puisque le retrait de toute observation incomplète permet de conserver toujours le même nombre d’observations, ce qui favorise la cohérence du modèle.

Cliquez sur 

pour revenir à la boite de dialogue principale.