Université de Montréal

Département de sociologie

L'analyse de régression multiple

Notes de cours

Table des matières

1. 0 Rappel des principes de base:

1.1 La corrélation

La corrélation

- est un *indice de la force* d'une relation linéaire ou linéarisée (après transformation) entre deux ou plusieurs variables.
- est un indice *standardisé* de la relation, ce qui permet de comparer les corrélations entre elles.
- donne aussi le *sens* (positif ou négatif) de la relation.

La corrélation égale:

a) la covariance – et donc l'indice de jusqu'à quel point l'augmentation sur une variable est accompagnée d'un mouvement similaire – positif ou négatif – sur l'autre variable. De façon à standardiser – amener le coefficient sur une échelle de moyenne 0 et de variance 1, la covariance est divisée par le produit des écarts-type de x et y

$$r=covarXY/s_x s_y$$

ou...

b) le coefficient de régression (b) divisé par l'écart-type de la variable dépendante

$$r = b/s_y$$

Le coefficient de détermination, soit la corrélation mise au carré, donne l'information la "variance expliquée". Il est égal au ration de

a) la somme des carrés (SC) expliquée – c'est-à-dire la somme des écarts mis au carré entre les valeurs *prédites* et la moyenne et
b) la somme des carrés totale, soit la somme des écarts mis au carré entre les valeurs *réelles* et la moyenne**.**

$$r^2 = \text{SC expliquée/ SC totale}$$

1.2 La ligne de régression, la régression simple

On est intéressé à la régression quand on veut savoir jusqu'à quel point on peut prédire la valeur d'une variable en connaissant la valeur d'une ou de plusieurs autres variables. L'équation de la régression simple est:

$$\overline{Y} = a + bx$$

"$\overline{Y}$" peut être conceptualisé comme la valeur attendue, la moyenne, pour une valeur de X donnée "E(Y|X)", l'espérance (E) de Y étant donné ce que nous connaissons de la valeur de x.

Comme il y a des écarts autour de la moyenne, chaque valeur de Y, y_i est égale à la valeur attendue de Y en fonction de la valeur de x à laquelle s'ajoute l'erreur, aussi appelé 'résidu':

$$y_i = a+bx_i+e_i$$

"a" peut être conceptualisé comme l'intercept de Y, soit la valeur moyenne que prend Y quand la valeur de X = 0;

"b" est le coefficient assigné à la variable indépendante X.
Il peut être compris comme le poids donné à la variable indépendante X, pour prédire la variable dépendante Y. Il dit jusqu'à quel point et dans quel sens la valeur moyenne de y est modifiée pour une augmentation de 1 de la valeur de x.

"e" peut être conceptualisé comme l'"erreur" comprenant l'erreur de mesure (voir alpha de Cronbach) ainsi que l'effet non contrôlé d'autres variables qui ne sont pas dans l'équation. La valeur de "e_i" pour un cas donné est l'écart entre la valeur y_i prédite par l'équation "$a +bx_i$" et la valeur réelle y_i.

2.0 La régression multiple:

Dans la régression multiple, on cherche la combinaison de coefficients (b) pour les variables indépendantes (X_i) qui amènerait les valeurs de Y prédites par l'équation aussi près que possible des valeurs de Y mesurées. L'équation est la suivante:

$$\overline{Y} = a + b_1x_1 + b_2x_2 + ... + b_nx_n$$

On cherche toujours à minimiser les écarts entre les valeurs prédites et les valeurs mesurées mais en recourant à plusieurs variables qui nous aident à prédire; ***la corrélation multiple*** *est un indice de la relation entre ces valeurs prédites et les valeurs mesurées.*

Notons que l'analyse de variance est un cas spécial d'une régression multiple dans laquelle les variables multi-nominales ayant k catégories seraient dichotomisées en k-1 variables.

Il y a quatre manières d'entrer les variables dans une régression multiple **(Tabachnik et Fidell, 2012, p. 136-144**). Ces quatre manières diffèrent par la manière dont les variables entrent dans l'équation et donc par la façon dont est traitée la variance commune à plusieurs variables.

- **La régression standard**: Toutes les variables sont entrées en même temps dans l'équation. La variance commune à plusieurs variables n'est attribuée à aucune des variables. On cherche à estimer le degré de relation entre chaque variable indépendante et la variable dépendante. *Ce type de régression permet de connaître la contribution unique (corrélation semi-partielle) de chaque variable* indépendante à la prédiction de la variable dépendante (V.D.).

- **La régression hiérarchique ou séquentielle**: Les variables sont entrées une à une ou par groupe de variables selon un ordre déterminé théoriquement par le chercheur. La variance commune à plusieurs variables est attribuée séquentiellement selon l'ordre d'entrée des variables. On cherche à estimer si et jusqu'à quel point une variable indépendante ou un groupe de variables indépendantes ajoute à la prédiction, au-delà des autres variables déjà dans l'équation. *Ce type de régression permet de connaître la contribution ajoutée d'une ou de plusieurs variables.*

- **La régression statistique ou pas-à-pas**: Les variables indépendantes entrent dans l'équation uniquement en fonction de critères statistiques (probabilité statistique de signification du coefficient de régression standardisé "Beta". On cherche la meilleure équation de prédiction, sans égard à la signification des variables. *Ce type de régression est utilisé surtout à titre exploratoire.*

➢ **La régression SETWISE**: Les variables sont entrées par bloc dont on compare la contribution globale. On cherche le meilleur ensemble de prédicteurs, par exemple si le revenu peut être mieux prédit par un ensemble de caractéristiques liées à l'éducation ou par un ensemble de variables relatives au contexte socio-économique du lieu de résidence.

2.1 Ce que l'on peut savoir avec une régression multiple (Tabachnik et Fidell, 2012, p.119-121):

- s'il existe une relation significative entre les prédicteurs et la V.D, c'est-à-dire si, dans la population, la relation est différente de 0.

$$H_0: r=0; \qquad H_1: r \neq 0$$

- si chacune des variables contribue de façon significative à la prédiction

$$H_0: b_i=0; \qquad H_1: b_i \neq 0$$

- si l'addition d'une variable (k) ou d'un ensemble de variables (Différence des R^2) à un ensemble existant contribue de façon significative à la prédiction

$$H_0: b_k=0; \qquad H_1: b_k \neq 0$$

- si une relation autre que linéaire (curvilinéaire, logarithmique,...) prédirait mieux qu'une relation linéaire (en transformant les variables et en comparant les coefficients, en analysant les résidus, ...)

L'analyse de régression peut être utilisée pour :

- prédire les valeurs de la V.D. dans un nouvel ensemble de données pour lesquelles seules les V.I ont été mesurées.

- effectuer des analyses de cheminement de causalité (surtout effectuées maintenant avec des procédures permettant l'évaluation simultanée des équations, soit les équations structurelles).

2.2 Considérations pratiques (Tabachnik et Fidell, 2012, p. 123-128):

Nombre de cas par variable.

Il devrait y avoir au minimum:

Régression standard et hiérarchique/séquentielle: **20 cas par variable**

Régression statistique: **40 cas par variable**, ceci parce que ce type de régression, fortement dépendante de l'échantillon, est moins stable et donc plus difficilement généralisable à la population.

La régression linéaire est relativement robuste mais elle est quand même sensible aux écarts aux postulats. Plus l'effet est présumé faible, moins la distribution des variables est normale, moins la fidélité est bonne, plus il faut de cas par variable.

Par ailleurs, lorsque l'on a beaucoup de variables et que certaines combinaisons de ces variables peuvent constituer une échelle, il devient d'autant plus judicieux de réduire le nombre de variables dans l'équation par la création d'échelles dont la fidélité peut être mesurée (C. Durand, *L'analyse factorielle et de fidélité*, http://www.mapageweb.umontreal.ca/durandc/menuMethodesQuantitatives.html#analyse_compo).

Valeurs extrêmes (outliers)

Les valeurs extrêmes peuvent avoir un impact très important sur l'équation de régression. Il est d'autant plus important de les identifier et d'agir en conséquence, c'est-à-dire transformer la variable ou retirer les cas de l'analyse selon les situations.

Multicollinéarité et singularité

On dit qu'il y a un problème de singularité lorsqu'une variable donnée est la combinaison d'une ou de plusieurs autres variables.

Comme on cherche à ce que chaque variable apporte le plus de variance **unique** possible, il est évident qu'une variable indépendante pouvant être prédite par les autres variables indépendantes ne nous intéresse pas, puisqu'elle n'ajoute rien à la prédiction. Si un tel cas se produit, il faut identifier la variable indépendante pouvant être prédite par les autres variables indépendantes et la retirer de l'analyse (sur des bases théoriques, logiques et statistiques).

Normalité, linéarité, homoscédasticité (homogénéité des variances), indépendance des résidus

Les postulats du modèle linéaire sont importants en régression multiple. Il est toutefois difficile sinon impossible d'examiner ces questions de façon multivariée en examinant les distributions univariées et bi-variées. L'analyse des résidus permet d'évaluer si les postulats sont respectés. Ceci dit, la régression est très robuste aux écarts aux postulats, surtout lorsque la taille de l'échantillon est importante et qu'il n'y a pas de relations clairement non linéaires.

2.3 Sommes des carrés, équations, test F, r^2:

La somme des carrés totale (SC, Somme des écarts de chaque valeur de Y à la moyenne de Y, $\overline{Y}$) égale la somme des carrés de la régression (Écarts de chaque valeur *prédite* de y à la moyenne de Y, $\overline{Y}$) additionnée à la somme des carrés des résidus (Écarts de chaque valeur de Y à la valeur *prédite* par l'équation, $y_{i\ prédit}$).

$$(Y-\overline{Y})= (Y'-\overline{Y})+(Y-Y') \quad \text{où Y' est la valeur prédite de Y}$$

et

$$SC_{total}= SC_{reg}+SC_{res}$$

De la même manière que pour l'analyse de variance, les degrés de liberté se répartissent en degrés de liberté expliqués par les V.I. (un degré de liberté pour chaque variable indépendante) et en degrés de libertés résiduels (N-k-1 où N est le nombre de cas, et k le nombre de V.I.).

$$DL_{total}= DL_{reg}+DL_{res}$$

La variance est la somme des carrés divisée par les degrés de liberté. Il s'agit d'une certaine manière de la moyenne des écarts à la moyenne mis au carré.

$$CM=SC/DL$$

La valeur du test F que l'on retrouve dans la présentation des résultats est le rapport entre la variance due à la régression et la variance résiduelle.

$$F=CM_{reg}/CM_{res}$$

Le coefficient de détermination, R^2, est le rapport de la somme des écarts à la moyenne au carré (SC ou somme des carrés) expliqués par la régression divisée par la somme des carrés totale. Cette valeur constitue un indice de la *proportion de la variance totale expliquée* par les variables qui sont dans l'équation.

$$r^2=SS_{reg}/SS_{total}$$

2.4 Les trois grands types d'analyse, utilité et conséquences

a) L'analyse de régression standard

Dans l'analyse de régression standard, toutes les variables indépendantes sont entrées en même temps dans l'analyse. Cette méthode permet
- d'évaluer la variance expliquée par un ensemble de variables;
- d'évaluer la contribution unique de chaque variable entre autres en comparant les coefficients de corrélation simple, de corrélation partielle et partie;
- d'estimer la signification statistique de la contribution de chaque variable lorsque toutes les variables sont dans l'analyse.

b) L'analyse de régression hiérarchique/séquentielle

Ce type d'analyse permet de répondre aux questions concernant la contribution d'une variable ou d'un ensemble de variables au-delà de la contribution des variables qui sont déjà dans l'équation. Il est très utile pour vérifier la présence d'effet de médiation.

Elle permet de répondre à des questions théoriques du type: Est-ce que la langue d'enseignement explique le revenu au-delà du niveau de scolarité, est-ce que le sexe contribue à l'explication du revenu, à niveau de scolarité égal?

L'analyse de régression hiérarchique/séquentielle est similaire à l'analyse de covariance et donnera les mêmes résultats. On aura tendance à utiliser l'analyse de covariance lorsqu'il y a plusieurs (mais pas trop de) variables multinominales et lorsqu'il y a des possibilités connues ou théoriques d'effets d'interaction. Il est plus facile d'analyser les effets d'interaction avec l'analyse de covariance et on n'a pas à créer des variables dichotomiques avec les variables multinominales. Toutefois, l'analyse de covariance est moins appropriée ou devient plus difficile à utiliser lorsque les variables sont nombreuses.

Ce qui nous intéresse fortement dans les résultats de l'analyse hiérarchique/séquentielle, c'est la différence de variance expliquée lorsqu'on entre de nouvelles variables ou des ensembles de variables dans l'analyse. Nous désirons savoir aussi si cet ajout est significatif, c'est-à-dire s'il est susceptible d'ajouter à l'explication du phénomène à l'étude dans la population.

Avec la régression hiérarchique/séquentielle, on émet des hypothèses et on les vérifie. Une première étape consiste à faire un modèle des relations postulées entre les variables, ce qui facilite les analyses subséquentes.

c) L'analyse de régression statistique.

Ce type d'analyse est souvent utilisé à titre exploratoire et a trop souvent été utilisée à titre d'analyse finale et définitive. Dans la régression statistique, c'est en fait le BETA (coefficient de régression standardisé) qui détermine quelle variable sera inclue dans l'analyse et quand elle le sera. Il suffit donc d'une fraction dans les calculs pour que, dans le cas où deux variables d'importance théorique et empirique équivalente reliées entre elles, une seule des deux soit inclue dans l'analyse. Il faut souligner que cette méthode est intéressante à titre exploratoire et qu'elle donnera les mêmes résultats finaux que les autres méthodes lorsque les variables indépendantes sont peu reliées entre elles.

Ce type d'analyse est fortement dépendant de l'échantillon et nécessite donc un plus grand nombre de cas par variable (normalement 40).

La régression statistique donne le meilleur ensemble de prédicteurs statistiques parmi les V.I. considérées; c'est la prédiction maximale avec les V.I. que l'on a, mais non pas la prédiction optimale, particulièrement sur le plan théorique.

REMARQUES:

- Quelle que soit la méthode utilisée, si on retrouve les mêmes prédicteurs dans l'équation finale, les coefficients de régression seront les mêmes. Ce qui distingue les méthodes, c'est l'ordre d'entrée des variables, l'identité des variables qui seront gardées dans l'équation de prédiction (particulièrement quand il y a multi-collinéarité) et le type de questions auxquelles elles permettent de répondre.

- Il faut se souvenir que l'équation de régression est une addition. On postule donc que les effets sont additifs.

- Plus la combinaison de prédicteurs est adéquate, moins les résidus seront importants. L'analyse des résidus est donc essentielle. Elle permet de vérifier la justesse, la qualité, de la prédiction, d'identifier les problèmes quant aux postulats de l'analyse (normalité, linéarité, homoscédasticité, absence d'auto-corrélation) et d'examiner les valeurs extrêmes.

Les informations qui nous intéressent sont:

- R et R^2.

- Test F de signification de R^2.

- Coefficients de régression (b), erreurs-type des coefficients et coefficients standardisés (BETA).

- Test T de signification des b_i (= b/es(b)).

- Changement de R^2 après l'ajout d'une variable -- régressions statistique ou hiérarchique/ séquentielle -- ou d'un groupe de plusieurs variables (régression hiérarchique/séquentielle).

- Corrélations de départ entre les variables indépendantes et dépendante.

- Corrélation semi-partielle et partielle (surtout en régression standard).

- Patrons et graphiques des résidus.

2.5 La régression, la présentation et l'interprétation des informations

Qu'est-ce que la corrélation multiple (R)?

La corrélation multiple est une indice standardisé, variant entre -1 et +1, de la force de la relation entre l'ensemble des variables indépendantes et la variable dépendante. C'est la corrélation entre les valeurs prédites et les valeurs réelles. La corrélation multiple s'interprète comme la corrélation simple: Plus la corrélation est élevée, plus la relation linéaire entre les variables indépendantes et la variable dépendante est élevée.

"Il existe une relation forte (r=.45) entre l'ensemble des variables indépendantes et le niveau de revenu".

Qu'est-ce que le coefficient de détermination (la corrélation multiple au carré)?

Le coefficient de détermination est un indice de la proportion de variance de la variable dépendante expliquée par les variables indépendantes qui sont dans l'équation. Ainsi, on dira que les variables entrées dans l'équation expliquent 25% de la variance de la variable dépendante.

"Le bloc des variables socio-démographiques explique 5% de la variance du niveau de revenu".

Qu'est que l'ajout de corrélation multiple au carré (ΔR^2)?

Ce qu'on appelle le changement de R^2 indique la proportion de l'explication de la variance de la variable dépendante ajoutée par la/les variables indépendantes qui sont entrées dans l'équation.

"Le lieu de résidence, urbain ou rural, explique 10% de la variance du niveau de revenu, au-delà de l'explication fournie par le bloc des variables socio-démographiques (5%)".

Que signifie le test F?

La valeur du test F indique si la variance ou l'ajout de variance expliquée sont significatifs, c'est-à-dire si la relation constatée est susceptible d'exister dans la population et n'est pas due simplement au hasard de l'échantillonnage.

Au-delà de la variance déjà expliquée par le bloc des variables socio-démographiques, le lieu de résidence ajoute de façon significative à la prédiction du niveau de revenu tel qu'en témoigne le test F ($F(dl_{reg}, dl_{res})=$, $p=.002$).

"On peut rejeter l'hypothèse que la relation constatée dans l'échantillon est due au hasard"

Qu'est-ce qu'un coefficient de régression?

Le coefficient de régression ‘ordinaire’ (non standardisé) indique quelle est l'augmentation ou la diminution prédite, en moyenne, dans la variable dépendante à chaque unité d'augmentation de la variable indépendante. Dans une régression multiple, il s'agit de l'augmentation prédite **toutes choses égales par ailleurs,** c'est-à-dire comme si toutes les autres variables avaient une valeur fixe. **Les coefficients des différentes variables ne peuvent être comparés entre eux puisqu'ils dépendent de l'échelle de mesure de chaque variable (sauf si les échelles de mesure sont les mêmes).**

Un coefficient de régression qui a une valeur de 2 indique qu’à chaque fois que la valeur de la variable indépendante augmente de 1, la valeur de la variable dépendante augmente de 2, toutes choses égales par ailleurs.

Si la variable dépendante est le niveau de revenu (sur une échelle ordinale de 1 à 10, par tranche de 20,000$) et la variable indépendante le niveau d’éducation (sur une échelle ordinale de 1 à 7):

"Le coefficient de régression "b" de 0,5 signifie qu'à chaque augmentation de 1 dans l’échelle du niveau d’éducation, le niveau de revenu prédit est de ½ point plus élevé; il faut donc une augmentation de 2 points sur l’échelle de niveau d’éducation pour que le niveau de revenu prédit soit de 1 point plus élevé, ce qui correspond à un niveau moyen de 20,000$ de plus."

Qu'est-ce qu'un coefficient standardisé (Beta)?

Le coefficient standardisé permet de comparer la contribution des variables entre elles puisqu'il s'agit du coefficient de régression ramené sur une échelle standard (variant de -1 à +1).

"Le coefficient standardisé de .5 pour la variable mesurant le niveau scolarité est le plus haut coefficient; cette variable est donc celle qui contribue le plus à la prédiction de la satisfaction en emploi."

Que signifient les tests T pour les coefficients?

Les valeurs des tests T pour les coefficients sont calculées par la division de la valeur du coefficient de régression "b" par son erreur-type. Cette valeur doit être plus grande que 2 (≈ 1.96 écarts-type) pour être significative. Elle indique si chacun des coefficients des variables présentes dans l'équation sont significatifs, c'est-à-dire si, quelque soit l'importance de la contribution de chaque variable, cette contribution est susceptible d'exister vraiment dans la population à laquelle on veut inférer les résultats. Il faut souligner que cette information s’inscrit dans l'univers des variables présentes dans l'équation; la contribution d'une variable est considérée – ou non – comme significative compte tenu de la présence des autres variables dans l'équation.

La valeur du test T pour le coefficient de régression de l'âge (T= "≥2", p=.03) indique que la contribution de cette variable à l'explication du niveau de revenu est significative.

"On peut rejeter l'hypothèse que la relation constatée dans l'échantillon est due au hasard"

Que signifie la corrélation semi-partielle (partie) dans la régression standard?

La corrélation semi-partielle dans la régression standard représente la contribution unique d'une variable à l'explication de la variable dépendante, compte tenu des autres variables présentes.

La corrélation semi-partielle (r=.02) entre l'âge et le niveau de revenu montre que l'explication apportée par l'âge seul est peu importante. La corrélation relativement forte (r=.50) entre l'âge et le niveau de revenu s'explique donc presque entièrement par les autres variables présentes dans l'équation, nommément le niveau de scolarité et surtout, la région de travail.

Qu'est-ce qu'un résidu?

Le résidu, c'est l'écart entre chaque valeur de la variable dépendante et la valeur que l'on a prédite sur cette variable étant donné les valeurs des variables indépendantes. Plus cet écart est important, moins la prédiction est adéquate; lorsqu'un résidu est plus grand que 3,16, on dit qu'il s'écarte anormalement de la distribution des résidus. Cette distribution devrait approcher celle de la distribution normale. Elle devrait aussi être la même quelles que soient les valeurs des variables indépendantes ou dépendante.

"Trois résidus sont supérieurs à 3,16; un est très supérieur. En examinant ce cas de façon plus poussée, il est apparu qu'il possédait des caractéristiques particulières.... Si le cas est retiré de l'analyse, les valeurs des coefficients sont légèrement modifiées, surtout pour la variable X_7 , et il n'y a plus de résidus plus grand que 3,16."

ou:
"Un examen attentif des résidus montre que ceux-ci se distribuent normalement et qu'aucun résidu ne présente une valeur statistiquement trop élevée. Ceci amène à conclure que la prédiction est valide et appropriée pour tous les patrons de réponse."

L'interprétation:

L'interprétation fait référence à la problématique de recherche, à la population, à la "vraie vie". Elle réfère aux hypothèses de départ et peut nous permettre, par exemple, de conclure sur des interventions à effectuer pour régler le problème qui était à la source de notre étude, les nouvelles recherches qu'il faudrait effectuer pour améliorer la compréhension de la situation, les raisons qui peuvent expliquer que les résultats présentés sont différents de ceux présentés par d'autres chercheurs auparavant.

"Les résultats ont montré que l'âge est un prédicteur important du niveau de revenu et qu'en fait, une bonne partie de l'explication attribuée à l'âge est dûe à l'acquisition d'un niveau de scolarité plus élevé chez les jeunes générations. Notre étude démontre la pertinence des politiques visant à faciliter l'accès à l'éducation pour réduire les problèmes de pauvreté dans la population de cette étude."

Commandes utiles pour la régression dans SPSS:

Dans SPSS, pour la relation entre deux variables:

On peut demander un diagramme de dispersion (dans le menu des graphiques).

Après avoir fait produire le graphique, on peut obtenir la droite de régression, demander le r^2 et l'intervalle de confiance de la droite; on peut modifier les largeurs, mettre des titres, etc. Cela vaut toutefois uniquement pour la relation entre deux variables.

Commandes de régression avec SPSS:

Voilà de quoi auront l'air les commandes une fois toutes les options, statistiques, graphiques, demandés ou édités:

→Régression standard

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA HISTORY ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT result
  /METHOD=ENTER pretest restot
  /SCATTERPLOT=(result ,*ZPRED ) (*ZPRED ,*ZRESID )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

******Note: Lorsque l'on veut une régression standard, il s'assurer d'avoir la mention "ZPP" dans la sous-procédure /STATISTICS , ce qui permet d'obtenir les corrélations semi-partielles, sinon la rajouter à la main.***

→Régression hiérarchique/ séquentielle

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA HISTORY CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT result
  /METHOD=ENTER restot  /METHOD=ENTER pretest
  /SCATTERPLOT=(result ,*ZPRED ) (*ZPRED ,*ZRESID )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

******Note: Lorsque l'on veut une régression hiérarchique, il faut s'assurer d'avoir la mention"CHANGE" dans la sous-procédure /STATISTICS, ce qui permet d'obtenir les informations sur la variance supplémentaire expliquée à chaque étape, sinon la rajouter à la main.***

→Régression statistique (pas à pas):

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA HISTORY CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT result
  /METHOD=STEPWISE pretest restot
  /SCATTERPLOT=(result ,*ZPRED ) (*ZPRED ,*ZRESID )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

******Note: Lorsque l'on veut une régression statistique, il faut s'assurer d'avoir la mention"CHANGE" dans la sous-procédure /STATISTICS, ce qui permet d'obtenir les informations sur la variance supplémentaire expliquée à chaque étape, sinon la rajouter à la main.***