**Régression multiple**

|  |  |
| --- | --- |
|  |  |

 **Rappel théorique**

En général, les modèles de régression sont construits dans le but d’expliquer (ou prédire, selon la perspective de l’analyse) la variance d’un phénomène (variable dépendante) à l’aide d’une combinaison de facteurs explicatifs (variables indépendantes). Dans le cas de la régression linéaire multiple, la variable dépendante est toujours une variable continue tandis que les variables indépendantes peuvent être continues ou catégorielles. La régression linéaire est appelée multiple lorsque le modèle est composé d’au moins deux variables indépendantes. À l’inverse, un modèle de régression linéaire simple ne contient qu’une seule variable indépendante. Comme il est excessivement rare, voire impossible, de prédire un phénomène à l’aide d’une seule variable, cette section porte sur la régression linéaire multiple. Cependant, tout le contenu s’applique également aux résultats d’une régression simple.

Nous allons donc voir maintenant comment il est possible d’expliquer (ou de prédire) la variance d’une variable dépendante à l’aide d’une combinaison linéaire de variables indépendantes à partir de la généralisation de l’équation algébrique utilisée dans le module sur la régression simple.

Les questions auxquelles la régression linéaire multiple permet de répondre sont nombreuses. Par exemple,

* *Quelles variables permettent de prédire les symptômes anxieux ?*
* *De combien le risque de chutes va diminuer chez les personnes âgées lorsqu’elles participent à des exercices de groupe, un suivi individuel et changent leurs habitudes de vie ?*
* *Est-ce que la satisfaction au travail varie en fonction de l’augmentation des défis à relever et de l’esprit d’équipe ?*
* *Quelle proportion de la variance du taux de décrochage est expliquée par la combinaison des variables prédictives ?*

**Hypothèse nulle**

L’hypothèse nulle est qu’il n’y a pas de relation linéaire entre la combinaison des variables indépendantes (X1, X2, X3… Xn) et la variable dépendante (Y).

L’hypothèse de recherche est l’inverse, soit que la combinaison des variables indépendantes est associée significativement à la variable dépendante.

**Les prémisses**

1. Les **types de variables à utiliser**:

Indépendantes : continue ou catégorielle (ordinale ou dichotomique)

Dépendante : continue

2. **Pas de variance égale à zéro** : la distribution des prédicteurs doit comprendre une certaine variance, donc ne doit pas être constante.

3. **Aucune multicolinéarité parfaite** : il ne doit pas y avoir de relation linéaire parfaite entre deux ou plusieurs variables indépendantes. Par conséquent, les corrélations ne doivent pas être trop fortes entre celles-ci. Cette prémisse peut être vérifiée avec le VIF (*Variance Inflation Factor*) indiquant si une variable indépendante a une une relation linéaire forte avec les autres. La règle arbitraire souvent appliquée veut qu’une valeur de cet indice plus grande que 10 indique la présence d’un tel problème.

4. **Pas de corrélation** entre les **variables indépendantes** et les **variables externes** : les variables d’influence doivent toutes être incluses dans le modèle.

5. **Homéocédasticité** (homogénéité des variances des résiduels) : la variance des valeurs résiduelles doit être similaire à tous les niveaux de la variable indépendante.

6. **Indépendance des erreurs** : les valeurs résiduelles ne doivent pas être corrélées entre les individus. Cette prémisse peut être vérifiée avec la statistique Durbin-Watson qui se situe entre 0 et 4, une valeur de 2 indiquant une absence de corrélation, moins de 2 une corrélation positive et plus de 2, une corrélation négative. La règle arbitraire cette fois est que la valeur ne doit pas être plus petite que 1 ou plus grande que 3.

7. **Distribution normale des résiduels** : bien que les variables indépendantes ne doivent pas nécessairement suivre une distribution normale, il importe que les résiduels en suivent une. Ils doivent donc avoir une moyenne de 0, la majorité des valeurs doivent s’en rapprocher. Cette prémisse peut être vérifiée en enregistrant les valeurs résiduelles dans la base de données et en effectuant le test de Kolmogorov-Smirnov ou de Shapiro-Wilks, disponible dans les options de la commande *Explorer*. Vous devez vous assurer que le test n’est pas significatif pour conserver l’hypothèse nulle de distribution normale.

8. **Indépendance de la variable prédite** : toutes les observations formant la distribution des valeurs de la variable dépendante sont indépendantes, viennent d’un individu différent.

9. **Relation linéaire entre** les **variables indépendantes** et la **variable dépendante** : la variation de la variable dépendante pour chaque augmentation d’une unité d’une variable indépendante suit une ligne droite.

**Le modèle multivarié**

De manière générale, les modèles statistiques se présentent globalement ainsi :

Observationi : (Modèlei) + erreuri

Chaque valeur de la variable dépendante (Observationi) peut être expliquée en partie par un modèle statistique. La partie que le modèle ne peut expliquer est l’erreur spécifique associée à cette valeur.

Dans le cas de la régression linéaire simple, ce modèle général peut se décliner plus précisément ainsi :

Yi : (b0 + b1X1) + εi

Où Y représente les valeurs possibles de la variable dépendante qui peuvent être expliquées par le modèle général de régression. Encore une fois, la portion qui ne peut être expliquée par le modèle est symbolisée par εi qui représente l’erreur commise par le modèle pour chaque valeur de Y.

L’équation de la régression linéaire multiple est en fait la généralisation du modèle de régression simple.

Yi : (b0 + b1X1 + b2X2 + … + bnXn) + εi

On observe que chaque variable indépendante (X) est multipliée par son propre coefficient bêta (b) qui sous sa forme standardisée correspond à sa contribution relative dans le modèle. La constante (b0) correspond à la valeur de la variable dépendante lorsque toutes les variables indépendantes égalent 0. On appelle aussi b0 l’ordonnée à l’origine. Associée de près à l’évaluation du modèle, l’indice de corrélation multiple *R2* représente le pourcentage de variance expliquée par le modèle (la combinaison des variables indépendantes).

**La conception d’un modèle de régression**

La conception d’un modèle de régression ne devrait jamais être prise à la légère. Elle devrait faire l’objet d’une réflexion préalable portant sur 1) le choix des variables indépendantes et 2) le choix de la méthode de régression.

**Le choix des variables indépendantes**

En tout temps, le choix des variables indépendantes doit être guidé par le principe de parcimonie qui veut qu’un bon modèle comprend un nombre optimal de variables et par la présence d’un lien théorique connu ou présumé avec la variable dépendante. Voici d’autres éléments à considérer lors du choix des variables indépendantes. Cette liste n’est pas exhaustive, mais souligne l’importance des éléments à considérer lors de cette étape.

1. La nature des objectifs ou des hypothèses de recherche : Les variables mises en cause dans l’énoncé d’un objectif ou d’une hypothèse doivent forcément se retrouver dans le modèle. L’énoncé peut également avoir un impact sur le choix de la méthode de régression.

2. La présence de variables confondantes : Il est possible que certaines variables n’apparaissant pas dans l’énoncé de l’objectif ou de l’hypothèse soient importantes dans un modèle dans la mesure où elles peuvent influencer les résultats. On appelle ces variables « confondantes » et leur inclusion dans le modèle permet de contrôler statistiquement leur effet.

3. La présence de corrélation avec la variable dépendante : Dans certains contextes, il est possible de choisir les variables indépendantes en fonction de leur degré d’association avec la variable dépendante. Des variables n’ayant pas de lien assez fort avec celle-ci pourrait être exclues du modèle.

4. La puissance statistique du devis : Cohen (1992) et Hair et al. (2005) ont bien démontré que le nombre d’observations détermine la quantité maximale de variables qu’un modèle peut supporter. Plus on a d’observations, plus on peut inclure de variables dans le modèle.

**Le choix de la méthode de régression**

La méthode de « construction » d’un modèle de régression nécessite également une réflexion préalable. En effet, la méthode choisie ne sera pas la même selon que l’on désire tester un modèle théorique précis, contrôler l’effet de variables confondantes ou tout simplement explorer une combinaison particulière de variables indépendantes. De même, la façon d’introduire les variables ou les blocs de variables indépendantes dans ce modèle doit faire également l’objet d’une justification rationnelle.

Dans un premier temps, on doit choisir une des deux stratégies suivantes : la modélisation globale ou la modélisation par blocs. Dans le premier cas, la combinaison de toutes les variables est évaluée globalement. Dans le second, les variables sont regroupées en bloc et les résultats évaluent le modèle global ainsi que la contribution de chaque bloc.

Dans un deuxième temps, on doit également déterminer la manière dont les variables indépendantes seront insérées dans le modèle global ou dans les blocs : par entrée forcée ou par entrée progressive.

Voyons maintenant la description plus précise des méthodes pour lesquelles vous pourriez opter une fois que les variables indépendantes ont été choisies.

1. La régression hiérarchique (*hierarchical regression*)

Cette méthode permet au chercheur de déterminer l’ordre d’entrée des variables dans le modèle à l’aide de la création des blocs de variables qui seront entrés de manière hiérarchisée dans le modèle. Ceci permet d’observer plus en détail comment se comporte le modèle. Les résultats indiquent l’apport de chaque bloc en termes de pourcentage de variance expliquée (*R2*). Pour les blocs constitués de plus d’une variable, il est possible de faire entrer celles-ci en un seul temps (entrée forcée) ou progressivement (voir plus bas).

2. La régression avec entrée forcée

Cette fois-ci, toutes les variables évaluées sont entrées au même moment et un test F évalue l’ensemble du modèle. Le choix des variables à inclure repose encore sur la théorie. Par contre, le chercheur n’influence pas l’ordre d’entrée des variables. Le modèle évalue donc leur effet combiné.

3. La régression avec entrée progressive

Contrairement aux deux autres méthodes, la sélection des variables à inclure est basée sur un critère mathématique. Une fois les variables indépendantes choisies, leur inclusion dans le modèle dépendra de leur contribution mathématique à son amélioration. Il existe trois méthodes progressives.

La première est la méthode **ascendante** (*forward*). Dans ce cas, le modèle initial ne contient que la constante (b0). Celui-ci servira de base de comparaison pour déterminer si l’ajout d’une variable contribue significativement à l’amélioration du modèle. SPSS choisit parmi les variables indépendantes soumises celle qui a la plus forte corrélation avec la variable dépendante. Il évalue si cet ajout est significatif. Si c’est le cas, il intègre une deuxième variable. Cette dernière a la plus forte corrélation partielle avec la variable dépendante. On parle de corrélation partielle puisque le calcul est effectué avec la variance de la variable dépendante qui reste à expliquer une fois que l’effet de la première variable est retiré. SPSS évalue ensuite si l’ajout de cette variable est significatif. Si c’est le cas, il la retient et détermine s’il peut ajouter un 3e prédicteur. Il cesse d’inclure des nouvelles variables lorsque l’augmentation de la valeur de *R2* n’est plus significative.

La deuxième est la méthode **pas-à-pas** (*stepwise*). Celle-ci ressemble beaucoup à la méthode ascendante, puisque le choix de la première variable est encore basé sur la corrélation la plus élevée et celui des variables suivantes sur la corrélation partielle. Toutefois, lorsque SPSS ajoute une variable au modèle, il évalue si elle apporte une contribution significative, mais également si celle qui contribuait le moins au modèle demeure significative. Si ce n’est pas le cas, il la retire. De cette manière, il est possible d’éliminer les variables redondantes.

Enfin, la dernière est la méthode **descendante** (*backward*). Dans ce cas, le modèle initial comprend toutes les variables, comme pour la régression forcée. SPSS va cette fois retirer la variable ayant la plus faible contribution au modèle si la variation du *R2* n’est pas significative en l’éliminant. La procédure va être répétée jusqu’à ce que toutes les variables conservées contribuent significativement à l’amélioration du *R2*.

**Quelle méthode privilégier ?**

Parmi toutes ces méthodes, laquelle devrions-nous privilégier ? De manière générale, on suggère qu’un modèle bien balisée par la théorie devrait utiliser une stratégie globale avec une méthode d’entrée forcée, hiérarchisée ou non. Pour les travaux de nature davantage exploratoire, les méthodes progressives sont adaptées. Parmi les trois présentées, on privilégiera la méthode descendante, car il y a plus de risques de commettre des erreurs de type II avec la méthode ascendante. Celle-ci ne tient pas compte des variables significatives lorsqu’elles sont combinées et peut donc plus facilement oublier une variable qui affecte la variable dépendante en présence d’un autre prédicteur.

Enfin, un bon modèle sera parcimonieux, constitué de variables ayant une pertinence théorique et expliquera une proportion satisfaisante de la variance de la variable dépendante.

La régression hiérarchisée est intéressante lorsque le modèle comporte plusieurs variables qui peuvent être théoriquement regroupées ou lorsque certaines variables doivent être contrôlées statistiquement (ex. : variables socioéconomiques). SPSS permet de regrouper ces variables en « blocs » dont l'ordre d'inclusion devrait représenter leur position relative (proximale ou distale) par rapport à la variable dépendante. Le premier bloc doit contenir les variables contrôles ou encore les variables proximales et les blocs subséquents comprennent les variables de plus en plus distales. SPSS donne les résultats pour le modèle global (toutes les variables) ainsi que l'apport spécifique de chaque bloc une fois l'effet du bloc précédent considéré.

**Le diagnostic des observations**

Une fois que la méthode de régression est choisie, il est important également de considérer si le modèle qu’on va obtenir est bien ajusté aux données ou s’il est influencé par la présence de valeurs extrêmes, qui s’écartent beaucoup des autres observations.

Comme pour toute analyse statistique, il est préférable d’examiner au préalable les distributions des variables qui seront mises en cause. Les procédures descriptives permettent entre autres d’identifier les valeurs extrêmes. Ces dernières influencent grandement le modèle, elles peuvent faire varier les coefficients beta de l’équation qui sera, de ce fait, moins précise. Il importe donc de savoir si des valeurs extrêmes sont présentes.

Une autre stratégie simple consiste à déterminer pour quelles observations les valeurs résiduelles sont importantes. En effet, si une valeur extrême est présente, son score prédit sera très différent de la valeur observée. En corollaire, on peut ajouter que plus les valeurs résiduelles de l’ensemble des observations sont petites, mieux le modèle de régression est ajusté aux données.

Les valeurs résiduelles sont calculées dans la même unité de mesure que la variable originale. Afin de faciliter la comparaison entre les modèles, on transforme ces valeurs en score Z (résiduels standardisés), ce qui nous permet plus facilement d’identifier quelles sont les valeurs très éloignées du modèle. On se base sur les balises de la courbe normale pour déterminer quelles sont les valeurs extrêmes. Dans un modèle bien ajusté, on s’attend à trouver

Moins de 5 % des résiduels standardisés ayant une valeur > 1,96 ou < -1,96
Moins de 1 % des résiduels standardisés ayant une valeur > 2,58 ou < -2,58
Aucun résiduel standardisé ayant une valeur de > 3,29 ou < -3,29

Sinon, on doit porter une attention très particulière aux résiduels ayant des valeurs de plus de 3 (3,29), puisque dans un échantillon normal, il est très peu probable que de tels écarts arrivent au hasard. Lorsque les seuils proposés sont dépassés, on peut penser que le modèle ne représente pas bien les données.

À noter qu’il est possible d’enregistrer les valeurs résiduelles et les résiduels standardisés dans des nouvelles variables dans la base de données dans les options disponibles dans SPSS.

On peut également enregistrer la distance de Cook qui nous indiquera l’influence de chaque observation sur le modèle total. Si la distance pour une observation est de plus de 1, elle influence probablement l’estimation des coefficients beta du modèle (Cook et Weisberg, 1982). Pour connaître son influence exacte, il faut refaire la régression sans cette observation et comparer les coefficients beta obtenus. La statistique *DFBeta* évalue la différence entre les deux. On utilise généralement la valeur *DFBeta* standardisée pour voir si la différence est importante. Les valeurs plus grandes que 1 indiquent une influence importante de l’observation sur les paramètres.

Le modèle de régression le plus précis et le mieux ajusté sera évidemment celui sans valeurs extrêmes et sans valeurs qui influencent grandement l’estimation des paramètres.

**Procédure 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.



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](http://spss.espaceweb.usherbrooke.ca/pages/stat-inferentielles/regression-simple/procedure-spss.php) 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 R2 correspondant et le R2 ajusté

Variation de R-deux : changement du R2 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).



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.

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.

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.

|  |  |
| --- | --- |
|  |  |

 **Interprétation**

|  |  |
| --- | --- |
|  |  |

 Dans notre exemple, nous voulons savoir quelles variables influencent le salaire annuel d’un employé (SALAIRE). La théorie nous indique que le nombre d’années de scolarité a une importante influence (EDUC). Nous désirons savoir si le sexe des employés (SEXE) et le nombre de mois d’expérience dans l’entreprise (DURÉE) exercent également une influence. Nous avons donc choisi un modèle de régression hiérarchique (comprenant deux blocs de variables) avec la méthode entrée pour la première étape (bloc 1 avec EDUC), mais la méthode ascendante pour la deuxième (bloc 2 avec SEXE et DURÉE), dans le but d'illustrer les différentes possibilités de modélisation.

**Statistiques descriptives**

Examinons d’abord les statistiques descriptives. Nous voyons que l’étude a été menée auprès de 474 employés qui gagnent en moyenne actuellement près de 35 000 $. Ils travaillent depuis environ sept ans pour leur entreprise (81 mois) et ont en moyenne 13 ans de scolarité. Bien entendu, la moyenne des hommes et des femmes n’est pas une donnée pertinente, mais elle indique que la variable a été codée 1-2. Idéalement, les variables dichotomiques doivent toujours être codées 0-1 pour en faciliter l'interprétation.



Le deuxième tableau fourni par SPSS concerne les corrélations entre les variables étudiées. Nous voyons qu’il y a une corrélation très élevée et significative entre le salaire et le nombre d’années de scolarité, ainsi qu’entre le sexe et le salaire. On doit porter attention aux relations entre les variables indépendantes. Si la corrélation entre deux de ces variables se situait à 0,9 (ou – 0,9), il y aurait un risque important de multicolinéarité. Nous aurions introduit deux variables qui mesurent sensiblement la même chose pour prédire le salaire actuel. Nous voulons éviter cette situation.

 

 **Variables introduites/éliminées**

Le tableau suivant présente les variables retenues dans les deux étapes du modèle. On constate que la variable EDUC est forcément présente puisque nous avions choisi la méthode Entrée. Pour le deuxième bloc du modèle, SPSS a retenu la variable SEXE avec notre critère de sélection (la probabilité F est significative à p < 0,05) et a exclu la variable DURÉE, car la valeur F associée au coefficient *b* n'atteint pas le seuil de signification.



**Étape 1 : Évaluation de la qualité du modèle de régression**

Tout comme la régression simple, l’interprétation débute en évaluant la qualité du modèle. On vérifie si la première étape du modèle explique significativement plus de variabilité qu’un modèle sans prédicteur (VI). Ensuite, il s'agit de s'assurer que toutes les variables introduites contribuent à améliorer significativement la variabilité expliquée par le modèle final.

**Analyse de variance**

Le tableau d’ANOVA nous donne cette information. Il nous permet de déterminer si nous rejettons l'hypothèse nulle (H0) ou non. Dans notre exemple, nous voulons savoir dans un premier temps si le nombre d'années de scolarité prédit mieux le SALAIRE que ne le fait un modèle sans prédicteur (avec seulement la moyenne) et dans un deuxième temps, si le nombre d’années de scolarité et le sexe prédisent mieux le SALAIRE qu’un modèle sans prédicteur. L'hypothèse nulle est donc que les deux modèles sont équivalents à la moyenne du salaire.



On constate à la lecture du tableau que selon la valeur F obtenue pour les deux modèles, on peut rejeter l’hypothèse nulle. En effet, les valeurs de 365,38 et de 225,51 sont significatives à p < 0,001, ce qui indique que nous avons moins de 0,1 % de chance de se tromper en affirmant que les modèles contribuent à mieux prédire le salaire que la simple moyenne.

**Étape 2 : Évaluation de l'ajustement du modèle de régression aux données**

Maintenant que l’on sait que le modèle est significatif, le tableau récapitulatif des modèles permet de déterminer la contribution de chaque bloc de variables. Ce tableau indique le R2 cumulatif à chaque étape du modèle (colonne R-deux)) ainsi que l'apport spécifique de chaque bloc (colonne Variation de R-deux).



Le tableau contient donc plusieurs informations utiles. Premièrement, la valeur de la corrélation multiple (R) correspond à l'agglomération des points dans la régression simple. Elle représente la force de la relation entre la VD et la combinaison des VI de chaque modèle. Des valeurs de 0,66 et 0,70 suggèrent que les données sont ajustées de manière satisfaisantes au modèle.

Ensuite, la signification du R2 est évaluée en fonction de l'apport de chaque étape. La variation de F associée au premier modèle est significative (p < -.001). Ce modèle explique donc une proportion significative de la variance de la variable SALAIRE. Nous sommes passés de R2 = 0 à R2 = 0,436. Le deuxième modèle fait passer le R2 de 0,436 à 0,489. Cette variation de 0,053 apparait comme significative. En effet, la valeur de F est calculée à partir de la varaition du R2 entre les étapes. SPSS détermine donc si la différence (0,053) entre le R2 du modèle 2 (0,489) et celui du modèle 1 (0,436) est significative. Cette fois-ci, c'est le cas (p < 0,001). Chaque étape contribue donc significativement à l'amélioration de l'explication de la variabilité de la VD.

La dernière colonne concerne le test de Durbin-Watson, il n'y a pas de seuil de signification associé, seulement la valeur de la statistique qui est acceptable lorsqu'elle se situe entre 1 et 3. Il est convenu que plus la valeur est près de 2, moins il y  a de problème au niveau de l'indépendance des erreurs. Avec une valeur de 1,96, nous pouvons croire que nous respectons cette prémisse.

**Étape 3 : Évaluation de la variabilité expliquée par le modèle de régression**

Enfin, on se rappelle que la valeur du R2, lorsqu'elle est multipliée par 100, indique le pourcentage de variabilité de la VD expliquée par le modèle (les prédicteurs). Les résultats suggèrent que 43,6 % du salaire est expliqué par le nombre d'années de scolarité et que 48,7 % du salaire est expliqué par la combinaison de la scolarité et du sexe de l'employé.

**Étape 4 : Évaluation des paramètres du modèle**

Maintenant que nous savons que notre modèle est significatif et que le deuxième est celui qui explique le plus de variance, il est possible de construire l’équation de régression pour prédire une valeur de Y. L’équation de base était la suivante :

Yi : (b0 + b1X1 + b2X2 + … + bnXn) + εi

Remplaçons maintenant les b par les coefficients fournis dans le tableau ci-dessous.

Yprédit = (-7900,99 + 3391,68educ – 8423,46sexe)

Pour un homme ayant complété 16 années de scolarité, nous obtiendrions un salaire prédit de...

Yprédit = (-7900,99 + 3391,68\*16 – 8423,46\*1)
Yprédit = 37 942,43 $ par année

**L**e signe du coefficient nous indique le sens de la relation. Dans notre cas, plus le nombre d’années de scolarité augmente, plus le salaire augmente. Nous voyons aussi que quand le sexe diminue (passant de 2 pour les femmes à 1 pour les hommes), le salaire augmente.

Le coefficient nous informe également sur le degré auquel chaque prédicteur influence la VD si tous les autres prédicteurs sont constants. Par exemple, chaque année de scolarité de plus est associée à 3 391,68 $ de plus annuellement.

L’erreur standard nous renseigne sur la variabilité du coefficient dans la population. Elle permet également de calculer la valeur de t. Cette dernière nous indique si le coefficient est significatif. Alors que le tableau sur le récapitulatif des modèles confirmait si chaque modèle était significatif, la signification de t nous permet de répondre à la question « est-ce que le *b* du prédicteur est différent de 0 ? », donc si chaque variable contribue significativement au modèle. Plus la valeur de t est élevée et plus celle de p est petite, plus le prédicteur contribue au modèle. Nous constatons donc que les deux variables sont significatives, mais que la variabilité expliquée par le nombre d’années de scolarité est plus importante que celle expliquée par le sexe.

La valeur du Beta standardisé (β) apporte aussi une information intéressante. Elle indique le changement en écart-type de la VD pour chaque augmentation d’un écart-type de la VI quand toutes les autres valeurs sont constantes. Par exemple, la valeur d’un écart-type du salaire est de 17 075,66 $ et celle d’un écart-type de scolarité est de 2,89. Nous pouvons donc savoir que l’augmentation d’un é.-t. de la scolarité (2,89) est associé à l’augmentation de 0,57 é.-t. du salaire (0,57\*17 075,66 = 9 733,13). Par conséquent, chaque fois que l’on étudie 2,89 années de plus, le salaire augmente de 9 733,13 $.

Pour les intervalles de confiance, nous voulons obtenir les valeurs les plus rapprochées pour que le modèle soit le plus près possible des données réelles de la population. Il ne faut absolument pas que la valeur 0 se situe entre les deux intervalles, car cela signifierait qu'une différence de 0 est une valeur possible. DAns ce cas, la valeur de t forcément non significative.



Ce tableau présente également la valeur des corrélations et des corrélations partielles. Ce sont ces valeurs sur lesquelles se base SPSS lorsqu’il choisit d’introduire des variables lorsque l’on sélectionne une méthode d'entrée progressive. La première variable est choisie à partir de la corrélation simple la plus forte (ici 0,661 pour EDUC). Le choix des variables suivantes est par contre basé sur la corrélation partielle, c'est-à-dire la plus forte corrélation entre les variables toujours disponibles et la partie de variance qui reste à expliquer une fois que l’on a retiré ce qui est expliqué par le premier prédicteur.

Finalement, la valeur VIF (ou la tolérance, soit l’inverse du VIF (1/VIF)) permet de vérifier la prémisse de multicolinéarité. Nous cherchons à obtenir une valeur VIF près de 1. Si elle est de 10, c’est problématique. Conséquemment, si la tolérance est équivalente à 0,1, il y a un problème sérieux . Probablement que les corrélations entre 2 variables prédictrices ou plus sont trop élevées.

**Les variables exclues**

Le tableau suivant présente les variables qui n’ont pas été retenues dans le modèle à chacune des étapes. La valeur *b* est estimée pour chaque variable si cette dernière était incluse. Si la valeur de t est significative, nous pouvons comprendre que l’ajout de cette variable contribuerait probablement à l’amélioration du modèle. SPSS évaluera d’abord l’apport de la variable ayant la valeur t la plus élevée dans une régression progressive. La corrélation partielle donne également un indice de la contribution au modèle. C’est la même valeur que l’on trouvait dans le tableau précédent.



**Le diagnostic des observations et la vérification des prémisses**

Ce dernier tableau est fourni grâce aux options sélectionnées préalablement. Il nous renseigne sur la présence de valeurs extrêmes qui influenceraient le modèle, donc sur la qualité de l’ajustement des données. Les valeurs extrêmes font varier les coefficients *b* et sont mal prédites par le modèle, donc elles sont associées à une valeur résiduelle importante. Comme nous avons vu précédemment dans le rappel théorique, nous ne voulons aucune valeur résiduelle standardisée de plus de 3,29 (ou de moins de -3,29), pas plus de 1 % de l’échantillon ayant une valeur de plus de 2,58 (ou de moins de -2,58) ainsi qu'un maximum de 5 % des observations ayant  une valeur de plus de 1,96 (ou de moins de – 1,96).



En observantt le diagnostic des observations, nous constatons que 7 individus ont des salaires de plus de 83 750 $. Ils s’écartent vraiment de la moyenne, la valeur résiduelle standardisée pour chacun est de plus de 3 écart-types. Les employés gagnant plus de 100 000 $ annuellement présentent un problème majeur. Il serait probablement judicieux de refaire l’analyse en excluant ces hauts salariés et de vérifier la variation des coefficients.

D’ailleurs, nous pouvons constater en regardant l’histogramme de la distribution des valeurs résiduelles que nous ne respectons pas la prémisse de normalité de distribution des erreurs. Nous souhaitons que la distribution suive une courbe normale, mais nous observons un pic prononcé ainsi que des valeurs éloignées de la courbe. Cette distribution n’est donc probablement pas normale. Nous pourrions confirmer avec le test de normalité de Shapiro-Wilks ou de Kolmogorov-Smirnoff. Ces tests sont disponbiles dans les options de la [procédure Explorer](http://spss.espaceweb.usherbrooke.ca/pages/stat-descriptives/procdure-explorer/procedure-spss.php). Cochez Graphes de répartition gaussiens avec tests. Cela nous confirme encore qu'il pourrait être judicieux de retirer ces valeurs extrêmes de l'analyse.

Finalement, nous pouvons tout de même jeter un coup d’œil aux prémisses d’homéodasticité et de linéarité avec le graphique de dispersion. Pour la première prémisse, les points doivent être répartis aléatoirement autour de 0 (ne pas former d’entonnoir), ce qui semble le cas ici, bien que les points soient répartis en colonnes. Pour la deuxième, nous voulons éviter que l’agglomération de points suive une courbe. Cette prémisse semble aussi respectée. Nous respectons donc la plupart des prémisses, le modèle est donc probablement valide, mais gagnerait certainement en précision en éliminant les valeurs extrêmes.

 