Théorique : LA REGRESSION LIGISTIQUE (Extrait)

La régression logistique propose de tester un modèle de régression dont la variable dépendante est dichotomique (codée 0-1) et dont les variables indépendantes peuvent être continues ou catégorielles. La régression logistique binomiale s'apparente beaucoup à la régression linéaire. Le poids de chaque variable indépendante est représenté par un coefficient de régression et il est possible de calculer la taille d'effet du modèle avec un indice semblable au coefficient de détermination (pseudo R^2). Toutefois, elle ne nécessite pas la présence d'une relation linéaire entre les variables puisque la variable dépendante est dichotomique.
Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à 0,5, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure à 0,5, il ne l'est pas.

Exemples de questions de recherche auxquelles peut répondre la régression logistique :

Est-ce que le nombre d'heures d'études, le niveau d'anxiété et le sexe permettent de prédire la réussite ou l'échec à un examen ?

Quelle est la probabilité de dépasser son poids santé en adoptant de mauvaises habitudes de vie ?

Généralement, les modèles de régression logistique comprennent plus d'une variable indépendante. Il s'agit donc d'une technique d'analyse multivariée.

Hypothèse nulle

L'hypothèse nulle générale est que la combinaison des variables indépendantes (le modèle) ne parvient pas à mieux expliquer la présence/absence de la variable dépendante qu'un modèle sans prédicteur. Comme c'était le cas pour la régression multiple, la confirmation de cette hypothèse nulle marque la fin de l'interprétation du modèle.

Lorsque cette hypothèse nulle est rejetée, ceci signifie qu'il y a au moins un prédicteur du modèle qui est associé significativement à la variable dépendante. Il faut alors interpréter les valeurs des coefficients du modèle (b_1, b_2, b_3…b_n) et déterminer lequel ou lesquels sont significatifs.

Les prémisses

1. Les **types de variables à utiliser** :
Indépendantes (prédicteurs) : continue ou catégorielles dichotomiques
Dépendante (prédite) : catégorielle dichotomique. Cette dernière doit être une vraie variable dichotomique et non une variable continue recodée en 2 groupes, ce qui serait associé à une importante perte d'information.
2. Inclure les **variables pertinentes** : toutes les variables pertinentes doivent être comprises dans le modèle et celles qui ne le sont pas, éliminées.

3. **Indépendance des observations (VD) et des résiduels** : un individu ne peut pas faire partie des deux groupes de la VD (par exemple avec des mesures pré-post-test).
4. Relation linéaire entre les VI et la transformation logistique de la VD
5. **Aucune multicolinéarité parfaite ou élevée** : il ne doit pas y avoir de relation linéaire parfaite, ni très élevée entre deux ou plusieurs prédicteurs. Par conséquent, les corrélations ne doivent pas être trop fortes entre ceux-ci.
6. **Pas de valeurs extrêmes des résiduels** : comme dans la régression multiple, des valeurs résidulelles standardisées plus élevées que 2,58 ou moins élevées que -2,58 influencent les coefficients du modèle et limitent la qualité de l'ajustement.
7. **Taille de l'échantillon** : l'échantillon doit être suffisant pour que l'on puisse procéder à l'analyse. On suggère minimalement 10 observations par variable indépendante (Hosmer et Lemeshow, 1989, voir également Cohen, 1992).
8. **Échantillon adéquat** pour les prédicteurs catégoriels : lorsqu'une VI catégorielle est croisée avec la VD, aucune cellule ne doit avoir moins d'une observation et un maximum de 20 % des cellules peuvent comprendre 5 observations ou moins.

Le modèle : parallèle entre la régression multiple et logistique

Nous avons vu que l'équation de la régression multiple est la suivante :

$Y_i : (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_i$

Pour la régression logistique, c'est la même chose, mais en ajoutant la transformation logarithmique. Par exemple, l'équation pour 1 prédicteur est la suivante :

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1)}}$$

où :
P(Y) est la probabilité que Y arrive
e est la base des logarithmes naturels
Les coefficients b_0 et b_1 représentent la combinaison linéaire du prédicteur et de la constante.

La régression à plusieurs prédicteurs est donc formulée ainsi :

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}}$$

Il faut toutefois se rappeler que même si la formule se ressemble, on ne peut pas appliquer une régression multiple quand la VD est dichotomique, parce qu'on ne respecte pas la prémisse de relation linéaire. La transformation logarithmique permet à l'équation de prendre une forme linéaire.

Le résultat obtenu à une régression logistique se situera toujours entre 0 et 1. Si la valeur est près de 0, la probabilité est faible que l'événement arrive, alors que si la valeur est près de 1, la probabilité est élevée.

La probabilité maximale

La droite de moindres carrés de la régression linéaire est construite à partir des coefficients qui minimisent la distance au carré entre les points (les valeurs observées) et la droite de régression. Le choix des coefficients de la régression logistique reposent plutôt sur l'obtention des valeurs prédites de Y situées le plus près possible des valeurs observées. Ces coefficients constituent les paramètres d'estimation de la probabilité maximale (*maximum-likelihood*) et mesurent le changement du ratio de probabilité (*odds ratio*).

La qualité d'ajustement du modèle : la probabilité log

Comme pour la régression linéaire, l'objectif de la régression logistique est que la variable ajoutée au modèle permette plus efficacement de prédire l'appartenance au groupe que ne le fait le modèle initial (sans prédicteur). La probabilité log (log likelihood), qui s'apparente à la somme des carrés résiduelle (SC_R), permet de comparer la valeur observée et prédite pour une personne et ainsi d'évaluer le degré d'imprécision du modèle. Cette probabilité indique quelle proportion de variance il reste à expliquer après avoir intégré le prédicteur au modèle. Lorsque la valeur de la probabilité log reste élevée, le modèle est peu ajusté aux données, puisqu'il demeure beaucoup de variance à expliquer.

La signification de la diminution de la probabilité log est évaluée dans une distribution χ^2.
La statistique χ^2 remplit donc le même rôle que la valeur F et nous indique si le modèle est significatif. Nous poouvons ainsi répondre à la question :Est-ce que la probabilité log avec seulement la constante est réduite de manière significative lorsque l'on ajoute les prédicteurs ?

Le meilleur prédicteur

Dans la régression logistique, le modèle de base est le plus grand nombre de cas, c'est-à-dire la catégorie (0 ou 1) qui obtient la fréquence la plus élevée. En effet, il ne serait pas judicieux d'utiliser la moyenne comme dans la régression linéaire, puisque la moyenne de 0 et de 1 ne ferait pas de sens. Ainsi, le modèle qui fournit la meilleure prédiction est l'événement qui arrive le plus souvent. Ce dernier est utilisé comme constante.

L'amélioration du modèle est donc calculée à partir de la probabilité -2 log de base (*log likelihood value* : -2LL), qui illustre la différence au carré entre le modèle de base (la constante ou l'événement qui arrive le plus souvent) et le modèle avec un ou plusieurs prédicteur (s).
$\chi^2 = 2[LL\ (modèle) – LL\ (base)]$

La différence est mise au carré, car le degré de signification du résultat est évalué à partir de la distribution χ^2 (ddl = k-1), où k représente le nombre de paramètres dans le modèle.

R et R^2

La statistique R n'est pas fournie par SPSS, mais peut être calculée à la main. Elle représente la corrélation partielle entre chaque VI et la VD et se situe toujours entre – 1 et 1. Lorsque la valence est positive, la valeur du prédicteur augmente en même temps que la probabilité Y, alors que lorsque la valence est négative, la probabilité Y diminue quand la

valeur du prédicteur augmente. Plus petite est la valeur calculée, plus faiblement le prédicteur contribue au modèle.

$$R = +/- \sqrt{\frac{Wald - (2 \times ddl)}{-2LL(original)}}$$

Par contre, la statistique R est calculée à partir de la statistique de Wald, qui sera décrite dans la prochaine section, et de ce fait, doit être interprétée avec prudence. La statistique Wald peut en effet se révéler imprécise dans certaines circonstances, notamment lorsque la taille de l'échantillon est importante. De plus, la statistique R ne peut pas être mise au carrée pour être interprétée comme le R^2 de la régression linéaire. Nous utilisons plutôt le Pseudo R^2 qui est obtenu à partir de la formule suivante :

$$R^2_{logit} = \frac{-2LL_{original} - (-2LL_{modèle})}{-2LL_{original}}$$

Celui-ci représente la proportion de la probabilité expliquée par l'amélioration du modèle. Sa valeur varie entre 0 et 1.

D'autres indicateurs peuvent aussi être utilisés. Ils sont fournis dans les tableaux de résultats. Le R^2_L de Hosmer et Lemeshow indique la réduction de la proportion de la valeur absolue de la probabilité log. En ce sens, il s'agit d'une mesure de l'amélioration de l'ajustement du modèle lorsqu'une variable est retirée. Sa valeur varie entre 0 (lorsque la variable indépendante ne permet pas de prédire Y) et 1 (lorsque la variable indépendante prédit parfaitement Y).

Ensuite, le R^2 de Cox et Snell (1989) et le R^2 Nagelkerke (1991), tous deux fournis dans un tableau SPSS, ces derniers s'apparentent aussi au R^2 de la régression linéaire. Le premier n'atteint jamais le maximum théorique de 1 et varie en fonction de la taille de l'échantillon. Le second est une modification du 1er pour obtenir une valeur théorique plus près de 1. Ils mesurent la force de l'association (la taille d'effet) et fournissent un indice de l'ajustement au modèle. Ils représentent un estimé de la variance expliquée par le modèle. Plus leur valeur est élevée, plus la probabilité prédite par le modèle s'approche de la valeur observée.

Le test de Hosmer et Lemeshow (1989)

Ce test évalue la présence de différences significatives entre les valeurs observées et les valeurs prédites pour chaque sujet. Nous cherchons évidemment à ce qu'il ne soit pas significatif. Par contre, il est très sensible à la taille de l'échantillon. De plus, il ne peut pas être calculé lorsque le modèle ne comprend qu'un prédicteur dichotomique. Il doit donc être utilisé à titre indicatif seulement.

Calcul de l'apport de chaque prédicteur : la statistique de Wald

Une fois que nous savons si le modèle est bien ajusté aux données, il est intéressant de connaître l'apport de chaque prédicteur à l'amélioration du modèle. Pour ce faire, nous avons

recours à la statistique de Wald. Elle occupe la même fonction que le test-t dans la régression. Elle indique si chaque coefficient beta (*b*) contribue significativement à l'amélioration du modèle, donc si sa valeur est différente de 0. Elle est évaluée dans une distribution dans une distribution χ^2. Cette statistique est calculée ainsi :

$$\text{Wald} = \frac{b}{\text{Err-t}_b}$$

Il est à noter que l'erreur-type est plus grande quand le coefficient *b* est élevé, il en résulte que la statistique Wald est sous-estimée et peut se révéler non significative même si elle l'est dans la réalité (erreur de type II).

Le changement de proportion : exp *b*

Occupant une fonction similaire à celle du coefficient *b* dans la régression linéaire, le coefficient exp *b* indique le changement de proportion (*odds ratio*) lorsque le prédicteur X augmente d'une unité. Lorsque la valeur est plus grande que 1, la probabilité augmente avec le changement.

Il faut savoir que la probabilité qu'un événement arrive (*odds*) est définie comme la probabilité qu'il arrive divisé par celle qu'il n'arrive pas :

$$\text{Odds} = \frac{\text{P(événement Y)}}{\text{P(pas événement Y)}}$$

Nous devons donc d'abord calculer $\text{P(événement Y)} = \frac{1}{1 + e^{-(b_0 + b_1 x_1)}}$

Puis
P(pas événement Y) = 1 – P (événement Y)

Si nous utilisons un exemple concret avec un prédicteur dichotomique pour faciliter la compréhension… nous disons que nous voulons évaluer la probabilité d'avoir la grippe P(Y) en ayant un vaccin (X = 1) et celle d'avoir la grippe sans utiliser le vaccin (X = 0).

Pour obtenir le résultat, nous devons donc remplacer la valeur de b_0 et de b_1 dans l'équation. Nous effectuons le calcul en remplacant X par 0 pour obtenir la probabilité originale (odds), soit le rapport entre la probabilité d'avoir la grippe et de ne pas avoir la grippe lorsque nous n'avons pas de vaccin.

Ensuite, nous reprenons le calcul avec le changement d'une unité du prédicteur, donc en remplacant X par 1. Cette probabilité représentera le rapport entre la probabilité d'avoir la grippe et celle de ne pas avoir la grippe lorsque nous recevons un vaccin.

À partir de ce moment, nous pourrons calculer le changement de proportion, soit la valeur de exp *b*.

$$\Delta \text{ odds} = \frac{\text{Odds après une unité de changement dans le prédicteur}}{\text{Odds original}}$$

Les méthodes de régression

Les méthodes de régression disponibles sont les mêmes que pour la régression linéaire. Toutefois, le critère de sélection pour les méthodes progressives est différent.

Vous pouvez donc opter pour la méthode Entrée et insérer toutes les variables prédictrices en même temps. De plus, si vous préférez sélectionner l'ordre d'entrée des variables, choisissez la méthode hiérarchique. Les paramètres seront calculés pour chaque bloc de variables.

Parmi les méthodes progressives, vous avez toujours le choix entre ascendante ou descendante. Dans la méthode ascendante, SPSS introduit la variable ayant le score le plus élevé en premier jusqu'à ce qu'aucune variable n'ait une statistique score significative (soit plus petit que 0,05). Dans la méthode descendante, le contraire se produit puisque le premier modèle évalué contient toutes les variables et SPSS retire celles qui ne contribuent pas significativement à l'amélioration de la prédiction. La différence avec la régression multiple est que SPSS évalue à chaque étape si certaines variables devraient être retirées en se basant sur
Le **rapport de vraisemblance** (*likelihood-ratio,* LR) : SPSS conserve la variable si le changement du LR est significatif quand la variable est retirée, ce qui indique que cette variable contribue à la qualité de l'ajustement
La **statistique conditionnelle** : il s'agit d'un critère moins exigeant que le LR, donc il est préférable de prioriser le 1er
La **statistique Wald** : cette fois, SPSS retire toutes les variables pour lesquelles la statistique Wald est inférieure à 0,1. Cette méthode peut être utilisée avec un petit échantillon. Sinon, il est préférable de privilégier le LR.

Interprétations :

Dans cet exemple, nous chercherons à identifier les variables qui permettent de prédire le plus efficacement la probabilité de vivre un épuisement professionnel chez les enseignants (ÉPUISEMENT). Nous vérifierons donc l'effet du stress généré par les élèves (ÉLÈVES), par les parents (PARENTS) et la direction (DIRECTION), le sentiment d'auto-contrôle (CONTRÔLE), les stratégies d'adaptation (ADAPTATION) et la présence d'un trouble anxieux (ANXIÉTÉ) sur la présence ou non d'un épuisement professionnel. Toutes les variables prédictrices évaluées sont continues, mise à part la variable anxiété qui est catégorielle.

Étape 0 : le modèle de base

Le premier tableau indique simplement que SPSS a conservé les mêmes valeurs que celles utilisées pour coder les variables, soit 0 pour les individus qui ne sont pas épuisés et 1 pour ceux qui le sont.

Codage de variables dépendantes

Valeur d'origine	Valeur interne
0 Non épuisé	0
1 Épuisé	1

Le tableau suivant illustre les valeurs utilisées pour la variable prédictrice catégorielle. Puisque nous avons choisi le contraste indicateur, nous conservons également les mêmes valeurs que pour coder la variable.

Codages des variables nominales

		Fréquence	Codage des paramètres (1)
anxiété	,00	406	,000
	1,00	61	1,000

Le troisième tableau présente l'historique des itérations pour le modèle de base. Nous retenons particulièrement la probabilité log (-2LL) initiale. Elle est de 530,107 et représente la probabilité que nous chercherons à améliorer (réduire) en ajoutant des variables prédictrices.

Historique des itérations[a,b,c]

Itération		-2log-vraisemblance	Coefficients Constant
Etape 0	1	530,875	-,981
	2	530,108	-1,071
	3	530,107	-1,073
	4	530,107	-1,073

a. La constante est incluse dans le modèle.

b. -2log-vraisemblance initiale : 530,107

c. L'estimation a été interrompue au numéro d'itération 4 parce que les estimations de paramètres ont changé de moins de ,001.

Le tableau de classement montre pour sa part que la prédiction en se basant sur la catégorie la plus fréquente permet de classifier correctement 74,5 % des participants.

Tableau de classement[a,b]

Observations			Prévisions: Épuisement professionnel: 0 Non épuisé	Prévisions: Épuisement professionnel: 1 Épuisé	Pourcentage correct
Etape 0	Épuisement professionnel	0 Non épuisé	348	0	100,0
		1 Épuisé	119	0	,0
	Pourcentage global				74,5

a. La constante est incluse dans le modèle.

b. La valeur de césure est ,500

Le tableau des variables dans l'équation nous indique la valeur du coefficient b_0. Dans notre cas, il est de – 1,073.

Variables dans l'équation

		B	E.S.	Wald	ddl	Sig.	Exp(B)
Etape 0	Constante	-1,073	,106	102,111	1	,000	,342

Enfin, le dernier tableau montre les valeurs de la statistique Score pour chaque variable prédictrice hors de l'équation qui s'apparente aux valeurs de corrélation partielle dans la régression multiple. Comme elles sont toutes significatives, elles contribueraient donc probablement toutes à l'amélioration du modèle.

Variables hors de l'équation

			Score	ddl	Sig.
Etape 0	Variables	contrôle	61,405	1	,000
		adapatation	133,012	1	,000
		enseignement	35,722	1	,000
		parents	6,306	1	,012
		direction	23,364	1	,000
		anxiété(1)	15,408	1	,000
	Statistiques globales		186,732	6	,000

Étape 1 : Évaluation de la signification du modèle de régression

Le tableau récapitulatif des modèles fournit les valeurs -2LL pour chaque étape du modèle. Nous pouvons déterminer si la probabilité – 2LL de chaque étape du modèle est inférieure à la probabilité – 2LL de base (530,11) et si cette différence est significative, ce qui nous indiquera si les termes de l'équation logistique finale prédisent mieux la probabilité de vivre un épuisement professionnel que ne le fait la probabilité initiale observée.

Récapitulatif des modèles

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	399,033[a]	,245	,361
2	364,179[a]	,299	,441
3	336,770[a]	,339	,500
4	324,710[b]	,356	,524

a. L'estimation a été interrompue au numéro d'itération 5 parce que les estimations de paramètres ont changé de moins de ,001.

b. L'estimation a été interrompue au numéro d'itération 6 parce que les estimations de paramètres ont changé de moins de ,001.

Par exemple, pour l'étape 1, nous pouvons calculer 530,11 – 399,033, ce qui donne 131,74. Cette valeur est évaluée dans une distribution χ^2 et sa signification est présentée dans le tableau tests de spécification du modèle. Nous constatons que l'étape (soit l'ajout de la variable adaptation) et le modèle complet sont significatifs. Bien sûr, à l'étape 1, le modèle ne comprend qu'une variable, donc nécessairement, la valeur χ^2 est identique pour les deux éléments. La ligne bloc ne sera examinée que dans une régression hiérarchique où on introduirait plus d'une variable (donc un bloc de variables) par étape.

Nous pouvons voir aux étapes suivantes que la ligne «étape» et la ligne «modèle» n'indiquent pas les mêmes valeurs. La ligne étape montre en effet la différence entre la probabilité -2LL de l'étape précédente et celle obtenue par l'ajout du nouveau prédicteur. Nous cherchons à ce qu'à chaque étape, le modèle présente une diminution significative du -2LL.

Tests de spécification du modèle

		Khi-Chi-deux	ddl	Sig.
Etape 1	Etape	131,074	1	,000
	Bloc	131,074	1	,000
	Modèle	131,074	1	,000
Etape 2	Etape	34,853	1	,000
	Bloc	165,928	2	,000
	Modèle	165,928	2	,000
Etape 3	Etape	27,409	1	,000
	Bloc	193,337	3	,000
	Modèle	193,337	3	,000
Etape 4	Etape	12,060	1	,001
	Bloc	205,397	4	,000
	Modèle	205,397	4	,000

À la lumière de ces deux tableaux, nous pouvons dire que le modèle final permet de prédire significativement mieux la probabilité de vivre un épuisement professionnel que le fait le modèle incluant seulement la constante.

Nous pouvons ensuite examiner le test de Hosmer-Lemeshow. Celui-ci indique s'il existe un écart important entre les valeurs prédites et observées. Nous constatons à la lecture du tableau qu'il existe une différence significative entre les valeurs prédites et observées pour les étapes 1 à 3, mais que lorsque la 4^{e} variable est introduite, les valeurs prédites et observées sont cohérentes.

Test de Hosmer-Lemeshow

Etape	Khi-Chi-deux	ddl	Sig.
1	50,018	8	,000
2	34,438	8	,000
3	15,972	8	,043
4	11,489	8	,176

Étape 2 : Évaluation de l'ajustement des données au modèle de régression

Ensuite, il faut évaluer la signification statistique des coefficients estimés des variables indépendantes conservées afin de s'assurer que chacune contribue à mieux prédire P(y) qu'un modèle qui ne l'inclurait pas. Pour ce faire, nous nous basons sur la statistique Wald. Cette dernière illustre la différence dans le modèle avant et après l'ajout de la dernière variable. On observe qu'à l'étape finale, tous les coefficients sont significatifs, même si plusieurs variables ont été introduites. On rejette donc pour chaque variable que le coefficient est égal à 0. Par conséquent, chacune contribue à l'amélioration du modèle.

Variables dans l'équation

		B	E.S.	Wald	ddl	Sig.	Exp(B)	IC pour Exp(B) 95% Inférieur	IC pour Exp(B) 95% Supérieur
Etape 1[a]	élèves	,086	,009	90,613	1	,000	1,090	1,071	1,110
	Constante	-3,384	,283	142,632	1	,000	,034		
Etape 2[b]	contrôle	,061	,011	31,316	1	,000	1,063	1,040	1,086
	élèves	,083	,009	77,950	1	,000	1,086	1,066	1,106
	Constante	-4,484	,379	139,668	1	,000	,011		
Etape 3[c]	contrôle	,092	,014	46,340	1	,000	1,097	1,068	1,126
	adaptation	-,083	,017	23,962	1	,000	,921	,890	,952
	élèves	,131	,015	76,877	1	,000	1,139	1,107	1,173
	Constante	-1,707	,619	7,599	1	,006	,181		
Etape 4[d]	contrôle	,107	,015	52,576	1	,000	1,113	1,081	1,145
	adaptation	-,110	,020	31,660	1	,000	,896	,862	,931
	élèves	,135	,016	75,054	1	,000	1,145	1,110	1,181
	direction	,044	,013	11,517	1	,001	1,045	1,019	1,071
	Constante	-3,023	,747	16,379	1	,000	,049		

a. Variable(s) entrées à l'étape 1 : élèves.

b. Variable(s) entrées à l'étape 2 : contrôle.

c. Variable(s) entrées à l'étape 3 : adaptation.

d. Variable(s) entrées à l'étape 4 : direction.

Le sens des coefficients *b* et de Exp(b) indiquent le sens de la relation. On constate donc que la relation est positive pour les variables contrôle, élèves et direction, soit que le faible sentiment de contrôle et le stress engendré par les élèves et la direction prédisent l'épuisement professionnel. Par contre, la relation est négative pour la variable adaptation, c'est donc dire que meilleures sont les stratégies d'adaptation de l'enseignant face au stress, moins il est probable qu'il vive un épuisement professionnel.

Le tableau suivant permet d'évaluer à chaque étape la présence d'un changement significatif de la probabilité -2LL lorsqu'une variable est retirée du modèle (la valeur doit être significative pour que la variable soit conservée).

Modèle si terme supprimé

Variable		Modèle log-vraisemblance	Modification dans -2log-vraisemblance	ddl	Signification de la modification
Etape 1	élèves	-265,054	131,074	1	,000
Etape 2	contrôle	-199,516	34,853	1	,000
	élèves	-236,237	108,294	1	,000
Etape 3	contrôle	-196,546	56,322	1	,000
	adaptation	-182,090	27,409	1	,000
	élèves	-231,036	125,301	1	,000
Etape 4	contrôle	-195,158	65,606	1	,000
	adaptation	-181,548	38,386	1	,000
	élèves	-224,912	125,113	1	,000
	direction	-168,385	12,060	1	,001

Le tableau des variables hors de l'équation est aussi produit pour chacune des étapes. Comme lors du modèle initial, on peut observer que la ligne statistique globale est significative pour les étapes 1 à 3, donc que l'ajout d'une variable contribuerait à améliorer le modèle. À chaque étape, la variable qui a été incluse par SPSS est celle ayant la variable score la plus élevée dans la mesure où elle était significative.

Variables hors de l'équation

			Score	ddl	Sig.
Etape 1	Variables	contrôle	42,181	1	,000
		adaptation	5,857	1	,016
		parents	,404	1	,525
		direction	,389	1	,533
		anxiété(1)	,333	1	,564
	Statistiques globales		72,701	5	,000
Etape 2	Variables	adaptation	25,911	1	,000
		parents	2,626	1	,105
		direction	1,086	1	,297
		anxiété(1)	,411	1	,521
	Statistiques globales		39,375	4	,000
Etape 3	Variables	parents	3,070	1	,080
		direction	11,957	1	,001
		anxiété(1)	2,358	1	,125
	Statistiques globales		15,180	3	,002
Etape 4	Variables	parents	3,514	1	,061
		anxiété(1)	,331	1	,565
	Statistiques globales		3,564	2	,168

Étape 3 : Évaluation de l'ajustement du modèle final

Nous savons maintenant que le modèle final est significatif et que chacune des variables indépendantes contribue significativement à mieux prédire P(y) qu'un modèle qui ne les inclut pas. Nous nous intéressons maintenant à savoir si le modèle est bien ajusté aux données. Pour ce faire, nous revenons au tableau récapitulatif du modèle pour voir les valeurs des R^2 de Cox et Snell et de Nalgelkerke. Comme le R^2 de la régression multiple, plus la valeur est élevée, mieux le modèle est ajusté aux données. Nous observons que la valeur augmente pour chaque étape et pouvons conclure que le modèle final est le mieux ajusté.

Récapitulatif des modèles

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	399,033[a]	,245	,361
2	364,179[a]	,299	,441
3	336,770[a]	,339	,500
4	324,710[b]	,356	,524

a. L'estimation a été interrompue au numéro d'itération 5 parce que les estimations de paramètres ont changé de moins de ,001.

b. L'estimation a été interrompue au numéro d'itération 6 parce que les estimations de paramètres ont changé de moins de ,001.

Il est également possible de calculer la valeur du Pseudo-R^2 pour obtenir un estimé de la variabilité expliqué.

$$R^2_{logit} = \frac{-2LL_{original} - (-2LL_{modèle})}{-2LL_{original}}$$

$$0,39 = \frac{530,107-324,710}{530,107}$$

Le modèle final prédit donc 39 % de la variance de la probabilité de vivre un épuisement professionnel.

Étape 4 : Évaluation de la justesse de l'ajustement du modèle

Il est maintenant possible d'examiner si le modèle permet de bien classer les sujets dans leur groupe d'appartenance à partir de l'équation finale. Nous nous rappelons que le hasard permettait de classer correctement 74,5 % des participants. Nous voyons que le pourcentage correct de classification passe de 78,8 % avec une seule variable indépendante et monte à 83,1 % pour l'étape 3. Il redescend minimalement à 82,9 % pour l'étape 4 où 92,2 % des enseignants non épuisés sont classés correctement, mais que seulement 55,5 % des épuisés le sont. Selon les résultats précédents, cette amélioration est significative.

Tableau de classement[a]

			Prévisions		
			Épuisement professionnel		
	Observations		,00	1,00	Pourcentage correct
Etape 1	Épuisement professionnel	,00	321	27	92,2
		1,00	72	47	39,5
	Pourcentage global				78,8
Etape 2	Épuisement professionnel	,00	318	30	91,4
		1,00	61	58	48,7
	Pourcentage global				80,5
Etape 3	Épuisement professionnel	,00	322	26	92,5
		1,00	53	66	55,5
	Pourcentage global				83,1
Etape 4	Épuisement professionnel	,00	321	27	92,2
		1,00	53	66	55,5
	Pourcentage global				82,9

a. La valeur de césure est ,500

Les résiduels

Finalement, on peut regarder le dernier tableau produit par SPSS, soit la liste des observations ayant une valeur résiduelle standardisée plus élevée que 2. Toujours dan l'optique de s'assurer que le modèle est bien ajusté aux données et qu'il prédit efficacement le groupe d'appartenance, nous conservons toujours les paramètres de la distribution normale, soit un maximum de

5 % des observations à l'extérieur des limites de ±1,96

1 % des observations à l'extérieur des limites de ±2,58

et nous portons une attention particulière à celle situées à plus 3 écart-types.

Nous constatons que 7 observations sur le total de 467 participants ont des valeurs résiduelles de plus de 3 écarts-types, ce qui représente 1 % de l'échantillon. Il pourrait donc être intéressant de réaliser à nouveau l'analyse sans ces participants afin de vérifier si les coefficients estimés par le modèle varient beaucoup. Si c'était le cas, ces individus pourraient être considérés comme influençant l'ajustement du modèle.

Liste par observation[b]

Observation	Etat sélectionné[a]	Observations	Prévisions	Groupe prédit	Variable temporaire	
		épuisement			Resid	ZResid
77	S	N**	,888	É	-,888	-2,817
178	S	N**	,945	É	-,945	-4,126
285	S	N**	,941	É	-,941	-3,996
353	S	É**	,094	N	,906	3,110
354	S	É**	,091	N	,909	3,152
355	S	É**	,021	N	,979	6,789
357	S	É**	,128	N	,872	2,615
363	S	É**	,108	N	,892	2,876
364	S	É**	,081	N	,919	3,362
368	S	É**	,135	N	,865	2,532
387	S	É**	,072	N	,928	3,579
388	S	É**	,102	N	,898	2,969
391	S	É**	,115	N	,885	2,770

a. S = observations sélectionnées, U = observations non sélectionnées et ** = observations mal classées.

b. Les observations avec des résidus studentisés supérieurs à 2,000 sont répertoriées.