# 1 Objectif

Régression logistique polytomique à variable dépendante ordinale.

La régression logistique est une technique très populaire pour analyser les dépendances entre une variable à expliquer (dépendante, endogène) binaire et une ou plusieurs variables explicatives (indépendantes, exogènes) quantitatives et qualitatives ordinales ou nominales[1].

L'approche peut être facilement généralisée à l'explication des valeurs prises par une variable dépendante qualitative nominale à K (K > 2) modalités. Il faut pour cela prendre une modalité de référence, la dernière par exemple, et estimer (K-1) LOGITS (Équation 1). La prédiction et l'interprétation des coefficients de la régression ne sont guères modifiées. On parle de régression logistique polytomique à variable dépendante nominale ; on parle aussi de régression logistique multinomiale en référence à la distribution utilisée pour modéliser la probabilité d'appartenance à un groupe[2].

$$\ln\frac{P(Y=k/X)}{P(Y=K/X)} = a_{0,k} + a_{1,k}X_1 + \cdots + a_{J,k}X_J, k = 1,\ldots,K-1$$

**Équation 1 - LOGITS pour le modèle multinomial, la dernière modalité est la référence**

La situation est un peu plus complexe lorsqu'il s'agit de modéliser une liaison impliquant une variable dépendante ordinale. Pléthores d'interprétations sont possibles, allant de l'impasse sur le caractère ordinal afin de revenir simplement au modèle multinomial, à l'assimilation de la variable à prédire à une variable quantitative, dans ce cas la régression linéaire multiple devrait suffire. Entre ces deux cas extrêmes existent différentes approches. Dans ce didacticiel, nous étudierons essentiellement les LOGITS adjacents et les LOGITS cumulatifs[3]. On parle alors de régression logistique polytomique à variable dépendante ordinale.

Pour étudier ces techniques, **nous utiliserons le logiciel R**, accessible librement en ligne[4]. Il s'agit d'un logiciel de statistique disposant d'un interpréteur de commande et d'un vrai langage de programmation. Il est particulièrement performant grâce au système des packages, des modules externes compilés, qui permettent de compléter sa bibliothèque de fonctions statistiques. Dans notre étude, nous utiliserons en priorité le package VGAM[5], il élargit de manière significative les dispositions de R en matière de régression généralisée.

Nous considérons que le lecteur est familiarisé avec R, notamment en ce qui concerne l'appréhension des fichiers et des vecteurs de données. Nous irons donc directement à l'essentiel en mettant l'accent, d'une part, sur les commandes relatives à la régression, d'autre part, sur la lecture et l'interprétation, en relation avec la théorie statistique, des sorties du logiciel.

[1] http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique.pdf

[2] http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique_polytomique.pdf

[3] http://www.stat.psu.edu/~jglenn/stat504/08_multilog/01_multilog_intro.htm

[4] http://www.r-project.org/

[5] http://www.stat.auckland.ac.nz/~yee/VGAM/ ; plus particulièrement, pour la description des méthodes décrites dans ce didacticiel : http://www.stat.auckland.ac.nz/~yee/VGAM/doc/Categorical.pdf

# 2 Données

Les données (hypertension.txt[6]) utilisées dans ce didacticiel proviennent d'une étude de cas publiée sur le web[7]. L'objectif est d'expliquer l'hypertension artérielle des patients à partir de leurs caractéristiques physiologiques, cliniques et comportementales : le sexe, fumer ou pas, effectuer régulièrement des exercices physiques, etc.

### 2.1.1 Variables

La variable dépendante est au départ la pression systolique (SYSTOLIC). Nous avons décidé de la découper en 4 niveaux en nous appuyant sur les valeurs limites couramment utilisées dans le domaine[8] :

- 1 : tension normale si PA systolique < 140 mm hg
- 2 : hypertension élevée si PA systolique > 140 mm g et <= 160 mm hg
- 3 : hypertension très élevée si PA systolique > 160 mm hg et <= 180 mm hg
- 4 : hypertension sévère si PA systolique > 180 mm hg

Pour expliquer l'hypertension, nous avons retenu 9 variables indépendantes :

| Variables indépendantes | Description |
|---|---|
| Homme | Sexe (1 : masculin ; 0 : féminin) |
| Fumeur | Fumeur (1 : oui ; 0 : non) |
| Exercice | Niveau d'activité physique (1 : faible ; 2 : moyen ; 3 : élevé) |
| Surpoids | Corpulence (1 : normal ; 2 : surcharge pondérale ; 3 : obèse) |
| Alcool | Consommation d'alcool (1 : faible ; 2 moyenne ; 3 : élevée) |
| Stress | Niveau de stress (1 : faible ; 2 : moyen ; 3 : élevé) |
| Sel | Consommation de sel (1 : faible ; 2 : moyen ; 3 : élevée) |
| Revenu | Niveau de revenu (1 : faible ; 2 : moyen ; 3 : élevé) |
| Education | Niveau d'éducation (1 : faible ; 2 : moyen ; 3 : élevé) |

### 2.1.2 Observations

Le fichier originel comprenait des individus ayant été traité contre l'hypertension, une variable indicatrice permettait de les distinguer. De fait, cette dernière masquait toutes les informations intéressantes. Les individus non malades se différenciaient avant tout parce qu'ils ont été traités. Nous avons donc décidé de ne conserver que les individus qui n'ont pas été soignés. Le fichier ainsi formé comporte **399 observations**.

[6] http://eric.univ-lyon2.fr/~ricco/cours/didacticiels/data-mining/hypertension.txt

[7] http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm

[8] http://fr.wikipedia.org/wiki/Hypertension_artérielle

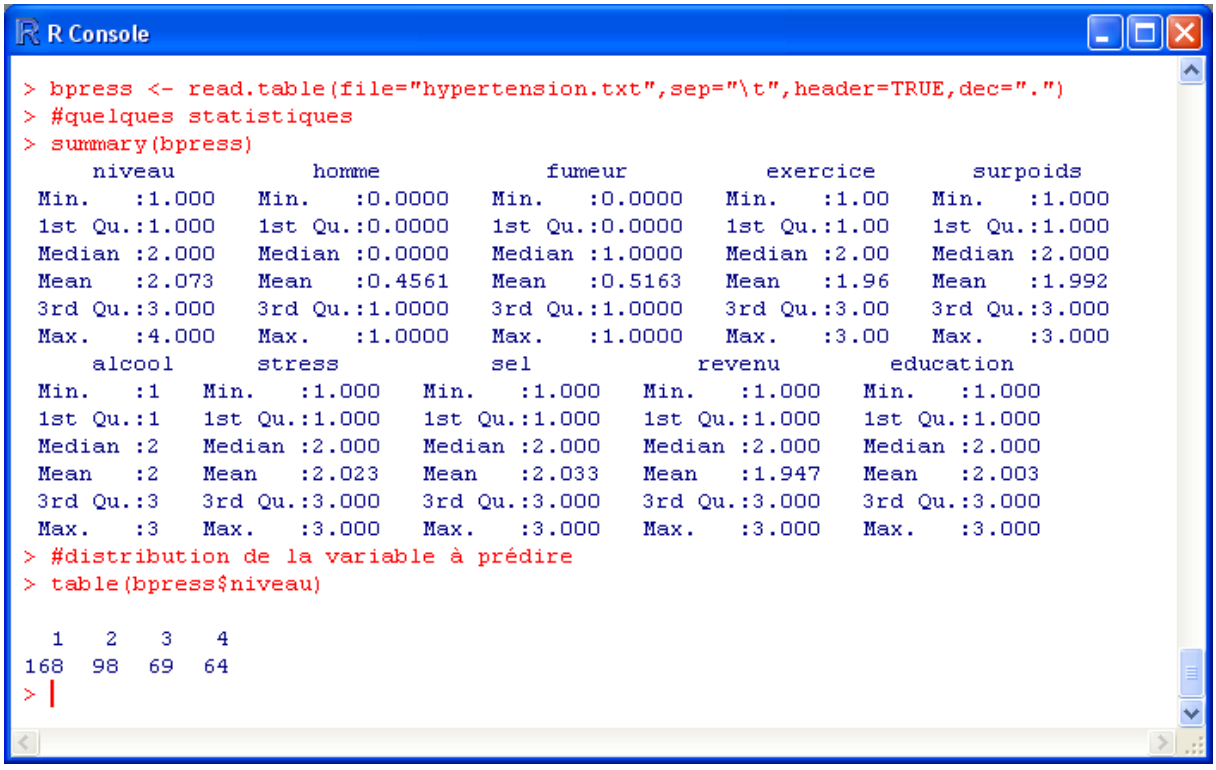
# 3 Quelques opérations préparatoires

## 3.1 Chargement des données et statistiques descriptives

Dans un premier temps, nous importons le contenu du fichier « hypertension.txt » dans une structure *data.frame*. Tout de suite, nous vérifions que tout s'est bien passé en demandant une description succincte des variables.

```
#chargement des données
bpress <- read.table(file="hypertension.txt",sep="\t",header=TRUE,dec=".")
#quelques statistiques
summary(bpress)
#distribution de la variable à prédire
table(bpress$niveau)
```

Quelques statistiques descriptives permettent de vérifier l'intégrité des données.




Toutes les colonnes étant décrites par des chiffres dans notre fichier, R considère que les variables sont toutes numériques. Une seconde commande *table()* est donc nécessaire pour obtenir la distribution de la variable à expliquer.

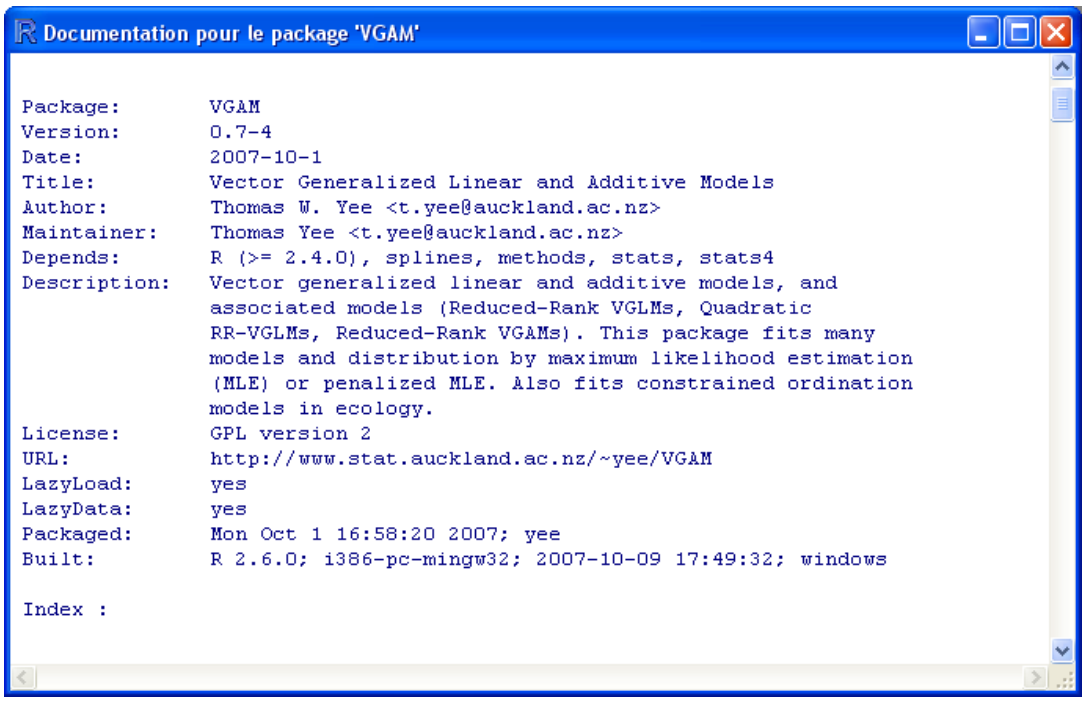
| Niveau | Effectif |
|---|---|
| 1 | 168 |
| 2 | 98 |
| 3 | 69 |
| 4 | 64 |
| Total | 399 |

## 3.2 Chargement de la bibliothèque VGAM

Dans un premier temps, la bibliothèque VGMA doit être installée correctement sur notre machine. Se référer à la documentation de R pour cette procédure. Puis nous la chargeons afin de pouvoir en exploiter les fonctions spécialisées. Il est possible d'obtenir une description courte de la bibliothèque et la liste de ses (innombrables) fonctions.

```
#chargement de la librairie VGAM
library(VGAM)
#description et liste des procédures
library(help=VGAM)
```

Une documentation approfondie est disponible sur le site de l'auteur du package.



# 4 Régression logistique polytomique – LOGITS adjacents

## 4.1 Principe des LOGITS adjacents

L'idée des LOGITS adjacents est de modéliser l'odds du passage d'une catégorie à l'autre avec une combinaison linéaire des variables explicatives (Équation 2). Dans notre cas, il s'agit de modéliser l'aggravation de l'hypertension c.-à-d. le passage du niveau k à (k+1).

$$\ln \frac{P(Y=k+1/X)}{P(Y=k/X)} = a_{0,k} + a_{1,k}X_1 + \cdots + a_{J,k}X_J, k = 1,\ldots,K-1$$

**Équation 2 - LOGITS adjacents**

La procédure vglm() du package VGAM permet d'en estimer les paramètres sur notre fichier de données.

Il y a (K – 1) = 3 équations à produire, avec ( J + 1) = 9 + 1 = 10 coefficients à estimer dans chaque équation. Il y a donc (K-1) x (J+1) = 3 x 10 = 30 paramètres à estimer en tout.

```
#modèle avec les variables
modele <- vglm(niveau ~ ., data = bpress, family = acat())
summary(modele)
```

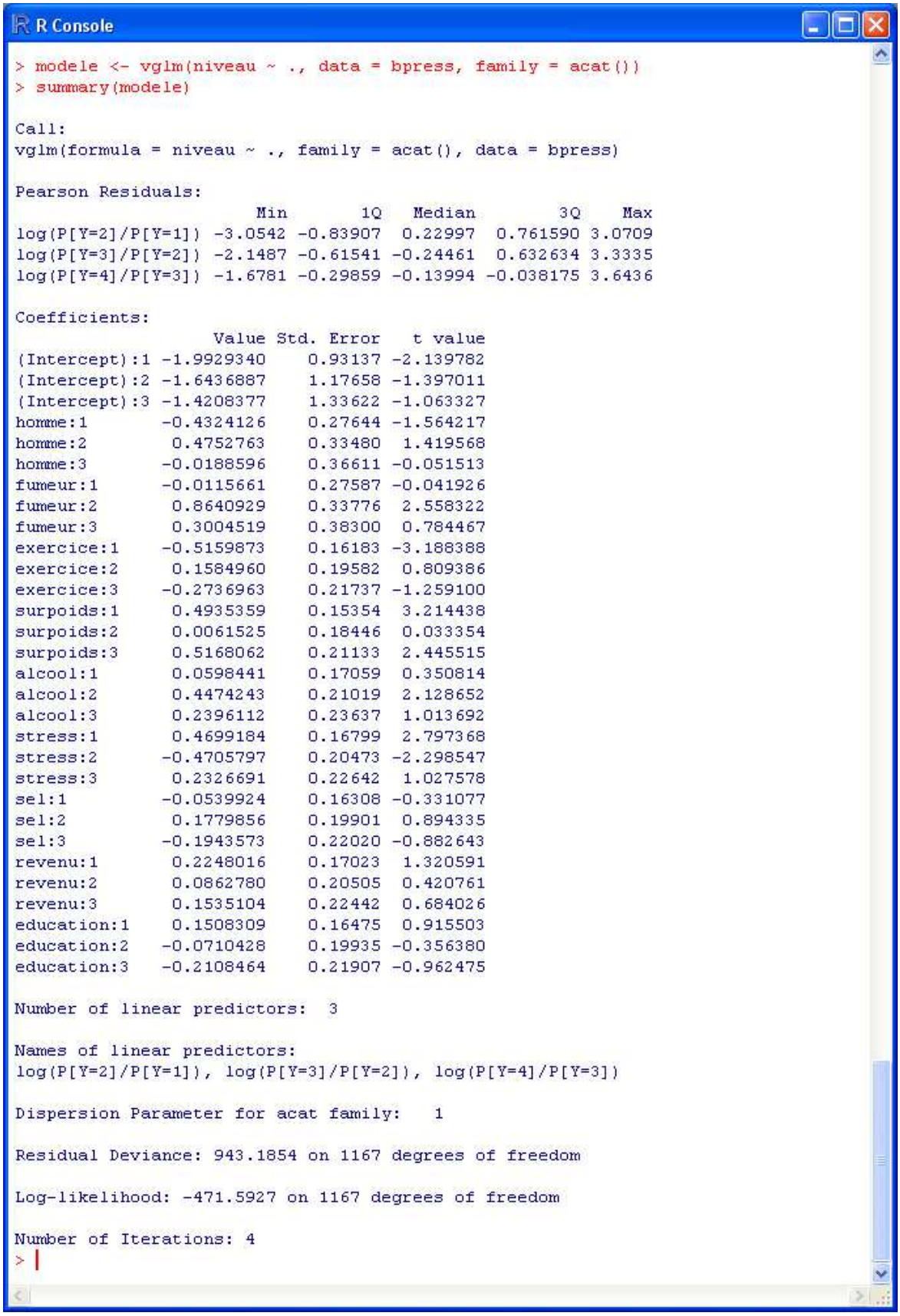
Dans la commande vglm(), l'option family permet de spécifier le type de modélisation que l'on veut mettre en œuvre. Dans notre exemple, nous choisissons la valeur acat() pour spécifier que l'on veut produire une estimation selon les LOGITS adjacents.

Les sorties du logiciel sont assez touffues (Figure 1). Essayons d'y discerner les informations importantes.

- Des statistiques sur les résidus de chaque équation sont proposées (**Pearson Residuals**). Nous constatons que la régression oppose bien le niveau (k + 1) au niveau précédent (k). Nous observons également que la qualité de l'estimation est à peu près équivalente, quelle que soit k, les résidus ne se démarquent pas beaucoup d'une équation à l'autre.

- Pour chaque LOGIT n°k, nous avons la liste des paramètres estimés (**Coefficients**). Nous ne nous attarderons pas trop sur ces coefficients dans un premier temps, tout au plus remarquerons nous que (1) l'écart type estimé (Std. Error), (2) le rapport entre le coefficient et son écart type (t value) sont directement calculés. Pour tester la significativité à 5% d'un coefficient dans une équation, nous comparerons[9] la « t value » avec la valeur 2. Si elle est plus grande, en valeur absolue, on peut conclure que le coefficient est significativement différent de 0.

[9] Si l'on veut vraiment pinailler, on le comparerait plutôt à 1.96 puisque le carré de la « t value » permet d'implémenter le test de Wald. Cette statistique « carré du t value » suit asymptotiquement une loi du KHI-2 à 1 degré de liberté. La valeur critique serait alors 3.84 → $1.96^2 = 3.84$. Mais il s'agit bien là d'un test asymptotique, le choix d'une valeur seuil 2 pour la valeur absolue de la « t value » ne dénature pas le test.



```
R Console
> modele <- vglm(niveau ~ ., data = bpress, family = acat())
> summary(modele)

Call:
vglm(formula = niveau ~ ., family = acat(), data = bpress)

Pearson Residuals:
                        Min       1Q   Median       3Q    Max
log(P[Y=2]/P[Y=1]) -3.0542 -0.83907  0.22997  0.761590 3.0709
log(P[Y=3]/P[Y=2]) -2.1487 -0.61541 -0.24461  0.632634 3.3335
log(P[Y=4]/P[Y=3]) -1.6781 -0.29859 -0.13994 -0.038175 3.6436

Coefficients:
                   Value Std. Error   t value
(Intercept):1 -1.9929340    0.93137 -2.139782
(Intercept):2 -1.6436887    1.17658 -1.397011
(Intercept):3 -1.4208377    1.33622 -1.063327
homme:1       -0.4324126    0.27644 -1.564217
homme:2        0.4752763    0.33480  1.419568
homme:3       -0.0188596    0.36611 -0.051513
fumeur:1      -0.0115661    0.27587 -0.041926
fumeur:2       0.8640929    0.33776  2.558322
fumeur:3       0.3004519    0.38300  0.784467
exercice:1    -0.5159873    0.16183 -3.188388
exercice:2     0.1584960    0.19582  0.809386
exercice:3    -0.2736963    0.21737 -1.259100
surpoids:1     0.4935359    0.15354  3.214438
surpoids:2     0.0061525    0.18446  0.033354
surpoids:3     0.5168062    0.21133  2.445515
alcool:1       0.0598441    0.17059  0.350814
alcool:2       0.4474243    0.21019  2.128652
alcool:3       0.2396112    0.23637  1.013692
stress:1       0.4699184    0.16799  2.797368
stress:2      -0.4705797    0.20473 -2.298547
stress:3       0.2326691    0.22642  1.027578
sel:1         -0.0539924    0.16308 -0.331077
sel:2          0.1779856    0.19901  0.894335
sel:3         -0.1943573    0.22020 -0.882643
revenu:1       0.2248016    0.17023  1.320591
revenu:2       0.0862780    0.20505  0.420761
revenu:3       0.1535104    0.22442  0.684026
education:1    0.1508309    0.16475  0.915503
education:2   -0.0710428    0.19935 -0.356380
education:3   -0.2108464    0.21907 -0.962475

Number of linear predictors:  3

Names of linear predictors:
log(P[Y=2]/P[Y=1]), log(P[Y=3]/P[Y=2]), log(P[Y=4]/P[Y=3])

Dispersion Parameter for acat family:   1

Residual Deviance: 943.1854 on 1167 degrees of freedom

Log-likelihood: -471.5927 on 1167 degrees of freedom

Number of Iterations: 4
>
```

**Figure 1 – LOGITS adjacents - Variables + Constantes**

- La dernière information importante est la déviance (**Deviance**), elle est égale à 943.1854. Plus elle est petite, meilleure sera l'approximation. Il ne faut pas tenir compte des degrés de libertés affichés ici. R considère qu'il y a une observation par « covariate pattern ». L'évaluation du modèle basé sur le résidu déviance n'est pas utilisable[10]. Nous utiliserons cette information surtout pour comparer des modèles. Notons qu'étrangement, le package utilise la formule (K-1) x N – (K-1) * (J+1) = 1167 pour obtenir les degrés de libertés.

## 4.2 Evaluation globale – Comparaison avec le modèle initial

Une question importante est la significativité de la régression : est-ce que les variables intégrées dans le modèle expliquent, globalement, les valeurs de la variable dépendante. Le plus simple dans ce cas est de produire la régression composée de la seule constante et de comparer les déviances des régressions avec et sans les variables : c'est le test du rapport de vraisemblance.

Pour produire le modèle initial, composé de la seule constante, nous écrivons sous R :

```
#modèle initial
modele.initial <- vglm(niveau ~ 1, data = bpress, family = acat())
summary(modele.initial)
```

Nous obtenons (Figure 2) :

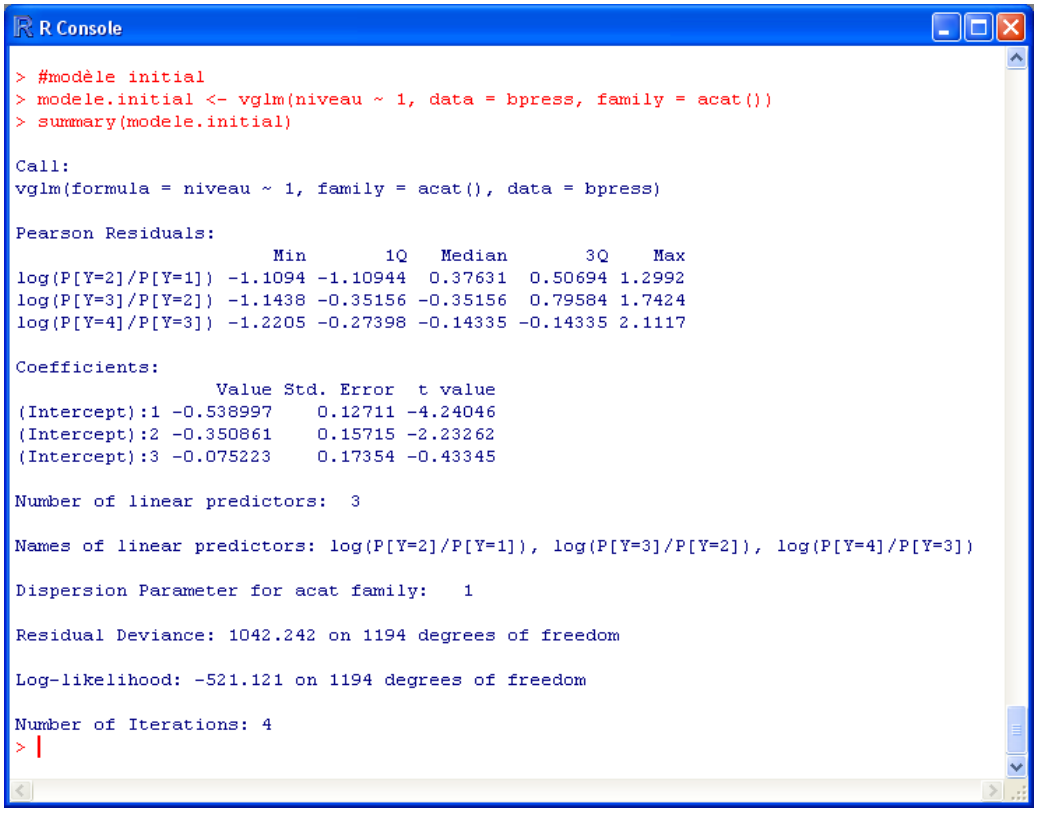



**Figure 2 - LOGITS adjacents - Constantes seulement**

[10] http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique.pdf

La déviance du modèle initial est : 1042.242.

La statistique du test du rapport de vraisemblance est égale à 1042.242 – 943.1854 = 99.0566. La statistique suit une loi du KHI-2, ses degrés de liberté sont (K-1) x J = 3 x 9 = 27 ; la probabilité critique (p-value) du test est <0.00001. Le modèle est globalement significatif au seuil de 5% (et bien en deçà).

## 4.3 Première interprétation des résultats

Maintenant commence le véritable travail : il faut ausculter en détail le modèle pour mettre en évidence les éventuelles causalités entre les variables indépendantes et la variable dépendante. Force est de constater que ça devient rapidement ingérable avec l'augmentation des niveaux de la variable Y et du nombre de variables explicatives. Une équation indique la propension au passage au niveau au dessus P(Y = k+1 / X) / P(Y = k / X) (Équation 2). Nous disposons de (K – 1) équations, il faudrait les lire simultanément pour comprendre la réalité du rôle de la variable, sa significativité, et le sens de la causalité, surcroît de chances d'augmenter le niveau d'hypertension par rapport au niveau actuel ou inversement.

Prenons le cas de la variable indicatrice FUMEUR pour illustrer cela. Nous recensons les informations produites par R (Figure 1) dans le tableau suivant :

| Niveau k de Y – Equation LOGIT n°k | Coefficient | « T value » (* significatif à 5%) |
|---|---|---|
| 1 | -0.116 | -0.042 |
| 2 | 0.864 | 2.558* |
| 3 | 0.300 | 0.784 |

Dans le cas de l'équation n°2, lorsqu'on est fumeur, toutes choses égales par ailleurs, l'augmentation du LOGIT est 0.864 c.-à-d.

$$\ln\frac{P(Y=3/\ fumeur=1,\ldots)}{P(Y=2/\ fumeur=1,\ldots)}-\ln\frac{P(Y=3/\ fumeur=0,\ldots)}{P(Y=2/\ fumeur=0,\ldots)}=0.864$$

Nous en déduisons

$$OR_{fumeur}(3/2)=e^{0.864}=2.37$$

Les fumeurs ont 2.37 fois plus de chances que les non-fumeurs d'être dans le niveau 3 plutôt que dans le niveau 2 c.-à-d. le niveau plus élevé.

Cette première lecture n'est pas évidente pour le non-initié (Voir Annexes 6.1 pour plus de détails). La situation se complique si l'on considère que la même variable fumeur n'est pas significative pour l'équation n°1 (passage du niveau 1 vers le niveau 2) et l'équation n°3 (passage niveau 3 vers niveau 4). Que faut-il penser de la variable FUMEUR finalement ?

Sauf si c'est le but explicite de l'étude, évaluer les influences niveau par niveau est très délicat, voire périlleux, il nous faut introduire une contrainte supplémentaire pour rendre les résultats de la régression exploitables.

## 4.4 Contraintes sur les pentes des LOGITS

Une simplification du modèle initial serait d'écrire les LOGITS adjacents de la manière suivante :

$$\ln \frac{P(Y=k+1/X)}{P(Y=k/X)} = a_{0,k} + a_1 X_1 + \cdots + a_J X_J, k = 1,\ldots,K-1$$

**Équation 3 - LOGITS adjacents et parallèles**

Nous supposons que les coefficients des co-variables sont les mêmes quel que soit le niveau étudié c.-à-d. chaque variable explicative agit de la même manière sur la variable expliquée quel que soit le niveau k.

Géométriquement, dans l'espace de représentation, les LOGITS forment des droites parallèles. Les valeurs des constantes, en modifiant l'origine pour chaque niveau k, permettent de les décaler.

La commande dans R est la suivante, nous introduisons la contrainte « parallèle » :

```
#modèle avec les variables + contrainte parallèle
modele <- vglm(niveau ~ ., data = bpress, family = acat(parallel=TRUE))
summary(modele)
```

Les résultats fournissent plusieurs éléments d'appréciation (Figure 3) :

- Le logiciel a bien calculé 3 équations log [P(Y=k+1/X) / P(Y=k/X)] avec pour contrainte des coefficients de pente identiques dans chaque équation.
- La déviance est égale à 970.1336. La statistique du test du rapport de vraisemblance, en comparaison avec le modèle trivial (Figure 2), est égal à LR = 1042.242 – 970.1336 = 72.1084. Elle suit une loi du KHI-2 à (9 + 3) – 3 = 9 degrés de liberté. La probabilité critique du test est < 0.00001. Le modèle est globalement significatif au seuil de 5%.
- Nous n'avons plus qu'un seul coefficient par variable, on peut mieux situer leur rôle. Reprenons l'exemple de FUMEUR. Il agit positivement (+0.4040) sur le surcroît de chances d'augmenter le niveau d'hypertension, l'odds ratio associé est EXP(0.4040) # 1.5 c.-à-d. un fumeur a 1.5 fois plus de chances d'être au niveau directement au dessus d'hypertension qu'un non fumeur, et ceci quel que soit son niveau actuel d'hypertension.
- Comme nous n'avons plus qu'un seul coefficient, évaluer sa significativité est immédiat à la lecture de la « t value ». Elle est largement au dessus du seuil 2 (pour un test approximativement à 5%), FUMER est néfaste pour l'hypertension.

Par rapport à la situation précédente, la lecture est certes améliorée. Mais il reste quand même une insatisfaction. Nous n'évaluons finalement que le risque de passage au cran supérieur de l'hypertension. La vision est locale. Pour en revenir aux fumeurs, nous avons quantifié qu'ils ont plus de chances d'être au stade suivant de l'hypertension. Mais il serait peut être plus parlant de situer le surcroît de risques sous un angle global c.-à-d. est-ce qu'un fumeur a globalement plus de chances d'être en dessous ou au dessus de tel ou tel niveau ? Les odds cumulatifs permettent de répondre à cette attente.

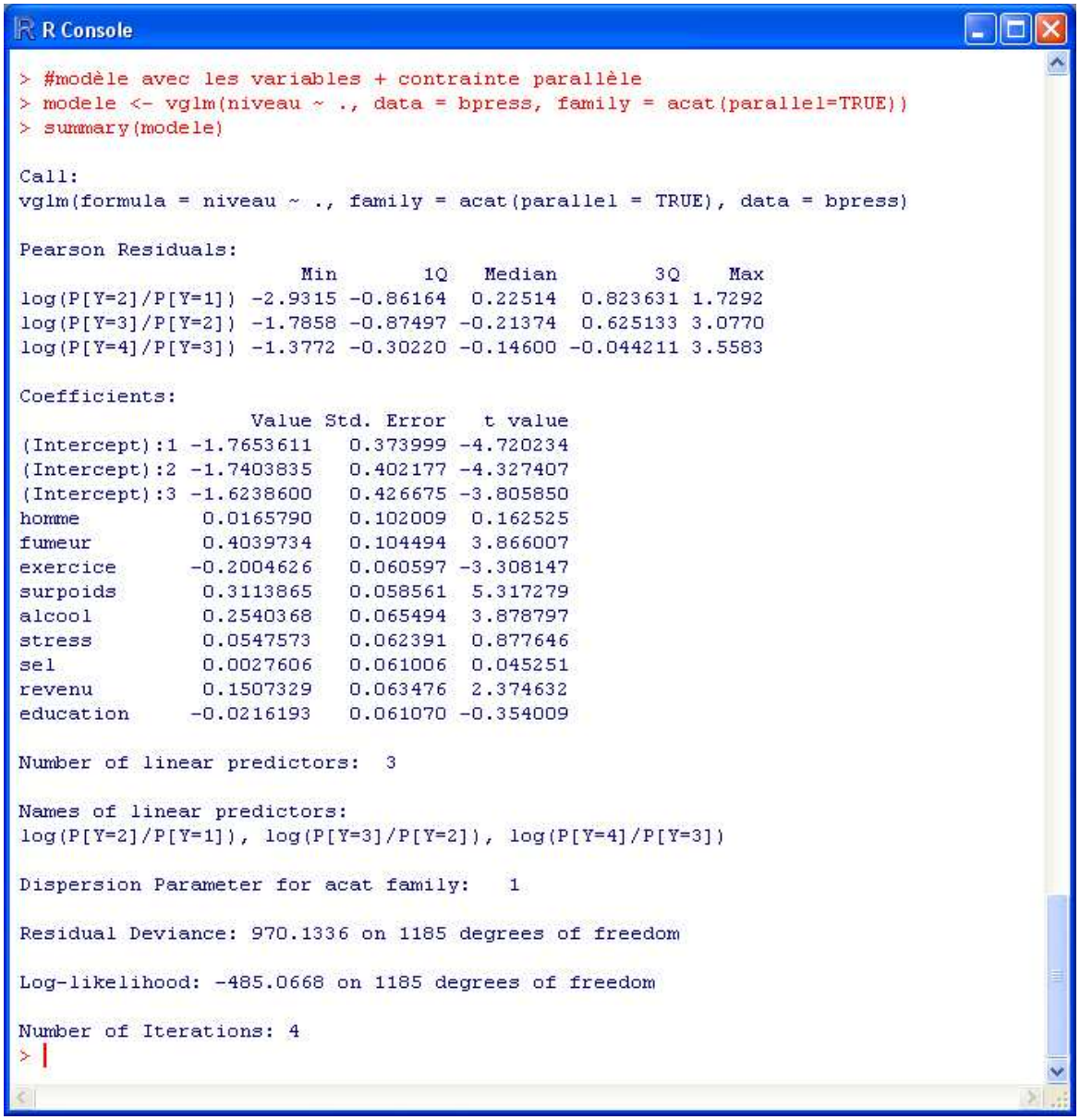



**Figure 3 - LOGITS adjacents et parallèles**

# 5 Régression logistique polytomique – LOGITS cumulatifs

## 5.1 Les odds proportionnels

Par rapport aux LOGITS adjacents, les LOGITS cumulatifs sont plus populaires auprès des praticiens. Ils permettent de **comparer une catégorie avec toutes les catégories qui lui sont inférieures** (ou supérieures selon l'écriture adoptée). Formellement, nous avons toujours (K-1) LOGITS (Équation 4).

$$\ln \frac{P(Y \leq k / X)}{P(Y > k / X)} = a_{0,k} + a_{1,k} X_1 + \cdots + a_{J,k} X_J, k = 1, \ldots, K-1$$

**Équation 4 - LOGITS cumulatifs**

La combinaison linéaire des variables indépendantes quantifie le surcroît de probabilité d'être en dessous, plutôt qu'au-dessus, du niveau k (Voir en annexes un exemple détaillé avec la variable FUMEUR).

De nouveau, nous pouvons introduire l'hypothèse de pentes identiques des LOGITS c.-à-d. l'effet d'une variable est le même quel que soit le niveau de la variable Y. Les coefficients des variables sont les mêmes dans tous LOGITS, seule la constante est propre au niveau. Nous obtenons un modèle très largement répandu : les LOGITS cumulatifs avec odds proportionnels[11] (Équation 5). C'est le modèle que nous étudierons dans cette section.

$$\ln \frac{P(Y \leq k / X)}{P(Y > k / X)} = a_{0,k} + a_1 X_1 + \cdots + a_J X_J, k = 1, \ldots, K-1$$

**Équation 5 - LOGITS cumulatifs, odds proportionnels**

## 5.2 Estimation des paramètres et évaluation globale de la régression

Pour évaluer globalement la régression, rien ne vaut un test du rapport de vraisemblance confrontant le modèle initial, composé des seules constantes, et le modèle intégrant l'ensemble des variables.

Par commodité, afin que la lecture des résultats soit cohérente avec la modélisation selon les LOGITS adjacents où l'on quantifiait le passage au niveau supérieur d'hypertension, nous modéliserons

$$\ln \frac{P(Y > k / X)}{P(Y \leq k / X)} = a_{0,k} + a_1 X_1 + \cdots + a_J X_J, k = 1, \ldots, K-1$$

Voici les commandes R :

```
#modèle avec la constante contrainte parallèle
modele <- vglm(niveau ~ 1, data = bpress, family = cumulative(parallel=TRUE,reverse=TRUE))
summary(modele)
#modèle avec les variables + contrainte parallèle
modele <- vglm(niveau ~ ., data = bpress, family = cumulative(parallel=TRUE,reverse=TRUE))
summary(modele)
```

Nous obtenons tout à tour le modèle composé de la constante (Figure 4) et le modèle intégrant l'ensemble des variables (Figure 5). Nous pouvons ainsi composer la statistique du rapport de vraisemblance LR = 1042.242 – 969.4615 = 72.7805. Elle suit une loi du KHI-2 à (J + K – 1) – (K – 1) = 9 – 4 – 1 – (4 – 1) = 9 degrés de liberté. La probabilité critique associée au test (< 0.00001) indique que les variables jouent un rôle significatif dans l'explication des valeurs prises par Y.

[11] Voir http://www.stat.psu.edu/~jglenn/stat504/08_multilog/40_multilog_proportion.htm ; voir aussi, J.P. Nakache et J. Confais, « Statistique Explicative Appliquée », Editions TECHNIP, 2003 ; pages 145 à 150.

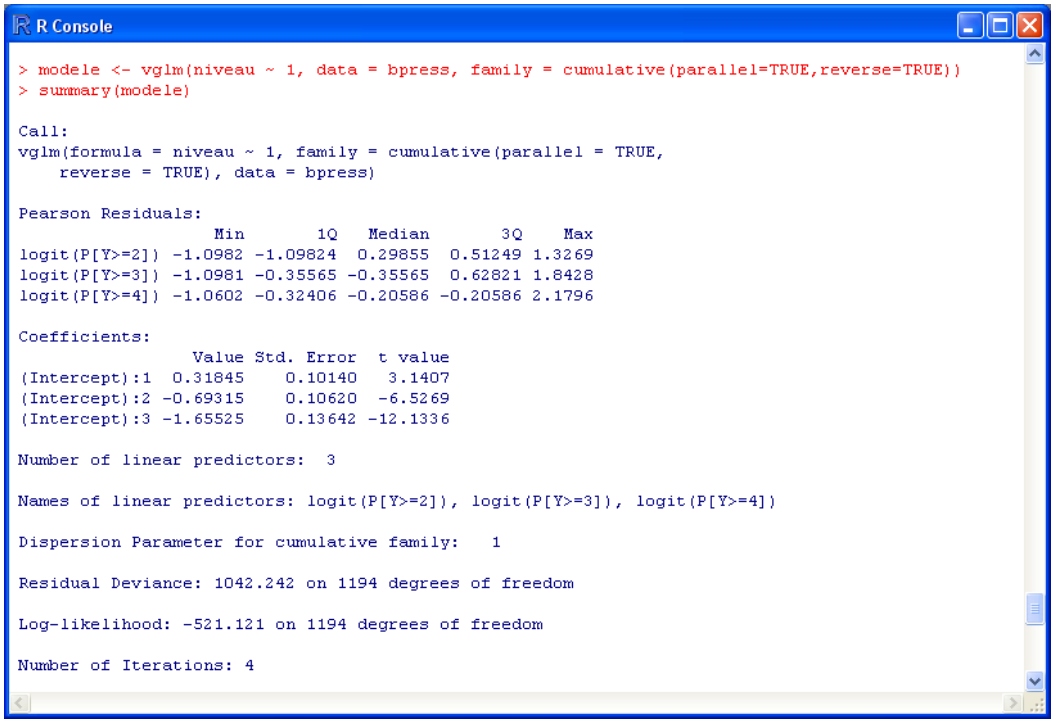



**Figure 4 - LOGITS cumulatifs - Constantes**

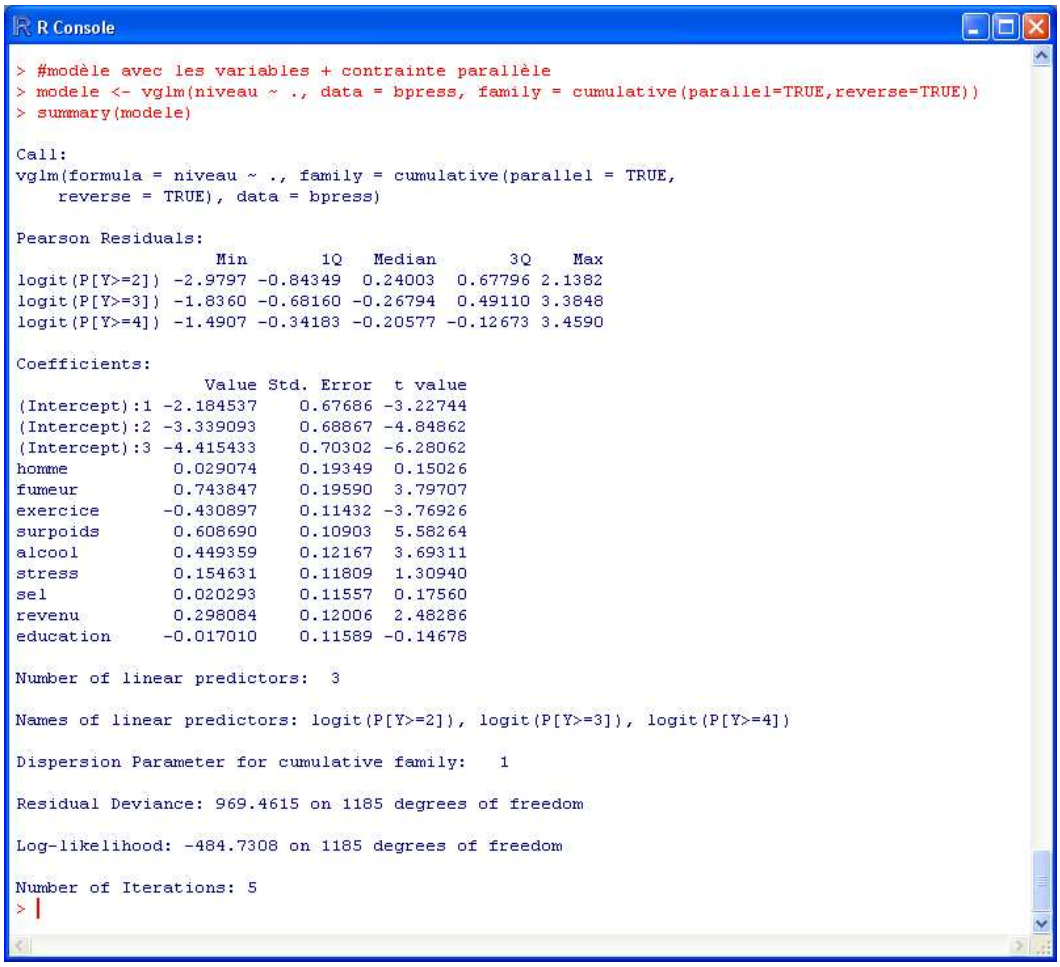



**Figure 5 - LOGITS cumulatifs - Variables + Constantes**

## 5.3 Sélection de variables

L'étape suivante serait l'interprétation des coefficients. Nous remarquons cependant que certaines variables sont significatives, d'autres non. Nous devons procéder à une phase de sélection de variables avant d'interpréter les coefficients.

Il y a plusieurs stratégies de sélection de variables. Nous nous en tiendrons à une procédure BACKWARD très simple : à partir de la régression intégrant toutes les variables, nous supprimons au fur et à mesure la variable la moins pertinente, non significative au sens de la valeur absolue de la « t value » (en la comparant à une valeur seuil prédéfinie, par ex. 2 pour un test à 5%), jusqu'à ce que toutes les variables non significatives aient été écartées.

Voici le code R associé.

```
#*****************************************
#sélection automatique de variables - backward
#*****************************************
backward_vglm <- function(x,seuil=2,...){
  #nombre de modalités de y
  K <- length(unique(x[,1]))
  #indicateur si recherche doit continuer
  okSearch <- TRUE
  #boucle de recherche
  while (okSearch == TRUE && length(x) > 1){
    #former la formule de régression
    str_formule <- paste(names(x[1]),"~")
    for (j in 2:length(x)){
     str_formule <- paste(str_formule,names(x[j]),"+")
    }
    str_formule <- substr(str_formule,1,nchar(str_formule)-2)
    print(str_formule)
    #transformer en formule
    formule <- as.formula(str_formule)
    #lancer la régression
    modele <- vglm(formule, data = x, ...)
    summary.modele <- summary(modele)
    #récupérer coefs, écarts-type et t-value
    mat.coef <- attr(summary.modele,"coef3")
    print(mat.coef)
    t.value <- mat.coef[,3]
    #récupérer le min. en valeur absolue et son indice
    t.min <- (min(abs(t.value[K:length(t.value)])))
    t.index <- which.min(abs(t.value[K:length(t.value)]))
    if (t.min < seuil){
      #non-significatif, on supprime la variable
      index <- t.index + 1
```

```
      x <- x[,-index]
    } else
    {
      okSearch <- FALSE
    }
  }
  #récupérer la dernière formule
  formule <- as.formula(str_formule)
  return (vglm(formule, data = x, ...))
}
```

Après l'appel de la procédure, la régression intégre 5 variables explicatives (Figure 6).

```
#modèle avec sélection de variables
modele <- backward_vglm(bpress,family = cumulative(parallel=TRUE,reverse=TRUE))
summary(modele)
```

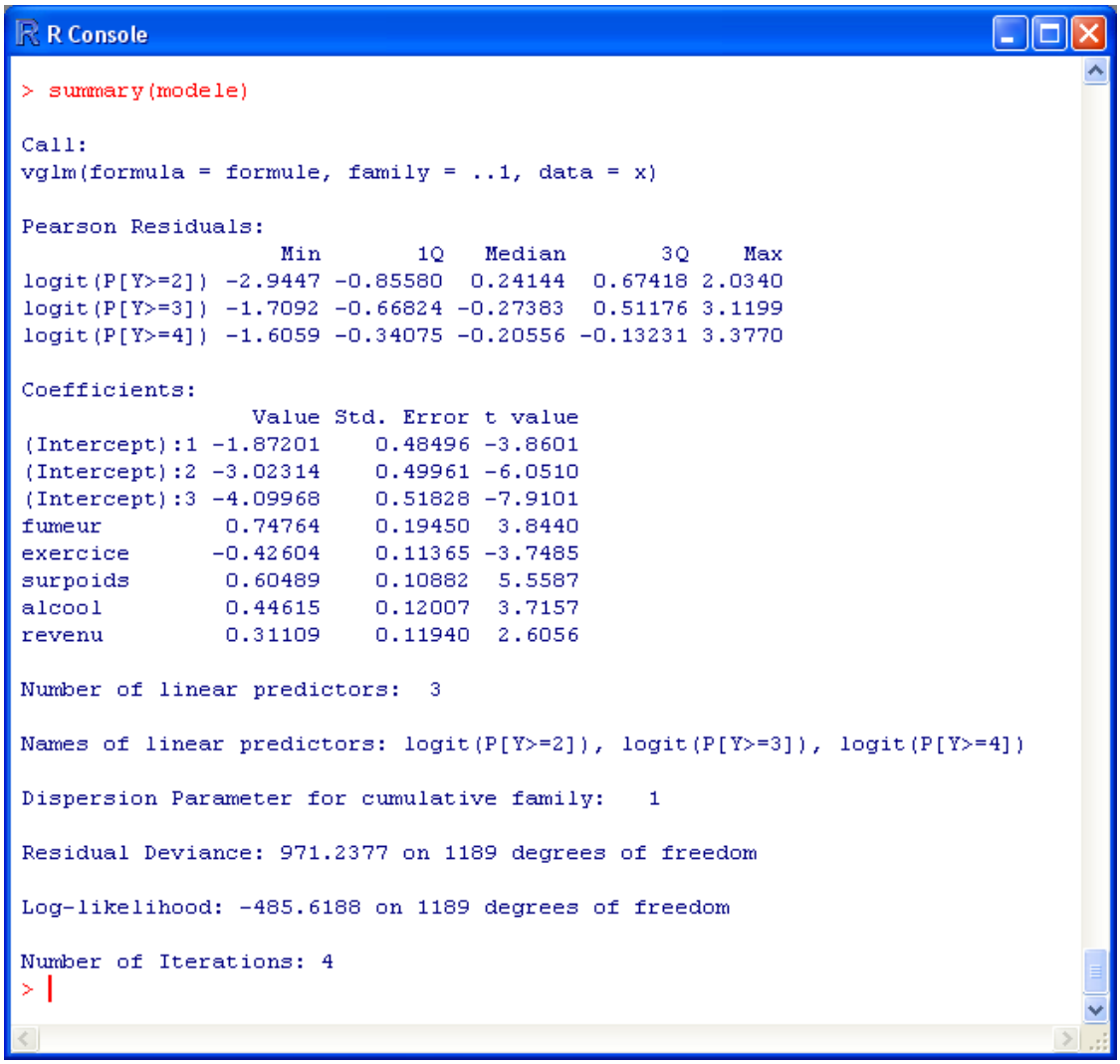



**Figure 6 - LOGITS cumulatifs - Sélection de variables**

Au final, seules les variables « fumeur », « surpoids », « alcool », « revenu », de manière nocive, et « exercice », de manière bénéfique, agissent sur le risque d'hypertension. Mis à part le variable « revenu », ces résultats correspondent peu ou prou à l'idée qu'un néophyte se fait de l'hypertension : trop manger, trop fumer, trop boire, ça aide pas.

## 5.4 Tester le caractère proportionnel des odds

Même si elle est très commode, l'hypothèse selon laquelle les coefficients des variables sont les mêmes pour tous les LOGITS peut ne pas être appropriée, faussant les calculs. Il importe d'en vérifier la compatibilité avec les données avant de se lancer dans des interprétations hasardeuses des coefficients.

L'auteur de la bibliothèque VGAM relève ce problème[12], sans que le test dit « Score Test for Proportional Odds Assumption », n'ait été programmé (semble-t-il). SAS, parmi d'autres logiciels commerciaux, le propose automatiquement dans ses sorties (Annexes 6.3). Dans notre cas, l'hypothèse de proportionnalité est sujette à caution à 5%, elle reste cependant compatible avec les données à 1%. Il faudrait réellement s'inquiéter pour des probabilités critiques très faibles, indiquant un rejet fort de l'hypothèse nulle. Il faudrait dans ce cas, pour se faire une idée de la nature du problème, comparer le modèle contraint, égalité des coefficients pour tous les LOGITS, avec le modèle non contraint, les coefficients sont estimés dans chaque équation LOGIT.

## 5.5 Prédiction

Un des aspects le plus séduisant du LOGIT cumulatif est que nous pouvons obtenir, relativement simplement, les probabilités des valeurs de la variable Y pour un individu donné.

Prenons un individu avec les caractéristiques suivantes (Fumeur = 1, Exercice = 1, Surpoids = 3, Alcool = 3, Revenu = 3). En appliquant les coefficients des LOGITS impliquant ces variables (Figure 6), nous obtenons la valeur C = 0.74764 x 1 – 0.42604 x 1 + 0.60489 x 3 + 0.44615 x 3 + 0.31109 x 3 = 4.40799. Il nous faut maintenant corriger cela avec les constantes et calculer les probabilités :

| | Niveau 4 | Niveau 3 | Niveau 2 |
|---|---|---|---|
| Constante | -4.09968 | -3.02314 | -1.87201 |
| S = C + Constante | 0.30831 | 1.38485 | 2.53598 |
| EXP(S) | 1.36112 | 3.99423 | 12.62880 |
| P(Y >= niveau) = exp(s) / [1 + exp(s)] | 0.57647 | 0.79977 | 0.92663 |
| P(Y = niveau) | 0.57647 | 0.22330 | 0.12686 |

Et par déduction, P(Y = 1) = 1 – 0.92663 = 0.07337.

Nous apprenons que ce monsieur a 57.647% de chances de présenter une hypertension de niveau 4 ; il a 79.977% de chances de présenter un niveau d'hypertension supérieur ou égal à 3 ; en tous les cas, il a 92.6% de chances d'être hypertendu (niveau >= 2). Vu ses caractéristiques, c'était couru d'avance.

Bien entendu, la probabilité qu'il ait un niveau d'hypertension >= 1 est 100%.

[12] http://www.stat.auckland.ac.nz/~yee/VGAM/doc/Categorical.pdf, page 5.

# 6 Annexes

## 6.1 Odds ratio pour une variable explicative binaire – LOGITS adjacents

Pour mieux comprendre les odds ratios, considérons que seule la variable FUMEUR intervient dans la régression. L'intérêt est que nous pouvons étudier simplement les probabilités et les rapports de probabilités à partir d'un tableau de contingence.

R produit le résultat suivant.

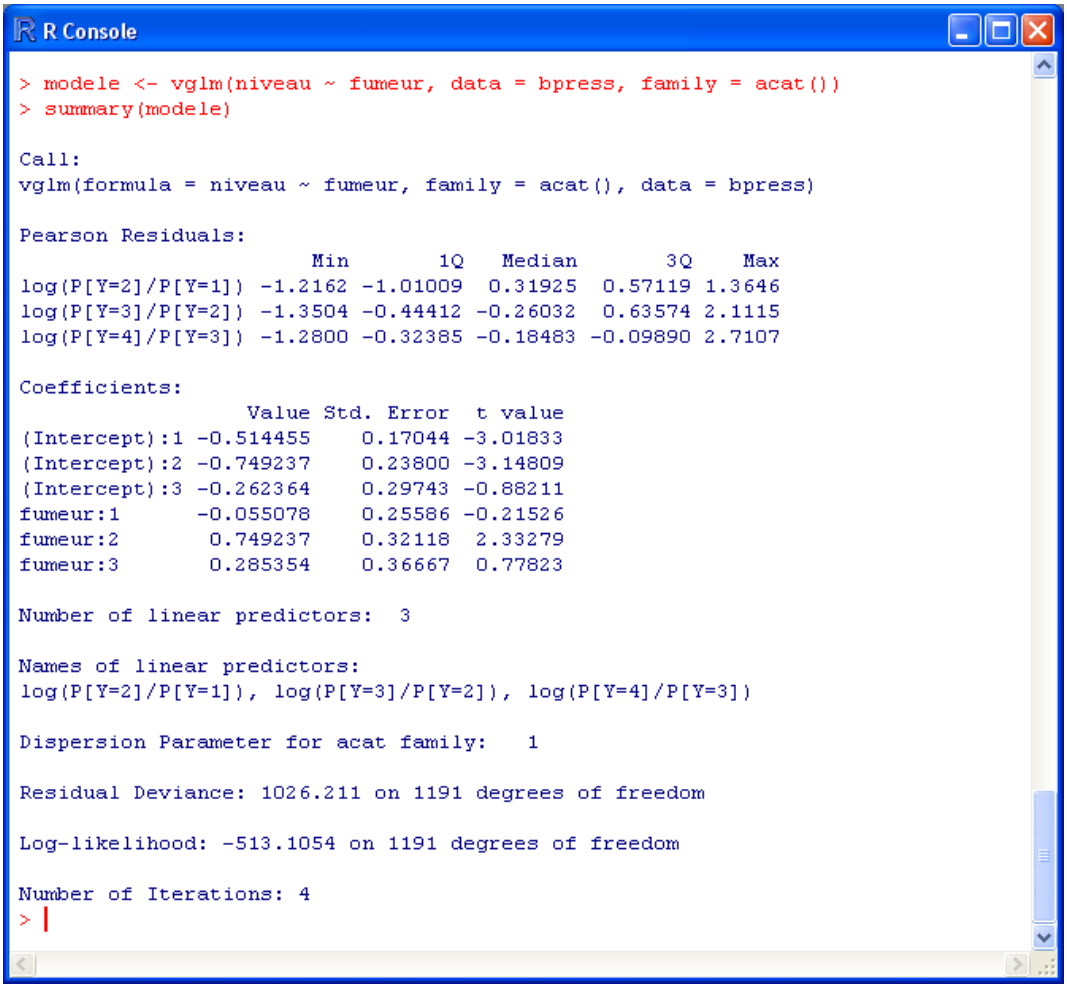



**Figure 7 - Adjacent LOGIT - Fumeur + Constante**

Sans l'intervention des autres variables, nous retrouvons, à peu près (valeur et significativité), les estimations produites par la régression sur l'ensemble des variables. Ceci laisse à penser qu'il y a peu de colinéarités entre les variables introduites dans cette étude.

Calculons maintenant le tableau croisé entre NIVEAU et FUMEUR (Tableau 1).

| Nombre de niveau | fumeur | | |
|---|---|---|---|
| niveau | 1 | 0 | Total |
| 4 | 44 | 20 | 64 |
| 3 | 43 | 26 | 69 |
| 2 | 43 | 55 | 98 |
| 1 | 76 | 92 | 168 |
| Total | 206 | 193 | 399 |

**Tableau 1 - Croisement Niveau Hypertension et FUMEUR**

Concentrons nous sur l'opposition niveau 3 / niveau 2. Que voyons nous dans ce tableau ?

- P(Y=3 / 0) = 26/193 = 0.1347 ; P(Y = 2 / 0) = 55/193 = 0.2849
- Odds[Y= 3/ Y = 2 ; X=0] = 0.1347 / 0.2849 = 0.4727. Lorsqu'on est non-fumeur, on a 0.4727 fois plus de chances d'être dans le niveau 3 que le niveau 2 (ou inversement 1/0.4727 = 2.1154 fois plus de chances d'être dans le niveau 2 que dans le niveau 3).
- Voyons ce qu'il en est pour les fumeurs : P(Y=3 / 1) = 43/206 = 0.21 ; P(Y = 2 / 1) = 43/206 = 0.21 ; et Odds[Y= 3/ Y = 2 ; X=0] = 1.0. Lorsqu'on est fumeur, on a autant de chances d'être dans le niveau 3 que dans le niveau 2.
- L'odds ratio est le rapport entre ces 2 odds c.-à-d. OR(Y=3/Y=2) = 1.0/0.4227 = 2.1154. On peut le lire comme le surcroît de chances des fumeurs, par rapport aux non fumeurs, d'être dans le niveau 3 plutôt que le niveau 2. On pourrait dire aussi que les fumeurs ont 2.1154 fois plus de chances que les non-fumeurs d'être dans le niveau 3 plutôt que dans le niveau 2....

L'odds ratio présent deux avantages décisifs : (1) la valeur calculée ne dépend pas de la répartition des modalités de la variable dépendante Y ; (2) la régression logistique fournit directement les odds ratios.

Revenons sur les résultats de la régression (Figure 7). Le coefficient de fumeur pour la 2ème équation est égal à 0.749237. Si nous formons l'exponentielle EXP(0.749237) = 2.1154 c.-à-d. l'odds ratio que nous avons produit à l'aide du tableau de contingence ci-dessus.

C'est vraisemblablement là un des facteurs de l'immense popularité de la régression logistique auprès des praticiens qui veulent analyser finement les causalités entre la variable expliquée et les variables explicatives, individuellement, globalement, et éventuellement en tenant compte des interactions,

## 6.2 Odds ratio - Variable indépendante binaire – LOGITS cumulatifs

Essayons de retrouver les résultats de la régression à partir de notre tableau de contingence. Pour faciliter la lecture, nous n'étudierons que le premier niveau c.-à-d. le surcroît de risque d'être affecté d'hypertension. De plus, pour être en accord avec les résultats du modèle LOGIT adjacents, nous allons inverser le rapport du LOGIT en estimant les paramètres de l'équation

$$\ln \frac{P(Y>1/X)}{P(Y\leq 1/X)} = a_{0,1} + a_{1,1}X_1 + \cdots + a_{J,1}X_J$$

Voici les résultats produits par *vglm* de R.

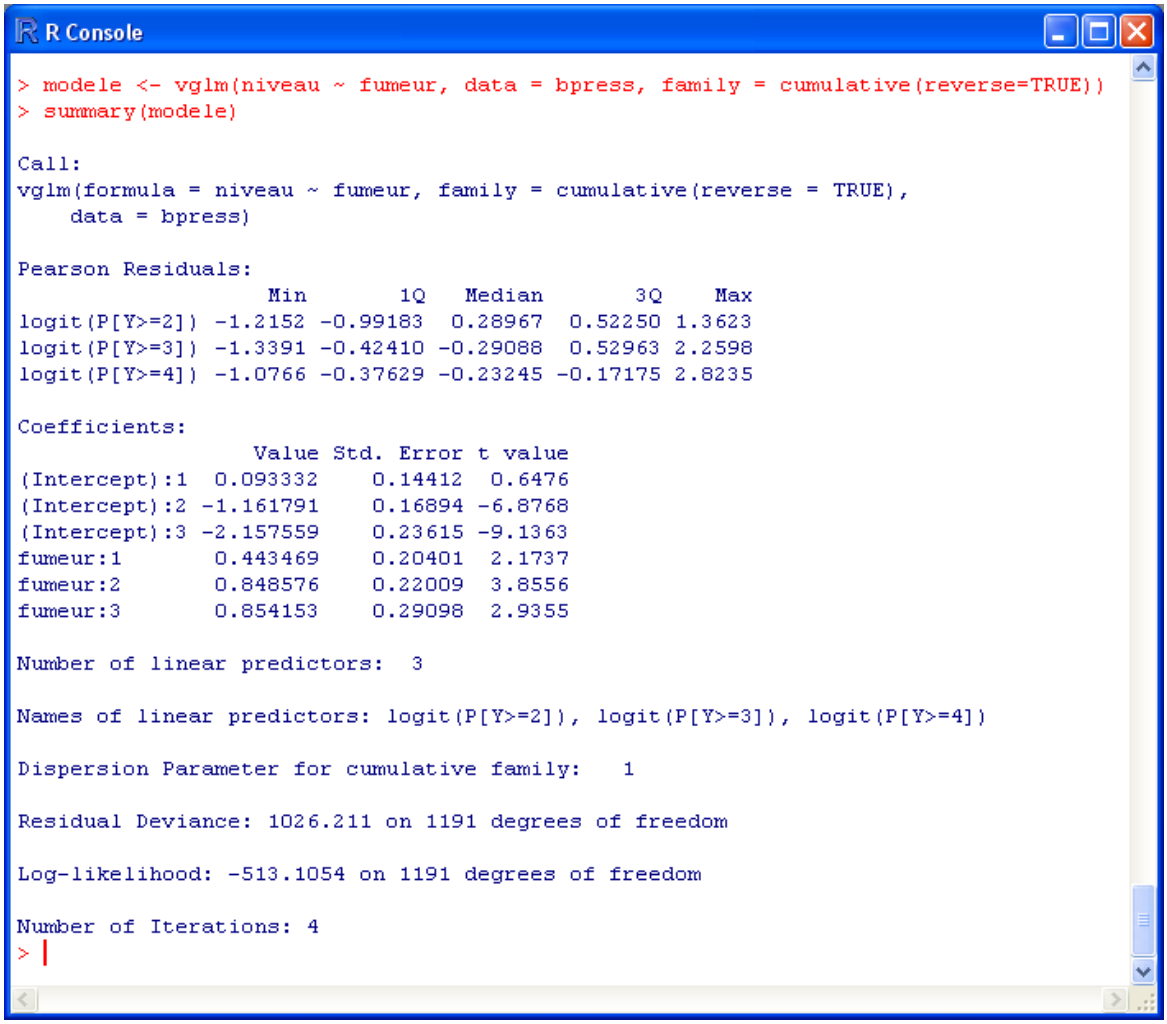



**Figure 8 - LOGITS cumulatifs - Fumeur + constante**

Pour la situation qui nous concerne, R présente l'équation sous la forme

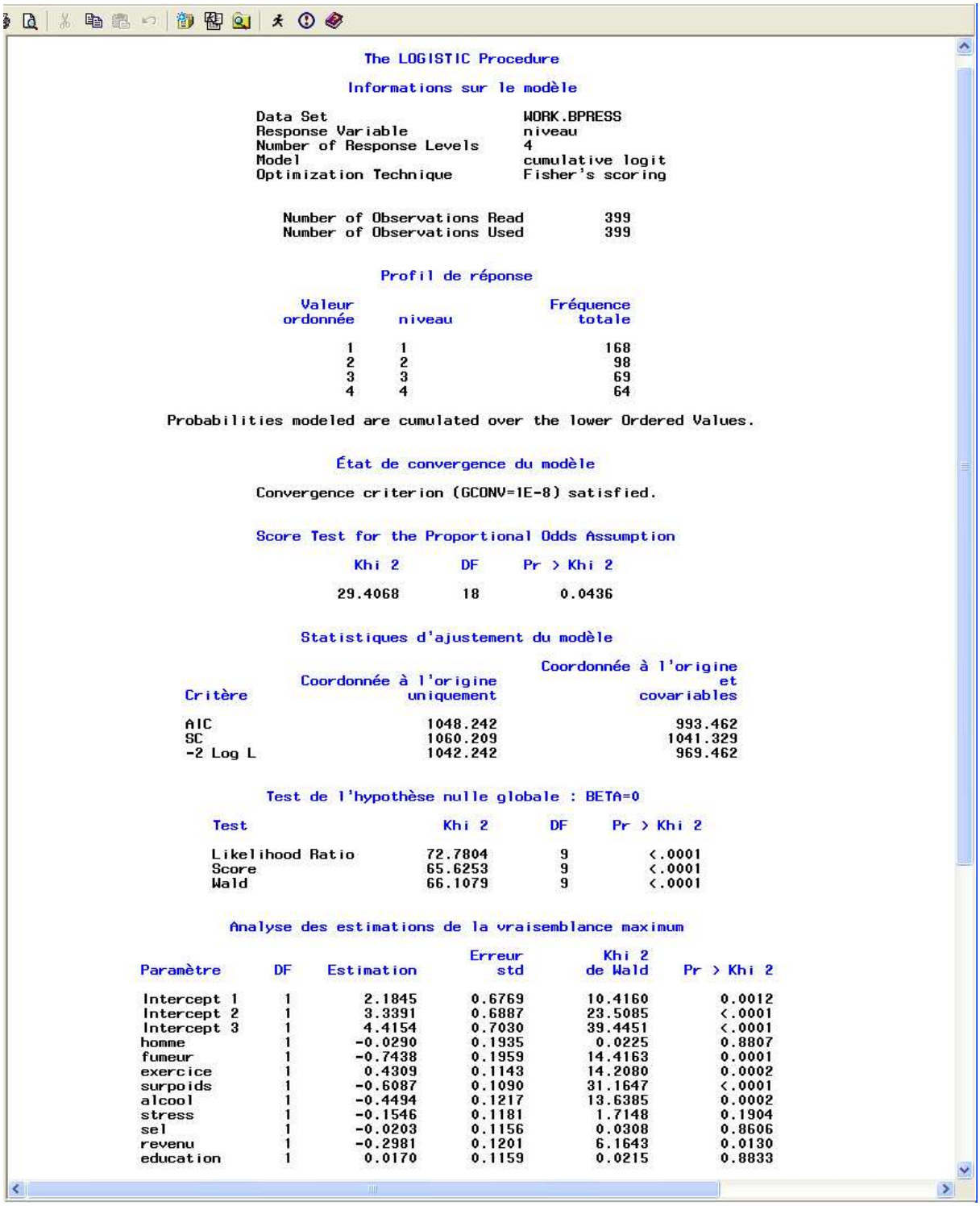
$$\ln \frac{P(Y \geq 2 / X)}{P(Y < 1 / X)} = a_{0,1} + a_{1,1}X_1 + \cdots + a_{J,1}X_J$$

Y prenant exclusivement des valeurs entières, il s'agit bien de la même chose, nous en conviendrons. Reprenons les valeurs de notre tableau de contingence (Tableau 1) :

- Chez les non fumeurs : P(Y>1/0) = (20+26+55)/193=0.5233 ; P(Y<=1/0) = 92/193 = 0.4767 ; Odds [Y >1 / Y<=1 ; 0] = 1.0978
- Chez les fumeurs : P(Y>1/1) = (44+43+43)/206=0.6311 ; P(Y<=1/1) = 76/206 = 0.3689 ; Odds [Y >1 / Y<=1 ; 1] = 1.7105
- L'odds ratio, le surcroît de chances chez les fumeurs, par rapport aux non-fumeurs, de présenter une hypertension est donc OR[Y>1/Y<=1] = 1.7105/10.978 = 1.5581
- C'est ce que nous indique le coefficient FUMEUR:1 de la régression → EXP(0.443469) = 1.5581 (Figure 8).

## 6.3 Résultats de SAS – LOGITS cumulatifs proportionnels

Les résultats produits par SAS (Figure 9) sont totalement en accord avec les résultats produits par le package VGAM de R (Figure 5). C'est plutôt rassurant. SAS se démarque en proposant un test pour évaluer l'hypothèse de proportionnalité des odds (« Score Test for the Proportional Odds Assumption).



The LOGISTIC Procedure

Informations sur le modèle

| | |
|---|---|
| Data Set | WORK.BPRESS |
| Response Variable | niveau |
| Number of Response Levels | 4 |
| Model | cumulative logit |
| Optimization Technique | Fisher's scoring |

| | |
|---|---|
| Number of Observations Read | 399 |
| Number of Observations Used | 399 |

Profil de réponse

| Valeur ordonnée | niveau | Fréquence totale |
|---|---|---|
| 1 | 1 | 168 |
| 2 | 2 | 98 |
| 3 | 3 | 69 |
| 4 | 4 | 64 |

Probabilities modeled are cumulated over the lower Ordered Values.

État de convergence du modèle

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

| Khi 2 | DF | Pr > Khi 2 |
|---|---|---|
| 29.4068 | 18 | 0.0436 |

Statistiques d'ajustement du modèle

| Critère | Coordonnée à l'origine uniquement | Coordonnée à l'origine et covariables |
|---|---|---|
| AIC | 1048.242 | 993.462 |
| SC | 1060.209 | 1041.329 |
| -2 Log L | 1042.242 | 969.462 |

Test de l'hypothèse nulle globale : BETA=0

| Test | Khi 2 | DF | Pr > Khi 2 |
|---|---|---|---|
| Likelihood Ratio | 72.7804 | 9 | <.0001 |
| Score | 65.6253 | 9 | <.0001 |
| Wald | 66.1079 | 9 | <.0001 |

Analyse des estimations de la vraisemblance maximum

| Paramètre | DF | Estimation | Erreur std | Khi 2 de Wald | Pr > Khi 2 |
|---|---|---|---|---|---|
| Intercept 1 | 1 | 2.1845 | 0.6769 | 10.4160 | 0.0012 |
| Intercept 2 | 1 | 3.3391 | 0.6887 | 23.5085 | <.0001 |
| Intercept 3 | 1 | 4.4154 | 0.7030 | 39.4451 | <.0001 |
| homme | 1 | -0.0290 | 0.1935 | 0.0225 | 0.8807 |
| fumeur | 1 | -0.7438 | 0.1959 | 14.4163 | 0.0001 |
| exercice | 1 | 0.4309 | 0.1143 | 14.2080 | 0.0002 |
| surpoids | 1 | -0.6087 | 0.1090 | 31.1647 | <.0001 |
| alcool | 1 | -0.4494 | 0.1217 | 13.6385 | 0.0002 |
| stress | 1 | -0.1546 | 0.1181 | 1.7148 | 0.1904 |
| sel | 1 | -0.0203 | 0.1156 | 0.0308 | 0.8606 |
| revenu | 1 | -0.2981 | 0.1201 | 6.1643 | 0.0130 |
| education | 1 | 0.0170 | 0.1159 | 0.0215 | 0.8833 |

**Figure 9 – CUMUL. LOGITS + Proportional Odds - Variables + Constantes (SAS)**