NOTES DE STATISTIQUE
ET D'INFORMATIQUE

2011/4

LA REGRESSION LOGISTIQUE
AVEC MINITAB

R. PALM et Y. BROSTAUX

Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique
appliquées à la bioingénierie
GEMBLOUX
(Belgique)

LA RÉGRESSION LOGISTIQUE AVEC MINITAB

R. Palm * et Y. Brostaux †

RÉSUMÉ

Cette note décrit les commandes BLOGISTIC, NLOGISTIC et OLOGISTIC de Minitab et propose une macro, qui peut être téléchargée, permettant de recoder les données en vue de l'ajustement du modèle logit basé sur les catégories adjacentes supérieures cumulées. Les différentes commandes sont illustrées par un exemple.

SUMMARY

This note describes the Minitab commands BLOGISTIC, NLOGISTIC and OLOGISTIC and proposes a macro, which can be downloaded, for coding the data in order to fit the continuation-ratio logits model. The commands are illustrated by an example.

1. INTRODUCTION

Comme la régression classique au sens des moindres carrés, la régression logistique a pour objectif l'étude de la relation entre une variable à expliquer et une ou plusieurs variables explicatives. La différence entre les deux approches est liée à la nature de la variable à expliquer : alors que la régression classique s'applique aux cas où la variable à expliquer est quantitative, en principe continue et d'erreur normale, la régression logistique concerne le cas où la variable à expliquer est qualitative.

En relation avec la nature de cette variable qualitative, trois situations différentes peuvent se rencontrer. Lorsque la variable à expliquer présente uniquement deux modalités, on a affaire à la régression logistique binaire. Si la variable à expliquer présente plusieurs modalités sans que l'ordre de celles-ci ne soit pris en compte, on utilise la régression logistique nominale. Enfin, lorsque la variable à expliquer présente plus de deux modalités et que ces modalités

*Professeur à l'Université de Liège, Gembloux Agro-Bio Tech.
†Chef de Travaux et Chargé de cours à l'Université de Liège, Gembloux Agro-Bio Tech.

sont ordonnées, on se trouve dans le cas de la régression logistique ordinale, qui elle-même présente plusieurs variantes.

Le logiciel Minitab offre trois commandes permettant d'appliquer ces trois méthodes et l'objectif de cette note est d'aider l'utilisateur en vue d'une utilisation correcte de ces commandes.

Après cette introduction (paragraphe 1), nous présentons d'abord les éléments communs aux trois méthodes (paragraphe 2). Ensuite, nous donnons des informations plus spécifiques pour chacune des méthodes, qui sont alors illustrées par un exemple (paragraphes 3, 4 et 5). Pour la régression logistique ordinale, le logiciel Minitab ne propose qu'une variante, basée sur les probabilités cumulées. Une deuxième variante, basée sur les catégories adjacentes supérieures cumulées peut cependant être réalisée facilement à partir d'une régression logistique binaire sur les données recodées. Cette méthode est décrite au paragraphe 6 et une macro, appelée OLOGISCASC, permettant le recodage des données est proposée.

L'objectif de cette note n'est donc pas de détailler de façon complète la régression logistique ni d'interpréter les résultats des modèles obtenus. Le lecteur trouvera des informations à ce sujet dans des ouvrages spécialisés, parmi lesquels on peut citer les livres d'Agresti [2002] et de Hosmer et Lemeshow [2000]. Pour la régression binaire, des informations sont également données dans les documents de Duyme et Claustriaux [2006] et de Palm *et al.* [2011]. Enfin, les principaux modèles utilisés en régression logistique sont encore décrits par Gillet *et al.* [2011].

Les données qui nous serviront à illustrer l'utilisation de Minitab proviennent de cette dernière publication. Elles concernent le niveau de dépérissement de 230 chênes et l'altitude des stations dans lesquelles ces chênes ont été observés. Le dépérissement a été évalué par l'aspect du houppier sur une échelle à quatre niveaux, le niveau 1 étant le plus faible et le niveau 4 le niveau le plus fort. Des modifications seront cependant apportées dans ces variables afin de permettre l'illustration des différentes situations examinées. Pour chaque exemple traité, une figure reprend la liste des commandes et une autre figure donne les résultats de la régression logistique. Des compléments d'information concernant les exemples sont donnés dans Gillet *et al.* [2011].

Le fichier initial de données, ainsi que la macro assurant le recodage des données et les différentes listes de commandes présentées dans les exemples sont à la disposition des lecteurs et accessibles à partir de la page :

http ://www.fsagx.ac.be/si/

en cliquant, dans le menu général, sur le lien « Macros » et en sélectionnant ensuite le thème relatif à la régression logistique.

2. INFORMATIONS GÉNÉRALES RELATIVES AUX TROIS COMMANDES

2.1. Présentation des données

Les données peuvent être présentées sous la forme d'un tableau de données brutes, à raison d'une ligne par individu observé, comme pour une régression classique. Si plusieurs individus présentent les mêmes valeurs pour l'ensemble des variables, une colonne supplémentaire reprenant les fréquences peut être introduite, évitant ainsi la répétition de lignes identiques.

Dans le cas de la régression logistique binaire, les données peuvent également être introduites sous la forme de nombres de succès et nombres d'essais. Dans ce cas, le fichier contient, en plus des variables explicatives, deux variables, l'une donnant les nombres de succès, c'est-à-dire le nombre d'apparitions de l'événement pris comme référence, l'autre donnant le nombre d'essais, c'est-à-dire le nombre d'observations pour une combinaison de variables explicatives. Le fichier ne comporte alors qu'une ligne par combinaison de variables explicatives.

2.2. Nature des variables explicatives

Les variables explicatives peuvent être quantitatives ou qualitatives. Dans Minitab, les premières sont appelées *covariables* et les secondes *facteurs*[1]. Dans le fichier de données, les facteurs peuvent être alphabétiques ou numériques mais, s'ils sont numériques, les nombres qu'ils contiennent sont considérés comme des codes correspondant aux variantes des facteurs. Les facteurs doivent être déclarés comme tels dans une des options de Minitab.

2.3. Spécification du modèle

La spécification du modèle se fait en régression logistique comme pour le modèle linéaire général (commande GLM de Minitab). Ainsi, l'utilisateur peut introduire des variables explicatives quantitatives (covariables) et qualitatives (facteurs), ces variables correspondant à des colonnes particulières du fichier. Il peut également introduire des termes d'interaction ou encore tenir compte d'une hiérarchie de facteurs.

2.4. Fonction de lien

Pour la régression logistique binaire et la régression logistique ordinale, Minitab propose les trois fonctions de lien suivantes :

- la fonction logit,
- la fonction probit (ou normit),
- la fonction complémentaire log-log (ou gompit),

1. En anglais : *covariates* et *factors*.

l'option par défaut étant l'utilisation de la fonction logit. Cette fonction de lien est la fonction la plus couramment utilisée en pratique, et, par la suite, nous ferons uniquement référence à celle-ci.

Pour la régression nominale, Minitab utilise toujours la fonction logit.

2.5. Modalité de référence pour la variable à expliquer

La régression logistique binaire a comme objectif de modéliser la probabilité d'apparition d'un événement. Ainsi par exemple, si les deux modalités sont A et B, on peut modéliser la probabilité d'apparition soit de la modalité A, soit de la modalité B. La modalité qu'on modélise est appelée *événement de référence*[2].

Pour la régression logistique nominale, l'événement de référence est l'événement qui sert de base de comparaison. Ainsi par exemple, si on a 4 modalités, codées 1, 2, 3 et 4 et qu'on souhaite modéliser les logits suivants :

$$\log_e(\pi_i/\pi_4) \quad (i = 1, ..., 3),$$

l'événement de référence correspond à la modalité 4. Dans cette relation, π_i correspond à la probabilité d'apparition de la modalité i et π_4 à la probabilité d'apparition de la modalité de référence.

Pour la régression ordinale, il n'y a pas d'événement de référence, mais il faudra spécifier la séquence des modalités qui correspond à l'ordre croissant de celles-ci.

En l'absence d'information donnée par l'utilisateur concernant la modalité de référence, Minitab considère que la modalité de référence est celle qui correspond à la dernière modalité lorsque celles-ci sont rangées par ordre croissant ou par ordre alphabétique, selon que la variable à expliquer est une variable numérique ou alphabétique. Ainsi, pour une variable explicative nominale numérique codée 1, 2 et 3, la modalité de référence sera, par défaut, la modalité 3. Si la variable est alphabétique et présente, par exemple, les trois modalités « rouge », « vert » et « jaune », la modalité de référence sera, par défaut, la modalité « vert ».

Pour la régression logistique ordinale, l'ordre croissant considéré par défaut est l'ordre donné par les valeurs des modalités ou l'ordre alphabétique, selon que la variable à expliquer est une variable numérique ou alphabétique. Ainsi, pour une variable explicative ordinale à trois niveaux codés 1, 2, 3, Minitab considère, par défaut, que les trois modalités sont dans l'ordre croissant. Si les trois modalités sont par contre codées « pas », « peu » et « fort », l'ordre croissant considéré par défaut serait : « fort », « pas », « peu ».

2. En anglais : *reference event*.

2.6. Modalités de référence pour les facteurs

En présence de variables explicatives qualitatives, l'utilisateur peut préciser quelle est la modalité de référence, c'est-à-dire celle qui n'apparaitra pas dans le tableau de résultats. Par défaut, Minitab considère que la modalité de référence est celle qui est classée la première lorsque les modalités sont ordonnées par valeurs croissantes ou par ordre alphabétique, selon que le facteur est représenté par une variable numérique ou par une variable alphabétique.

Ainsi, si un facteur est représenté par une variable alphabétique à deux modalités, notées « F » et « M », la modalité de référence sera par défaut la modalité « F » et, dans la sortie des résultats, un coefficient apparaitra en regard de la modalité « M ».

2.7. Estimation des paramètres

L'utilisateur peut introduire des *valeurs initiales*[3] pour les paramètres du modèle. Il peut également modifier le nombre maximum d'itérations lors de l'estimation des paramètres par la méthode du maximum de vraisemblance, ce nombre étant fixé, par défaut, à 20. Ces options peuvent être utiles si la convergence de la solution n'a pas été obtenue avec les options par défaut.

L'utilisateur peut également imposer des valeurs pour les paramètres. Cette option permet notamment de calculer les probabilités relatives aux événements pour des données autres que celles qui ont servi à l'ajustement du modèle, ces paramètres ayant été préalablement enregistrés dans une colonne du fichier (paragraphe 2.8). On notera cependant que cette estimation n'est possible que si on dispose de valeurs de la variable à expliquer, comme c'est par exemple le cas pour des données de validation externe.

2.8. Visualisation et enregistrement des résultats

Des options particulières permettent de contrôler le volume des résultats qui apparaissent dans la session. Pour les exemples présentés dans les paragraphes suivants, cette possibilité a été utilisée pour limiter la taille des figures.

L'utilisateur peut également demander l'enregistrement de différents résultats et notamment des coefficients des équations de régression et des probabilités estimées.

3. RÉGRESSION LOGISTIQUE BINAIRE

3.1. Commande BLOGISTIC

Si la fonction de lien est la fonction logit, le modèle s'écrit :

3. En anglais : *starting estimates.*

$$\log_e(\pi_1/\pi_0) = a + \boldsymbol{x}\boldsymbol{b},$$

π_1 représentant la probabilité d'appartenance à la modalité de référence et π_0 la probabilité d'appartenance à l'autre modalité ; $\boldsymbol{x}$ est le vecteur des variables explicatives, a est l'ordonnée à l'origine et $\boldsymbol{b}$ est le vecteur des coefficients de régression estimés.

La régression logistique binaire est réalisée avec la commande BLOGISTIC. Les informations concernant les données, le modèle et la plupart des options ont été présentées au paragraphe 2.

Si deux ou plusieurs individus sont caractérisés par le même vecteur de variables explicatives, Minitab offre la possibilité d'associer la probabilité estimée à chaque individu (option par défaut), ou, au contraire, uniquement au premier de ces individus, le code de données manquantes étant attribué aux autres.

Une option spécifique à la régression binaire concerne le test de HOSMER et LEMESHOW. Ce test est basé sur la comparaison de fréquences observées et attendues d'appartenance aux deux classes quand les données sont réparties en groupes. Par défaut, le nombre de groupes est fixé à 10, mais l'utilisateur peut modifier ce nombre.

3.2. Exemple

La variable à expliquer, appelée *Deper* dans le fichier Minitab est, au départ, une variable qualitative ordinale, à quatre niveaux croissants codés 1, 2, 3 et 4 (paragraphe 1).

Pour illustrer la régression binaire, cette variable a subi un recodage : le code « Très faible » est attribué au niveau initialement codé 1, le code « Faible à très fort » est attribué aux niveaux initialement codés 2, 3 et 4. La variable recodée s'appelle *Dep-binaire*. L'appartenance à la première classe est choisie arbitrairement comme l'événement de référence.

La figure 1 donne les commandes permettant de recoder les données et de réaliser la régression logistique binaire, en prenant comme événement de référence le code « Très faible » et en demandant l'enregistrement, dans le fichier de données, des probabilités relatives à la modalité de référence et des coefficients de l'équation de régression. La figure 2 donne les résultats de la régression.

4. RÉGRESSION LOGISTIQUE NOMINALE

4.1. Commande NLOGISTIC

Pour une variable explicative à k modalités et en considérant arbitrairement que la $k^{ième}$ modalité est la modalité de référence, le modèle s'écrit :

```
# Codage de la variable Deper en variable binaire
  Code (1) "Tres faible" (2:4) "Faible a tres fort" 'Deper' c3
 name c3 'Dep_binaire'

# Regression logistique binaire
 Name c4 "EPRO1" c5 "COEF1"
 Blogistic 'Dep_binaire' = Altitude;
   Logit;
   Reference 'Dep_binaire'  "Tres faible";
   Eprobability 'EPRO1';
   Coefficients 'COEF1';
   Brief 1.
 Name c4 'Pi_TFaible'
```

Figure 1 – Commandes pour le recodage de la variable à expliquer et pour la régression logistique binaire.

```
Binary Logistic Regression: Dep_binaire versus Altitude

Link Function: Logit


Response Information

Variable     Value               Count
Dep_binaire  Tres faible            47  (Event)
             Faible a tres fort    183
             Total                 230


Logistic Regression Table

                                                    Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant     1.57141   0.633517   2.48  0.013
Altitude  -0.0095331  0.0021457  -4.44  0.000   0.99   0.99   0.99


Log-Likelihood = -103.428
Test that all slopes are zero: G = 26.075, DF = 1, P-Value = 0.000
```

Figure 2 – Résultats pour la régression logistique binaire.

$$\log_e(\pi_j/\pi_k) = a_j + \boldsymbol{x}\boldsymbol{b}_j \quad (j = 1, \cdots, k-1).$$

La régression logistique nominale est réalisée par la commande NLOGISTIC, dont les caractéristiques principales ont été décrites au paragraphe 2.

Dans le cas où l'utilisateur demande les probabilités estimées π_j pour les différentes modalités il devra être attentif à l'ordre dans lequel celles-ci sont enregistrées : la première probabilité correspond à la modalité de référence et les autres probabilités sont données dans l'ordre dans lequel apparaissent les fonctions logistiques dans la sortie de Minitab.

4.2. Exemple

La figure 3 donne les commandes pour les données relatives au dépérissement du chêne, en considérant les quatre niveaux de dépérissement comme quatre modalités, sans prendre en compte l'ordre de celles-ci. La modalité 4 est prise comme référence et les coefficients des équations de régression sont enregistrés ainsi que les probabilités d'appartenance aux groupes. Les colonnes contenant ces probabilités sont renommées pour éviter toute confusion dans les modalités. Les résultats sont repris à la figure 4.

5. RÉGRESSION LOGISTIQUE ORDINALE POUR LES PROBABILITÉS CUMULÉES

5.1. Commande OLOGISTIC

Pour une variable dépendante qualitative ordinale y à k modalités ordonnées de manière croissante, les probabilités modélisées par Minitab sont les probabilités cumulées jusqu'à la modalité j :

$$\pi_1 + \cdots + \pi_j = P(y \leq j) \quad (j = 1, \cdots, k-1).$$

Si la fonction de lien est la fonction logit, le modèle s'écrit :

$$\log_e\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_k}\right) = a_j + \boldsymbol{x}\boldsymbol{b} \quad (j = 1, \cdots, k-1).$$

On considère donc que les différents logits ont le même vecteur de coefficients de régression mais des ordonnées à l'origine différentes.

Les informations concernant la présentation des données et du modèle, ainsi que la plupart des options ont été examinées au paragraphe 2.

```
# Regression logistique nominale
 Name c3 "COEF1" c4 "EPROB1" c5 "EPROB2" c6 "EPROB3" c7 "EPROB4"
 NLogistic 'Deper' = Altitude;
   Reference 'Deper' 4;
   Coefficients 'COEF1';
   Eprobability 'EPROB1'-'EPROB4';
   Brief 1.
Name  c4 "Pi_4" c5 "Pi_3" c6 "Pi_2" c7 "Pi_1"
```

Figure 3 – Commandes pour la régression logistique nominale.

```
Nominal Logistic Regression: Deper versus Altitude

Response Information

Variable  Value  Count
Deper     4         23  (Reference Event)
          3         49
          2        111
          1         47
          Total    230


Logistic Regression Table

                                                    Odds   95% CI
Predictor         Coef   SE Coef      Z     P   Ratio Lower Upper
Logit 1:(3/4)
Constant        2.5449   0.97611   2.61  0.009
Altitude     -0.004436  0.002283  -1.94  0.052  1.00  0.99  1.00
Logit 2:(2/4)
Constant        4.9582   0.93485   5.30  0.000
Altitude     -0.008987  0.002233  -4.02  0.000  0.99  0.99  1.00
Logit 3:(1/4)
Constant       6.35134    1.0832   5.86  0.000
Altitude    -0.0165789  0.002979  -5.57  0.000  0.98  0.98  0.99


Log-Likelihood = -260.950
Test that all slopes are zero: G = 46.556, DF = 3, P-Value = 0.000
```

Figure 4 – Résultats pour la régression logistique nominale.

5.2. Exemple

La figure 5 donne les commandes pour les données relatives au dépérissement du chêne, avec enregistrement des probabilités et des probabilités cumulées et la figure 6 donne les résultats obtenus. Ici aussi, les variables correspondant aux probabilités ont été renommées.

6. RÉGRESSION LOGISTIQUE ORDINALE BASÉE SUR LES CATÉGORIES ADJACENTES SUPÉRIEURES CUMULÉES

6.1. Macro OLOGISCASC

Pour une variable dépendante y à k modalités ordonnées de manière croissante, les probabilités modélisées dans cette approche sont les probabilités conditionnelles :

$$P(y = j|y \geq j) = \pi_j/(\pi_j + \cdots + \pi_k) \quad (j = 1, \cdots, k-1).$$

Si la fonction de lien est la fonction logit, on a :

$$\log_e\ [\pi_j/(\pi_{j+1} + \cdots + \pi_k)] = a_j + \boldsymbol{xb} \quad (j = 1, \cdots, k-1),$$

en considérant, comme dans le paragraphe précédent, que les différents logits ont le même vecteur de coefficients de régression mais des ordonnées à l'origine différentes. Minitab ne propose pas de commande spécifique pour cette régression mais les résultats peuvent être obtenus en ajustant un modèle binaire (commande BLOGISTIC) sur les données recodées. Le principe consiste à empiler les données correspondant à $k-1$ régressions binaires et à ajouter des variables artificielles, qui sont utilisées comme variables explicatives supplémentaires. Ce recodage est expliqué au paragraphe 6.2, dans le cas du dépérissement des chênes. En pratique, il peut être réalisé par la macro OLOGISCASC proposée aux utilisateurs (paragraphe 1).

6.2. Exemple

Le tableau 1 donne la structure des données recodées dans le cas du dépérissement des chênes. Chaque ligne du tableau correspond, dans le fichier réel des données recodées, à un nombre de lignes fonction des nombres d'observations n_j présentes dans les différentes catégories.

Les deux premières lignes du tableau 1 correspondent à une première régression binaire traduisant la dichotomie $Deper = 1$ et $Deper > 1$; les deux suivantes correspondent à la dichotomie $Deper = 2$ et $Deper > 2$ et les deux dernières à la dichotomie $Deper = 3$ et $Deper = 4$. Au total, le fichier sera donc constitué de :

```
# Modele logit basé sur les probabilités cumulées
 Name c3 "COEF1" c4 "EPROB1" c5 "EPROB2" c6 "EPROB3" c7 "EPROB4" &
     c8 "CUMP1" c9 "CUMP2" c10 "CUMP3"
 OLogistic 'Deper' = Altitude;
   Logit;
   Order 1 2 3 4;
   Coefficients 'COEF1';
   Eprobability 'EPROB1'-'EPROB4';
   Cumprobability 'CUMP1'-'CUMP3';
   Brief 1.
 Name  c4 "Pi_1" c5 "Pi_2" c6 "Pi_3" c7 "Pi_4"
 Name  c8 "P(<=1)" c9 "P(<=2)" c10 "P(<=3)"
```

Figure 5 – Commandes pour la régression logistique ordinale pour les probabilités cumulées.

```
Ordinal Logistic Regression: Deper versus Altitude

Link Function: Logit


Response Information

Variable  Value  Count
Deper     1         47
          2        111
          3         49
          4         23
          Total    230


Logistic Regression Table

                                                    Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Const(1)     1.18366   0.414712   2.85  0.004
Const(2)     3.65144   0.476404   7.66  0.000
Const(3)     5.24258   0.547170   9.58  0.000
Altitude  -0.0081690  0.0012598  -6.48  0.000   0.99   0.99   0.99

Log-Likelihood = -261.658
Test that all slopes are zero: G = 45.140, DF = 1, P-Value = 0.000
```

Figure 6 – Résultats pour la régression logistique ordinale pour les probabilités cumulées.

Tableau 1 – Structure du tableau des données recodées pour le modèle basé sur les catégories adjacentes supérieures cumulées.

Nombre de lignes	*Deper*	z	u_1	u_2
n_1	1	1	1	0
$n_2 + n_3 + n_4$	2, 3 ou 4	0	1	0
n_2	2	1	0	1
$n_3 + n_4$	3 ou 4	0	0	1
n_3	3	1	0	0
n_4	4	0	0	0

$$n_1 + 2n_2 + 3n_3 + 3n_4$$

lignes. En plus des variables $Deper, z, u_1$ et u_2, le fichier réel comporte une colonne, notée x, reprenant, pour chacune des lignes, l'altitude. Les variables artificielles u_1 et u_2 permettent d'identifier les individus relatifs à la première $(u_1 = 1, u_2 = 0)$, à la deuxième $(u_1 = 0, u_2 = 1)$ et à la troisième $(u_1 = 0, u_2 = 0)$ régression binaire.

Une régression logistique binaire est alors calculée en considérant que la variable à expliquer est la variable z. Le modèle de régression ajusté à ces données est le suivant :

$$\log_e \left[\frac{P(z=1)}{P(z=0)}\right] = a + bx + b_1 u_1 + b_2 u_2,$$

$P(z = 1)$ et $P(z = 0)$ étant les probabilités que z soit égal à 1 et à 0.

La figure 7 reprend les commandes permettant l'exécution de la macro OLOGISCASC et la réalisation, sur la base de ce recodage, de la régression logistique binaire, avec l'enregistrement des probabilités estimées. Les résultats obtenus sont donnés à la figure 8.

La régression logistique binaire obtenue s'écrit (figure 8) :

$$\log_e \left[\frac{P(z=1)}{P(z=0)}\right] = 3{,}4897 - 0{,}006730x - 2{,}7302u_1 - 0{,}6356u_2.$$

Il en résulte que les équations pour la régression ordinale s'écrivent :

```
# Codage des données
%OLOGISCASC c1 c2 ;
 OUTYX c11 c12 ;
 OUTU c13 c14.
name c11 'z' c12 'x' c13 'u1' c14 'u2'

# Regression logistique binaire
 Name c15 "EPRO1"
 Blogistic 'z' =  x u1 u2;
   Logit;
   Reference 'z'  1;
   Eprobability 'EPRO1';
   Brief 1.
```

Figure 7 – Commandes pour le codage des données et la régression logistique binaire sur les données recodées.

```
Binary Logistic Regression: z versus x; u1; u2

Link Function: Logit

Response Information

Variable  Value  Count
z         1        207  (Event)
          0        278
          Total    485

Logistic Regression Table

                                                    Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant     3.48972   0.530792   6.57  0.000
x         -0.0067302  0.0011025  -6.10  0.000   0.99   0.99   1.00
u1          -2.73015   0.342234  -7.98  0.000   0.07   0.03   0.13
u2         -0.635561   0.316430  -2.01  0.045   0.53   0.28   0.98

Log-Likelihood = -262.407
Test that all slopes are zero: G = 137.107, DF = 3, P-Value = 0.000
```

Figure 8 – Résultats de la régression logistique binaire sur les données recodées.

$$\log_e [\pi_1/(\pi_2+\pi_3+\pi_4)] = (3{,}4897 - 2{,}7302) - 0{,}006730x,$$

$$\log_e [\pi_2/(\pi_3+\pi_4)] = (3{,}4897 - 0{,}6356) - 0{,}006730x,$$

et
$$\log_e (\pi_3/\pi_4) = 3{,}4897 - 0{,}006730x.$$

Les probabilités reprises dans la colonne EPR01 sont les probabilités :

$$P(Deper = j | Deper \geq j),$$

la valeur de j se déduisant des variables artificielles u_1 et u_2 :

$$\begin{array}{llllll} j=1 & \text{si} & u_1=1 & \text{et} & u_2=0 \\ j=2 & \text{si} & u_1=0 & \text{et} & u_2=1 \\ j=3 & \text{si} & u_1=0 & \text{et} & u_2=0. \end{array}$$

7. QUELQUES INFORMATIONS COMPLÉMENTAIRES

Les trois commandes de Minitab pour la régression logistique ont été décrites et illustrées à l'aide d'un exemple. Les calculs ont été réalisés avec Minitab 16, version anglaise [Minitab, 2010]. Pour chaque cas, la liste des commandes a été donnée. Ces listes de commandes ont été générées directement à partir des écrans. En particulier, les noms des variables enregistrées dans le fichier de travail ont été générés de manière automatique. Toutefois, pour plus de clarté, ils ont ensuite été modifiés afin d'être mieux adaptés aux situations concrètes.

Comme signalé dans l'introduction, l'objectif de cette note n'était pas de présenter dans le détail les méthodes de régression logistique, ni de commenter ou d'interpréter les résultats. Des références bibliographiques ont été données au paragraphe 1 pour permettre à l'utilisateur d'approfondir, si nécessaire, ses connaissances sur le sujet. Par ailleurs, les exemples ont été commentés dans l'article de Gillet *et al.* [2011].

Enfin, nous rappelons que les données de départ, le fichier reprenant les commandes utililsées et la macro OLOGISCASC sont disponibles à partir de l'adresse suivante : http ://www.fsagx.ac.be/si/.

BIBLIOGRAPHIE

Agresti A. [2002]. *An introduction to categorical data analysis.* New-York, Wiley, 710 p.

Duyme F., Claustriaux J. J. [2006]. La régression logistique binaire. *Notes Stat. Inform.* (Gembloux) 2004-6, 26 p.

Gillet A., Brostaux Y. et Palm R. [2011]. Principaux modèles utilisés en régression logistique. *Biotechnol. Agron. Soc. Environ.* **15** (3), 425-433.

Hosmer D. W. et Lemeshow S. [2000]. *Applied logistic regression.* New-York, Wiley, 392 p.

Minitab [2010]. Meet Minitab 16. Document PDF. <http ://www.minitab.com>

Palm R., Brostaux Y. et Claustriaux J. J. [2011]. Inférence statistique et critères de qualité de l'ajustement en régression logistique binaire. *Notes Stat. Inform.* (Gembloux) 2011/5 , 32 p.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant de l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de la Faculté universitaire des Sciences agronomiques et du Département de Biométrie, Gestion des données et Agrométéorologie du Centre wallon de Recherches agronomiques (Gembloux - Belgique).

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLOUX (Belgique)
E-mail : sima.gembloux@ulg.ac.be

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

http://www.fsagx.ac.be/si/

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

PALM R. [2008]. Détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée selon la norme ISO 5725-2. *Notes Stat. Inform.* (Gembloux) 2008/1, 22 p.

CHARLES C., LECHARLIER L., RENAUD F. [2008]. Introduction à LATEX. *Notes Stat. Inform.* (Gembloux) 2008/2, 21 p.

CHARLES C. [2008]. Introdution à OCTAVE. *Notes Stat. Inform.* (Gembloux) 2008/3, 19 p.

PALM R., BROSTAUX Y. [2009]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 2009/1, 17 p.

CHARLES C. [2011]. Introduction aux ondelettes. *Notes Stat. Inform*. (Gembloux) 2011/1, 22 p.

CHARLES C. [2011]. Introduction aux applications des ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/2, 35 p.

PALM R., BROSTAUX Y. et CLAUSTRIAUX J. J. [2011]. Macros Minitab pour le choix d'une transformation pour la normalisation de variables. *Notes Stat. Inform.* (Gembloux) 2011/3, 15 p.

D/2011/2371/4