NOTES DE STATISTIQUE
ET D'INFORMATIQUE

2011/5

INFÉRENCE STATISTIQUE
ET CRITÈRES DE QUALITÉ DE L'AJUSTEMENT
EN RÉGRESSION LOGISTIQUE BINAIRE

R. PALM, Y. BROSTAUX et J. J. CLAUSTRIAUX

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

INFÉRENCE STATISTIQUE ET CRITÈRES DE QUALITÉ DE L'AJUSTEMENT EN RÉGRESSION LOGISTIQUE BINAIRE

R. PALM [*], Y. BROSTAUX [†] et J. J. CLAUSTRIAUX [‡]

RÉSUMÉ

Les principes de l'inférence statistique basée sur la fonction de vraisemblance sont rappelés et ensuite appliqués à la régression logistique binaire. Différents critères globaux d'ajustement de la régression logistique binaire sont également présentés.

Les diverses notions sont illustrées par un exemple traité par le logiciel Minitab et par le logiciel SAS.

SUMMARY

The principles of statistical inference based on the maximum likelihood function are described and applied to binary logistic regression. Several goodness of fit statistics are also given.

These topics are illustrated by an example processed by Minitab and SAS softwares.

1. INTRODUCTION

La régression logistique est une méthode statistique qui vise à mettre en relation une variable à expliquer de nature qualitative avec une ou plusieurs variables explicatives. Elle offre plusieurs variantes en fonction du nombre et de la nature des modalités de la variable à expliquer. Lorsque celle-ci présente uniquement deux modalités, il s'agit de la *régression logistique binaire* [1]. Si la variable à expliquer présente plusieurs modalités sans que l'ordre de celles-ci ne soit pris

[*] 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.
[‡] Professeur ordinaire à l'Université de Liège, Gembloux Agro-Bio Tech.

1. En anglais : *binary logistic regression.*

en considération, on utilise la *régression logistique nominale*[2]. Enfin, lorsque la variable à expliquer présente plus de deux modalités et que ces modalités sont ordonnées, on se trouve dans le cas de la *régression logistique ordinale*[3].

L'objectif de cette note est d'aider l'utilisateur débutant ou occasionnel de la régression logistique à mieux comprendre les résultats fournis par les logiciels statistiques. Nous nous focalisons sur les principaux tests statistiques et sur les mesures globales de la qualité des ajustements réalisés. Nous n'abordons pas les problèmes de construction de modèles et de choix de variables ni l'examen détaillé des résidus et des mesures de l'influence des observations. Nous nous limitons également à la régression logistique binaire, bien que différentes notions qui sont présentées puissent s'appliquer ou être étendues aux autres cas de régression logistique.

Le lecteur trouvera des informations complémentaires sur la régression logistique dans les ouvrages spécialisés, parmi lesquels on peut citer les livres d'Agresti [2002] et de Hosmer et Lemeshow [2000]. Une présentation très synthétique de la régression logistique binaire est également disponible dans Duyme et Claustriaux [2006]. Enfin, des informations plus directement en relation avec le logiciel SAS sont données par Allison [1999].

Après cette introduction (paragraphe 1), nous rappelons la notion de fonction de vraisemblance et son utilisation pour l'estimation de paramètres (paragraphe 2). Le modèle logit pour données binaires est alors présenté (paragraphe 3). Le paragraphe 4 a trait à la fonction de vraisemblance dans le cas de la régression logistique. Le paragraphe 5 est consacré aux tests de signification des coefficients de régression. Les critères globaux d'ajustement utilisés en régression logistique binaire sont alors décrits (paragraphe 6). Nous clôturons enfin par quelques informations complémentaires (paragraphe 7).

Les différentes notions sont illustrées par un exemple traité, d'une part, avec Minitab [Minitab, 2010] et, d'autre part, avec SAS [SAS Institute Inc, 2010]. Les listes de commandes Minitab et la procédure SAS utilisées pour générer les résultats repris dans les figures illustrant cette note sont regroupées en annexe. Les données retenues concernent le niveau de dépérissement de 230 chênes observés dans deux régions naturelles et l'altitude des stations dans lesquelles ces chênes ont été observés. Elles proviennent d'une étude de Gillet [2005] et ont déjà été utilisées antérieurement pour illustrer les différents modèles de régression logistique [Gillet *et al.* 2011] et leur analyse avec Minitab [Palm et Brostaux, 2011]. Le dépérissement a été évalué par l'aspect du houppier sur une échelle à quatre niveaux. La variable à expliquer, qui est donc, au départ, une variable qualitative ordinale, a été recodée sous forme binaire. La première modalité correspond au dépérissement « très faible » et la seconde modalité au dépérissement « faible à très fort ». L'appartenance à la deuxième classe a été choisie arbitrairement comme la modalité de référence. Nous supposons également que les 230 arbres observés peuvent être considérés comme constituant un échantillon aléatoire et simple d'arbres choisis parmi tous les arbres atteints de dépérissement dans la zone considérée.

2. En anglais : *polytomous nominal logistic regression.*
3. En anglais : *polytomous ordinal logistic regression.*

2. PRINCIPE DU MAXIMUM DE VRAISEMBLANCE

2.1. Inférence relative à une proportion

L'inférence statistique en régression logistique repose très largement sur la fonction de vraisemblance et l'objectif de ce paragraphe est de rappeler ce qu'est cette fonction et comment elle est utilisée pour estimer des paramètres, calculer des limites de confiance et réaliser des tests d'hypothèses. Nous envisageons d'abord un problème particulièrement simple qui ne fait intervenir qu'un seul paramètre et pour lequel les calculs peuvent être réalisés sans recourir à des programmes très spécifiques. Il s'agit de l'estimation d'une proportion à partir d'un échantillon aléatoire et simple et des problèmes d'inférence associés. La généralisation à d'autres situations sera abordée au paragraphe 2.2.

En relation avec l'exemple présenté dans l'introduction, on se propose d'estimer la proportion de chênes caractérisés par un dépérissement faible à très fort parmi les chênes atteints, ainsi que les limites de confiance correspondantes. De plus, on souhaite vérifier si cette proportion doit être considérée ou non comme différente de 50 %. Sur les 230 arbres atteints de dépérissement qui ont été examinés, 183 sont atteints d'un dépérissement faible à très fort et 47 d'un dépérissement très faible.

Soit π la proportion à estimer, c'est-à-dire la proportion d'arbres atteints d'un dépérissement faible à très fort parmi les chênes atteints dans la zone considérée. L'expression, en fonction de π, de la probabilité d'observer 183 individus présentant un dépérissement faible à très fort parmi 230 arbres observés est appelée fonction de vraisemblance et est notée $L(\pi)$. Cette probabilité est donnée par la loi binomiale de paramètre n et π :

$$L(\pi) = \mathrm{P}(X = x) = \mathrm{C}_n^x \, \pi^x (1-\pi)^{n-x},$$

X étant la variable aléatoire décrivant le nombre d'individus, parmi les n individus prélevés, qui ont un dépérissement faible à très fort [4]. On a donc, pour $n = 230$ et $x = 183$:

$$L(\pi) = \mathrm{C}_{230}^{183} \, \pi^{183} (1-\pi)^{47}.$$

Le facteur C_{230}^{183} est un nombre et ne dépend donc pas de π. Dans la mesure où, par la suite, ce sont essentiellement des valeurs relatives, telles que des rapports de valeurs de la fonction de vraisemblance pour différentes valeurs de π, qui vont nous intéresser, on peut éliminer cette constante. On a alors :

$$L(\pi) = \pi^x (1-\pi)^{n-x} = \pi^{183} (1-\pi)^{47}.$$

L'estimation du paramètre π au sens du maximum de vraisemblance consiste à déterminer la valeur de π, notée $\hat{\pi}$, qui rend maximum $L(\pi)$. Cette estimation

4. Nous désignons par C_n^x le nombre de combinaisons de n objets pris par groupes de x objets au sens de l'analyse combinatoire. Ce nombre est souvent désigné par $\binom{n}{x}$.

peut être obtenue en dérivant $L(\pi)$ par rapport à π et en recherchant la valeur $\hat{\pi}$ qui annule cette dérivée. Le problème est cependant sensiblement simplifié si on remplace la fonction $L(\pi)$ par son logarithme. La transformation logarithmique étant une transformation monotone croissante, la valeur $\hat{\pi}$ qui rend maximum $L(\pi)$ est aussi la valeur qui rend maximum $\log_e L(\pi)$:

$$\log_e L(\pi) = x \log_e(\pi) + (n - x) \log_e(1 - \pi).$$

La dérivée du logarithme de la vraisemblance s'écrit :

$$\frac{d \log_e L(\pi)}{d \pi} = \frac{x}{\pi} - \frac{n - x}{1 - \pi}.$$

En annulant cette dérivée :

$$\frac{x}{\hat{\pi}} - \frac{n - x}{1 - \hat{\pi}} = 0,$$

on obtient :

$$\hat{\pi} = \frac{x}{n}.$$

La proportion observée est donc une estimation du maximum de vraisemblance d'une proportion théorique. La figure 1 donne la représentation graphique du logarithme de la vraisemblance pour l'exemple considéré. Elle atteint son maximum pour :

$$\hat{\pi} = 183/230 = 0{,}7957 \quad \text{ou} \quad 0{,}80.$$

On considère donc que, parmi les chênes atteints de dépérissement, 80 % présentent un dépérissement faible à très fort.

Il a été démontré que l'erreur-standard d'un paramètre estimé par la méthode du maximum de vraisemblance est égale à l'inverse de la racine carrée de l'information i, définie de la manière suivante :

$$i = -\mathrm{E}\left[\frac{d^2 \log_e L(\pi)}{d \pi^2}\right].$$

Il s'agit donc de calculer l'espérance mathématique de la dérivée seconde par rapport à π de la fonction de vraisemblance. Pour calculer cette espérance mathématique, on remplace, dans la fonction de vraisemblance x par X et on trouve :

$$i = -\mathrm{E}\left[-\frac{X}{\pi^2} - \frac{n - X}{(1 - \pi)^2}\right] = \frac{\mathrm{E}(X)}{\pi^2} + \frac{\mathrm{E}(n - X)}{(1 - \pi)^2}.$$

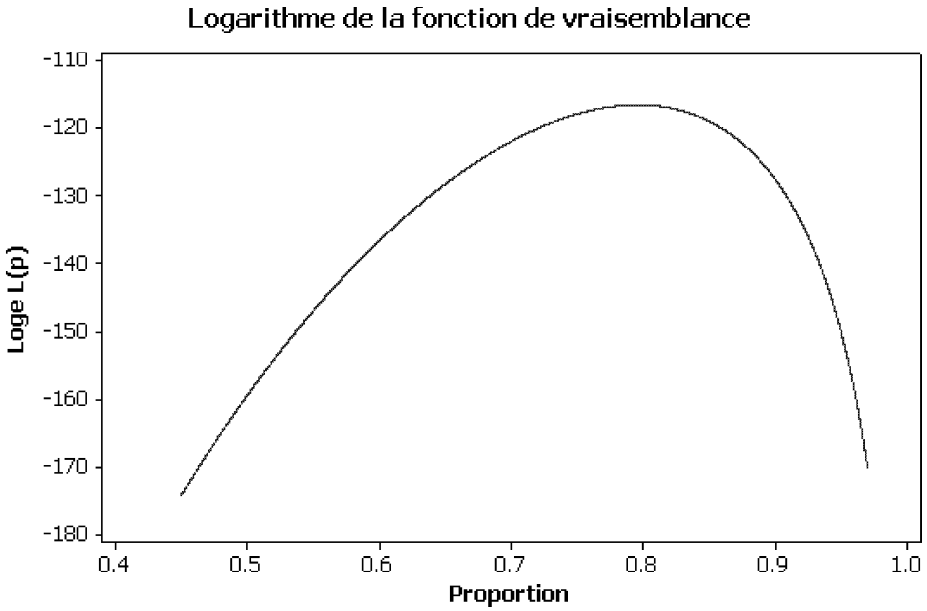

Figure 1 – Logarithme de la fonction de vraisemblance en fonction de la proportion d'arbres atteints d'un dépérissement faible à très fort.

X étant une variable binomiale de paramètres n et π.

On a :

$$\mathrm{E}(X) = n\,\pi \quad \text{et} \quad \mathrm{E}(n - X) = n(1 - \pi),$$

et il en résulte que :

$$i = n/[\pi(1 - \pi)].$$

L'erreur-standard de la proportion estimée est égale à :

$$e\,s(\hat{\pi}) = 1/\sqrt{i} = \sqrt{\pi(1 - \pi)/n}.$$

On retrouve bien la formule classique de l'écart-type d'une proportion qu'on peut d'ailleurs obtenir très simplement en considérant que la proportion est une transformation linéaire de la variable binomiale.

Une des propriétés des estimateurs du maximum de vraisemblance est que leur distribution d'échantillonnage est asymptotiquement normale. Cette propriété peut être utilisée pour construire un intervalle de confiance par la méthode de l'erreur-standard. Les limites de confiance sont données par la relation :

$$\hat{\pi} \pm u_{1-\alpha/2} \sqrt{\hat{\pi}(1-\hat{\pi})/n},$$

$u_{1-\alpha/2}$ étant le percentile $1-\alpha/2$ de la variable normale réduite. Pour un degré de confiance de 95 %, ce percentile est égal à 1,96.

Pour les données relatives aux chênes, on obtient :

$$0{,}7957 \pm 1{,}96 \sqrt{0{,}7957(1-0{,}7957)/230},$$

soit 0,7436 et 0,8478.

L'intervalle défini ci-dessus est appelé intervalle de WALD. Il correspond à toutes les valeurs π_0 telles que :

$$\frac{|\hat{\pi} - \pi_0|}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} < u_{1-\alpha/2},$$

le dénominateur de cette expression étant l'écart-type de la proportion estimée si $\pi = \hat{\pi}$.

Une autre solution consiste à remplacer ce dénominateur par l'écart-type de la proportion estimée si $\pi = \pi_0$. L'intervalle de confiance est alors constitué de toutes les valeurs π_0 telles que :

$$\frac{|\hat{\pi} - \pi_0|}{\sqrt{\pi_0(1-\pi_0)/n}} < u_{1-\alpha/2}.$$

Cette deuxième solution, qui est analytiquement plus compliquée, est appelée méthode du score.

Pour les données du dépérissement, le calcul de l'expression ci-dessus, en faisant varier π_0, conduit aux deux limites de confiance suivantes :

$$0{,}7389 \quad \text{et} \quad 0{,}8427.$$

Une troisième approche est basée sur le rapport de vraisemblance. Cette approche est plus complexe du point de vue des calculs mais est simple dans son principe : font partie de l'intervalle de confiance toutes les valeurs π_0 telles que :

$$2\left[\log_e L(\hat{\pi}) - \log_e L(\pi_0)\right] < \chi^2_{1-\alpha},$$

ou

$$\log_e L(\pi_0) > \log_e L(\hat{\pi}) - \tfrac{1}{2}\chi^2_{1-\alpha},$$

$\chi^2_{1-\alpha}$ étant le percentile $1-\alpha$ de la distribution χ^2 à 1 degré de liberté. Pour un degré de confiance de 95 %, ce percentile est égal à 3,84.

Pour l'échantillon observé, la valeur maximum du logarithme de la fonction de vraisemblance vaut $-116{,}465$. Les limites de confiance correspondent donc aux deux valeurs π_0 telles que :

$$\log_e L(\pi_0) = -116{,}465 - 3{,}84/2 = -118{,}385,$$

soit 0,7404 et 0,8443. Ces valeurs se déduisent de la figure 1, ou, si on souhaite davantage de précision, à partir des valeurs qui ont été calculées pour établir cette figure.

Pour tester l'hypothèse nulle :

$$H_0 : \pi = \pi_0,$$

contre l'alternative :

$$H_1 : \pi \neq \pi_0,$$

les trois approches ci-dessus peuvent être envisagées. Dans chaque cas, on rejette l'hypothèse nulle si la valeur π_0 ne se trouve pas dans l'intervalle de confiance.

Pour la méthode de WALD et la méthode du score, cela revient à calculer respectivement la statistique :

$$u_{obs} = \frac{|\hat{\pi} - \pi_0|}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \quad \text{ou} \quad u_{obs} = \frac{|\hat{\pi} - \pi_0|}{\sqrt{\pi_0(1-\pi_0)/n}},$$

et à rejeter l'hypothèse nulle si $u_{obs} > u_{1-\alpha/2}$. Le carré d'une variable normale réduite étant une variable χ^2 à 1 degré de liberté, on peut aussi rejeter l'hypothèse nulle si :

$$(u_{obs})^2 = \chi^2_{obs} > \chi^2_{1-\alpha}.$$

Pour l'exemple et pour $\pi_0 = 0{,}5$ on a, pour le test de WALD :

$$u_{obs} = \frac{0{,}7957 - 0{,}5}{\sqrt{0{,}7957(1-0{,}7957)/230}} = 11{,}12$$

et pour le test du score :

$$u_{obs} = \frac{0{,}7957 - 0{,}5}{\sqrt{0{,}50(1-0{,}5)/230}} = 8{,}97$$

et, dans les deux cas, on rejette indiscutablement l'hypothèse nulle.

Pour la méthode du rapport de vraisemblance, on calcule :

$$\chi^2_{obs} = -2\log_e\left[\frac{L(\pi_0)}{L(\hat{\pi})}\right] = -2\left[\log_e L(\pi_0) - \log_e L(\hat{\pi})\right]$$

et on rejette l'hypothèse nulle si $\chi^2_{obs} > \chi^2_{1-\alpha}$, la variable χ^2 ayant un degré de liberté.

Pour $\pi_0 = 0{,}5$ on a :

$$\log_e L(0{,}5) = \log_e(0{,}5^{230}) = -159{,}424$$

et on trouve :

$$\chi^2_{obs} = -2[-159{,}424 - (-116{,}465)] = 85{,}92$$

et on rejette l'hypothèse nulle car :

$$\chi^2_{obs} = 85{,}92 > \chi^2_{1-\alpha} = 3{,}84.$$

2.2. Test de WALD, test du score et test du rapport de vraisemblance en régression logistique.

Les trois méthodes qui ont été présentées dans le cas particulier du calcul des limites de confiance et du test de conformité d'une proportion peuvent être généralisées à d'autres situations. Elles peuvent notamment être utilisées pour l'inférence statistique multivariée.

La variable normale réduite et la variable χ^2 à un degré de liberté qui interviennent dans les cas d'un paramètre seront alors remplacées par des variables χ^2 dont le nombre de degrés de liberté est fonction du nombre de paramètres impliqués dans l'inférence statistique.

Les trois méthodes ont en commun d'exploiter la normalité asymptotique des estimateurs du maximum de vraisemblance. Pour certaines applications, elles conduisent au même résultat. Ainsi, en régression multiple ordinaire, les trois méthodes d'inférence sont équivalentes.

Pour d'autres applications, et notamment pour la régression logistique binaire, les différences entre les méthodes seront d'autant plus marquées que les échantillons sont de taille réduite et que les modèles contiennent plus de paramètres. Les discordances entre les méthodes résultent notamment du caractère dissymétrique de la fonction de vraisemblance.

La méthode de WALD est la plus couramment utilisée à cause de la simplicité de sa mise en oeuvre. Elle est proposée dans Minitab et SAS, comme nous le verrons au paragraphe 5. La méthode du rapport de vraisemblance est cependant préférable pour les problèmes de régression logistique. Elle est disponible en option dans SAS pour le calcul des intervalles de confiance des paramètres (paragraphe 5.2).

3. MODÈLE LOGIT POUR DONNÉES BINAIRES

3.1. Fonction de lien et modèle

Lorsque la variable à expliquer y possède deux modalités, codées par exemple $y = 1$ et $y = 2$ ou $y = A$ et $y = B$, l'objectif est de modéliser la probabilité d'appartenance à l'une des deux catégories, appelée *succès* ou *événement*[5], en fonction d'une ou plusieurs variables explicatives, $x_1, \ldots, x_p$. Cette probabilité d'appartenance est notée $\pi(\boldsymbol{x}_i)$, $\boldsymbol{x}_i$ étant le vecteur des valeurs prises par les variables explicatives pour un individu i.

Les probabilités $\pi(\boldsymbol{x}_i)$ évoluent cependant de manière non linéaire en fonction de $\boldsymbol{x}_i$ et la variance de ces probabilités varie avec $\boldsymbol{x}_i$. Pour cette raison, on effectue une transformation de la probabilité du succès $g[\pi(\boldsymbol{x}_i)]$. Cette transformation s'appelle fonction de lien. Pour alléger les notations, la probabilité de succès pour un individu sera fréquemment notée par la suite simplement π_i et la fonction de lien g_i.

Plusieurs fonctions de lien existent, mais la plus couramment utilisée est la fonction logit :

$$g_i = \text{logit}\,(\pi_i) = \log_e[\pi_i/(1-\pi_i)].$$

Le modèle de régression s'écrit alors :

$$g_i = \alpha + \boldsymbol{x}_i\,\boldsymbol{\beta},$$

où α et $\boldsymbol{\beta}$ sont des paramètres à estimer. La transformation inverse permet ensuite de retrouver les probabilités estimées en fonction de $\boldsymbol{x}_i$:

$$\pi_i = \exp(g_i)/[1+\exp(g_i)].$$

Cette expression est analogue à la densité de probabilité de la loi logistique, ce qui justifie le nom donné à la fonction de lien et, par extension, à ce type de régression. Cette fonction de lien est en effet la plus fréquemment retenue parce qu'elle conduit à une interprétation simple des coefficients de régression, par l'intermédiaire des *odds ratios*, qui seront présentés au paragraphe 5.2, mais aussi pour des raisons théoriques [COLLETT, 1979].

Les variables explicatives peuvent être quantitatives ou qualitatives. Les premières possèdent en général un grand nombre de valeurs différentes ; les secondes ont, au contraire, un nombre limité et connu *a priori* de modalités. Le modèle peut également contenir des termes d'interaction entre variables.

Si le modèle ne comporte que des variables quantitatives, le vecteur $\boldsymbol{\beta}$ comporte autant d'éléments que de variables, ou, exprimé autrement, à chaque variable quantitative correspond un coefficient. Pour une variable qualitative à q

5. En anglais : *success* ou *event*.

modalités, on a $q-1$ coefficients affectés à $q-1$ modalités. Pour la $q^{ième}$ modalité, qui constitue la modalité de référence, aucun coefficient n'est repris dans les sorties des logiciels, mais ce coefficient peut être obtenu à partir des $q-1$ autres coefficients. On notera cependant que les valeurs données aux coefficients de régression et à l'ordonnée à l'origine sont différentes mais équivalentes, pour le logiciel Minitab et le logiciel SAS, comme nous le verrons au paragraphe 5.2.

3.2. Profils

Une combinaison particulière des variables explicatives définit un profil et deux individus i et k ayant le même profil ont le même vecteur de variables explicatives.

En relation avec ces profils, différentes situations peuvent se rencontrer en pratique. Les individus peuvent tous avoir un profil différent. Les données sont alors nécessairement présentées sous forme non groupée : la matrice des variables explicatives et le vecteur de la variable à expliquer comportent n lignes, n étant le nombre d'individus. Cette situation se rencontre typiquement lorsqu'une ou plusieurs variables explicatives quantitatives sont présentes dans le modèle.

Lorsque plusieurs individus ont le même profil, les données peuvent être regroupées selon les J profils différents et les m_j observations relatives à un profil j se répartissent en y_j succès et $m_j - y_j$ échecs. Cette situation se rencontre typiquement lorsque le modèle de régression ne comporte que des variables explicatives qualitatives, mais il peut aussi se rencontrer avec des variables quantitatives, lorsque celles-ci ne peuvent prendre qu'un nombre limité de valeurs différentes, par exemple du fait de la précision des données. Ce cas diffère cependant du cas précédent car le nombre de profils n'est pas connu *a priori* mais a tendance à augmenter avec le nombre d'observations.

L'utilisateur devra être attentif à cette distinction entre données individuelles et données groupées lors de l'examen des résultats fournis par les logiciels statistiques : même si Minitab et SAS acceptent indifféremment les données individuelles ou les données groupées, des différences peuvent se présenter dans les résultats dans le cas de profils multiples, comme nous le verrons au paragraphe 6.

Pour bien marquer la différence entre données individuelles et données groupées, nous utilisons l'indice i pour les données individuelles ($i = 1, \ldots, n$) et l'indice j pour les données groupées ($j = 1, \ldots, J$).

4. FONCTION DE VRAISEMBLANCE POUR LA RÉGRESSION LOGISTIQUE BINAIRE

4.1. Fonction de vraisemblance et estimation des paramètres

Pour des données individuelles et en considérant que la variable explicative y est égale à l'unité en cas de succès et à zéro en cas d'échec, la fonction de vraisemblance s'écrit :

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right]$$

et son logarithme est égal à :

$$\log_e L(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log_e \pi_i + (1-y_i)\log_e(1-\pi_i)\right],$$

un des deux termes de l'expression étant systématiquement nul, puisque $y_i = 0$ ou $y_i = 1$, selon que l'individu i est caractérisé par un échec ou par un succès.

Pour les données groupées, on a :

$$L(\alpha, \boldsymbol{\beta}) = \prod_{j=1}^{J} \left[\pi_j^{y_j}\ (1-\pi_j)^{m_j - y_j}\right]$$

et $$\log_e L(\alpha, \boldsymbol{\beta}) = \sum_{j=1}^{J} [y_j\ \log_e\ \pi_j + (m_j - y_j)\ \log_e(1-\pi_j)],$$

y_j étant le nombre de succès pour le profil j et m_j le nombre d'individus présentant la modalité j.

On peut constater que, en présence de profils multiples, on obtient le même résultat, que les calculs soient réalisés à partir de données individuelles ou à partir de données groupées : la fonction de vraisemblance est donc invariante au regroupement des données qui ont le même profil.

On se rappellera que les π_i ou π_j sont fonction des $\boldsymbol{x}_i$ ou $\boldsymbol{x}_j$, par l'intermédiaire des coefficients de régression, ceux-ci étant déterminés de manière à maximiser la fonction $\log_e L(\alpha, \boldsymbol{\beta})$.

Des informations concernant cette maximisation et les éventuels problèmes de non-convergence sont données par Allison notamment [1999].

4.2. Vraisemblance pour le modèle nul

La fonction $\log_e L(\alpha, \boldsymbol{\beta})$ est donc une fonction des coefficients de régression. Elle peut être évaluée pour n'importe quelles valeurs de α et $\boldsymbol{\beta}$.

Outre les valeurs $\hat{\alpha}$ et $\hat{\boldsymbol{\beta}}$ qui maximisent cette fonction, un intérêt particulier concerne le cas $\boldsymbol{\beta} = \mathbf{0}$, qui correspond au modèle nul, c'est-à-dire au modèle ne faisant intervenir aucune variable explicative :

$$g_i = \hat{\alpha}.$$

Pour ce modèle nul, la proportion estimée $\hat{\pi}$ est donc constante, puisque le modèle ne présente pas de variables explicatives et cette constante est égale à la proportion de succès pour l'ensemble des n données (paragraphe 2.1) :

$$\hat{\pi} = n_1/n,$$

n_1 étant le nombre de succès et :

$$\hat{\alpha} = \text{logit}(\hat{\pi}) = \log_e[\hat{\pi}/(1-\hat{\pi})].$$

La valeur du logarithme de la fonction de vraisemblance de ce modèle nul vaut :

$$\begin{aligned} \log_e L(\hat{\alpha}) &= n_1 \ \log_e (n_1/n) + n_0 \ \log_e(n_0/n) \\ &= n_1 \ \log_e(n_1) + n_0 \log_e(n_0) - n \ \log_e(n), \end{aligned}$$

n_0 étant le nombre d'échecs :

$$n_0 = n - n_1.$$

Pour les données relatives au dépérissement du chêne, 183 arbres appartiennent à la catégorie « dépérissement faible à très fort », sur un total de 230 observations. Le logarithme de la vraisemblance pour le modèle nul vaut donc :

$$\log_e \ L(\hat{\alpha}) = 183 \ \log_e(183) + 47 \ \log_e(47) - 230 \ \log_e(230) = -116{,}465.$$

La vraisemblance du modèle nul est utilisée pour réaliser le test global de signification des variables d'un modèle, comme nous le verrons au paragraphe 5.1.

4.3. Vraisemblance du modèle saturé

Un modèle saturé est un modèle qui reproduit parfaitement les observations réalisées pour un ensemble de variables explicatives données.

Pour des données groupées, les π_j, pour le modèle saturé, sont donc égaux à y_j/m_j et le logarithme de la fonction de vraisemblance est égal à :

$$\log_e \ L(s) = \sum_{j=1}^{J} y_j \ \log_e \ \left(\frac{y_i}{m_j}\right) + (m_j - y_j) \ \log_e \left(\frac{m_j - y_j}{m_j}\right).$$

Pour les données relatives au chêne, considérons le modèle de régression logistique faisant intervenir comme seule variable explicative la région dans laquelle l'arbre a été observé. Le tableau 1 donne, pour chacune des deux régions,

Tableau 1 – Répartition des observations en fonction du degré de dépérissement et de la région naturelle.

Catégorie	Ardenne	Condroz	Totaux
Très faible	14	33	47
Faible à très fort	134	49	183
Totaux	148	82	230

le nombre d'arbres à dépérissement très faible et le nombre d'arbres à dépérissement faible à très fort.

Pour ce modèle, on n'a que deux profils (Ardenne et Condroz) et donc deux valeurs de probabilité de succès. Le modèle saturé est par conséquent un modèle qui redonne les deux proportions observées. Si $j = 1$ pour l'Ardenne et $j = 2$ pour le Condroz, on a :

$$\hat{\pi}_1 = 134/148 = 0{,}9054 \quad \text{et} \quad \hat{\pi}_2 = 49/82 = 0{,}5976.$$

Le logarithme de la fonction de vraisemblance est donné par :

$$\begin{aligned} \log_e L(s) \quad = \quad & 134 \ \log_e(134/148) + 14 \ \log_e(14/148) \\ & + 49 \ \log_e(49/82) + 33 \ \log_e(33/82) = -101{,}597. \end{aligned}$$

Nous reviendrons sur ce modèle au paragraphe 5.2.

Pour des données non groupées, le modèle saturé est tel que les n valeurs π_i sont égales à y_i : la probabilité de succès pour un individu est égale à l'unité si l'individu correspond à un succès et elle est nulle si l'individu correspond à un échec. Le logarithme de la fonction de vraisemblance est par conséquent égal à :

$$\log_e \ L(s) = \sum_{i=1}^{n} \left[y_i \ \log_e(y_i) + (1 - y_i) \ \log_e(1 - y_i) \right].$$

On constate que pour chaque individu, les deux termes sont systématiquement nuls car :

$$\begin{aligned} &\text{si} \quad y_i = 1 \quad \text{alors} \quad \log_e(y_i) = 0 \quad \text{et} \quad 1 - y_i = 0, \\ &\text{si} \quad y_i = 0 \quad \text{alors} \quad \log_e(1 - y_i) = 0. \end{aligned}$$

Il en résulte que $\log_e L(s)$ est toujours égal à zéro pour des données individuelles.

La vraisemblance du modèle saturé n'est donc pas invariante au regroupement des données ayant les mêmes profils. En effet, ce regroupement conduit à des termes qui ne seront plus systématiquement nuls. Cela peut se comprendre car le modèle saturé pour des données individuelles doit reproduire n proportions observées, alors que pour les données groupées il ne doit reproduire que J proportions.

La vraisemblance du modèle saturé intervient dans la définition de la déviance (paragraphe 6.1).

5. TESTS DE SIGNIFICATION DES COEFFICIENTS DE RÉGRESSION

5.1. Test global de signification des variables

Comme en régression ordinaire, on peut tester la nullité simultanée de tous les coefficients des variables intervenant dans le modèle, la nullité d'un coefficient particulier quand les autres variables sont présentes dans le modèle et, enfin, la nullité des coefficients d'un sous-ensemble de variables quand les autres variables sont présentes dans le modèle. Nous examinons d'abord la première situation, les deux autres feront l'objet des paragraphes 5.2 et 5.3.

La nullité simultanée de tous les coefficients de régression du modèle peut être testée par le rapport de vraisemblance, en calculant la quantité :

$$G = -2 \log_e \left[\frac{L(\hat{\alpha})}{L(\hat{\alpha}, \hat{\boldsymbol{\beta}})} \right] = 2 \left[\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) - \log_e L(\hat{\alpha}) \right].$$

Cette quantité G est basée sur la comparaison de la vraisemblance du modèle nul, c'est-à-dire du modèle ne faisant intervenir aucune variable explicative, et du modèle avec les variables explicatives. Si l'hypothèse nulle est vraie, la distribution de G est approximativement une distribution χ^2 à p degrés de liberté, p étant le nombre de paramètres dans le modèle, à l'exclusion de l'ordonnée à l'origine. On rejette l'hypothèse de nullité simultanée de tous les coefficients des variables du modèle si $G > \chi^2_{1-\alpha}$.

Les deux autres méthodes présentées au paragraphe 2 (test de WALD et test basé sur le score), peuvent aussi être utilisées. Elles conduisent également à une valeur χ^2_{obs} provenant d'une distribution χ^2 à p degrés de liberté et au rejet de l'hypothèse de nullité des coefficients si cette valeur χ^2_{obs} est supérieure à $\chi^2_{1-\alpha}$. Pour des échantillons de taille faible ou modérée, le test du rapport de vraisemblance est cependant préférable.

A titre d'illustration, les données relatives au dépérissement ont été exprimées en fonction de l'altitude et de la région naturelle, en utilisant, d'une part, la commande BLOGISTIC de Minitab et, d'autre part, la procédure PROC LOGISTIC de SAS. Le code relatif à ces exécutions est donné en annexe. Une partie des résultats obtenus est reprise dans les figures 2 et 3.

```
Log-Likelihood = -99.804
Test that all slopes are zero: G = 33.322, DF = 2, P-Value = 0.000

Logistic Regression Table
                                                    Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant   -0.808214   0.690910  -1.17  0.242
Altitude   0.0049716  0.0027065   1.84  0.066   1.00   1.00   1.01
Region
 Ardenne     1.22693   0.473007   2.59  0.009   3.41   1.35   8.62
```

Figure 2 – Paramètres estimés et tests de signification des variables *Altitude* et *Région* (Minitab).

La première partie de ces figures concerne les tests globaux de signification simultanée des deux coefficients de régression. L'autre partie sera discutée au point suivant.

Pour le modèle ajusté, la figure 2 donne le logarithme de la vraisemblance :

$$\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = -99{,}804.$$

D'autre part, nous avons vu que le logarithme de la vraisemblance pour le modèle nul est égal à (paragraphe 4.2) :

$$\log_e L(\hat{\alpha}) = -116{,}465.$$

On a donc :

$$G = 2[-99{,}804 - (-116{,}465)] = 33{,}322,$$

ce qui correspond bien à la valeur donnée par Minitab.

La probabilité associée à cette valeur observée d'une variable χ^2 à 2 degrés de liberté est très faible et, par conséquent, on rejette l'hypothèse de nullité simultanée des deux coefficients de régression. Cela signifie concrètement qu'il est opportun d'utiliser soit les deux variables soit une des deux variables pour modéliser le dépérissement. La solution à retenir sera envisagée au point suivant.

SAS donne également les résultats de ce test, mais reprend aussi les résultats des tests du score et de WALD (figure 3). On constate qu'il y a peu de différence entre les valeurs numériques pour le test du score ($\chi^2_{obs} = 32{,}895$) et le test du rapport de vraisemblance ($\chi^2_{obs} = 33{,}322$), mais que la valeur du test de WALD est plus différente ($\chi^2_{obs} = 27{,}563$). Pour cet exemple, la conclusion pratique est la même pour les trois tests : on rejette indiscutablement l'hypothèse de nullité simultanée des deux coefficients.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.3218	2	<.0001
Score	32.8947	2	<.0001
Wald	27.5627	2	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1947	0.8247	0.0558	0.8133
Altitude		1	0.00497	0.00271	3.3743	0.0662
Region	Ardenne	1	0.6135	0.2365	6.7283	0.0095

Profile Likelihood Confidence Interval for Parameters

Parameter		Estimate	95% Confidence Limits	
Intercept		-0.1947	-1.8242	1.4202
Altitude		0.00497	-0.00017	0.0105
Region	Ardenne	0.6135	0.1633	1.0965

Wald Confidence Interval for Parameters

Parameter		Estimate	95% Confidence Limits	
Intercept		-0.1947	-1.8110	1.4215
Altitude		0.00497	-0.00033	0.0103
Region	Ardenne	0.6135	0.1499	1.0770

Profile Likelihood Confidence Interval for Odds Ratio

Effect		Unit	Estimate	95% Confidence Limits	
Altitude		1.0000	1.005	1.000	1.011
Region	Ardenne vs Condroz	1.0000	3.411	1.386	8.961

Wald Confidence Interval for Odds Ratios

Effect		Unit	Estimate	95% Confidence Limits	
Altitude		1.0000	1.005	1.000	1.010
Region	Ardenne vs Condroz	1.0000	3.411	1.350	8.619

Figure 3 – Paramètres estimés et tests de signification des variables *Altitude* et *Région* (SAS).

5.2. Test de signification d'un coefficient particulier

Le test de signification d'une variable explicative quantitative particulière x_j peut se faire par le test de WALD en calculant la valeur :

$$u_{obs} = \frac{|\hat{\beta}_j|}{e\, s(\hat{\beta}_j)},$$

et en rejetant, pour un test bilatéral, l'hypothèse de nullité de β_j si $u_{obs} \geq u_{1-\alpha/2}$ ou encore si $u^2_{obs} \geq \chi^2_{1-\alpha}$, la variable χ^2 ayant un degré de liberté. Dans cette relation, $e\, s(\hat{\beta}_j)$ est l'erreur-standard du coefficient de régression estimé $\hat{\beta}_j$ de la variable x_j.

Le test peut également être réalisé par le rapport de vraisemblance, en déterminant la quantité :

$$\chi^2_{obs} = -2 \log_e \left[\frac{L(\hat{\alpha}, \hat{\boldsymbol{\beta}}_{réd})}{L(\hat{\alpha}, \hat{\boldsymbol{\beta}})} \right] = 2 \left[\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) - \log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}_{réd}) \right].$$

Dans cette relation, $L(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ est la vraisemblance pour le modèle complet et $L(\hat{\alpha}, \hat{\boldsymbol{\beta}}_{réd})$ la vraisemblance pour le modèle dont on a éliminé la variable x_j. On rejette l'hypothèse de nullité de β_j si $\chi^2_{obs} \geq \chi^2_{1-\alpha}$, la variable χ^2 ayant un degré de liberté.

Pour une variable explicative qualitative à deux modalités, le test de signification de la variable se fait comme décrit ci-dessus, l'introduction d'une telle variable dans le modèle conduisant à l'estimation d'un seul paramètre supplémentaire.

Une variable explicative qualitative à q modalités conduit, par contre, à $q-1$ paramètres supplémentaires. Le test de signification de chacun de ces paramètres peut se faire comme ci-dessus. Dans ce cas, le rejet de l'hypothèse de nullité du coefficient pour une modalité donnée signifie que la modalité en question est significativement différente de la modalité de référence.

Si on souhaite tester globalement l'effet de la variable qualitative, on teste la nullité simultanée des coefficients des $q-1$ modalités. Il s'agit alors d'un test de nullité d'un groupe de coefficients. Ce problème sera abordé au paragraphe 5.3.

Minitab donne uniquement les tests de WALD. Dans la deuxième partie de la figure 2, la colonne intitulée z reprend les valeurs u_{obs}, soit 1,84 pour l'altitude et 2,59 pour la région. Les probabilités associées sont égales à 0,066 et 0,009. On rejette donc l'hypothèse de nullité de l'effet de l'altitude quand on prend en considération la région, mais on ne rejette pas l'hypothèse de nullité de l'effet région quand on prend en considération l'altitude. Pratiquement, cela signifie que l'altitude peut être éliminée du modèle.

Minitab ne donne pas les valeurs χ^2 des tests du rapport de vraisemblance ni du score. Les valeurs relatives aux premiers tests peuvent cependant s'obtenir en ajustant plusieurs modèles de régression. Pour un coefficient β_j donné, la différence entre les valeurs G des modèles avec et sans la variable x_j correspond au χ^2_{obs}.

Ainsi par exemple, pour tester la signification de l'altitude, la différence de valeur G du modèle à deux variables et du modèle avec comme seule variable la région est égale à :

$$33{,}322 - 29{,}736 = 3{,}586.$$

La valeur 33,322 est donnée dans la figure 2 et la valeur 29,736 peut se retrouver à partir de la vraisemblance du modèle nul (paragraphe 4.2) et de la vraisemblance du modèle saturé quand on ne prend en considération que la région (paragraphe 4.3) :

$$2[-101{,}597 - (-116{,}465)] = 29{,}736.$$

Quand on ne dispose que de la variable région, le maximum du logarithme de la vraisemblance est en effet égal au logarithme de la vraisemblance du modèle saturé car le modèle avec la variable région donne des probabilités estimées par région qui sont égales aux proportions estimées.

A titre de vérification, la figure 4 donne les résultats de l'ajustement du modèle lorsqu'on prend en compte uniquement la variable région.

La procédure LOGISTIC de SAS donne également les résultats des test de WALD, les valeurs χ^2_{obs} de WALD étant égales aux carrés des valeurs z données par Minitab, ainsi que les limites de confiance des paramètres estimés (figure 3). Elle donne aussi les limites de confiance calculées par le rapport de vraisemblance, ce qui permet de tester la signification des variables par cette méthode : un coefficient de régression est significatif si ses deux limites de confiance sont du même signe. On conclut donc, comme pour la méthode de WALD, que l'altitude n'est pas significative mais que la région est, par contre, significative.

```
Log-Likelihood = -101.597
Test that all slopes are zero: G = 29.736, DF = 1, P-Value = 0.000

Logistic Regression Table
                                                  Odds     95% CI
Predictor      Coef   SE Coef     Z      P  Ratio  Lower  Upper
Constant   0.395313  0.225191  1.76  0.079
Region
 Ardenne    1.86347  0.359991  5.18  0.000   6.45   3.18  13.05
```

Figure 4 – Paramètres estimés et signification de la variable *Région* (Minitab).

A propos des paramètres estimés, on notera que, en présence de variables qualitatives, les résultats donnés par Minitab et par SAS diffèrent. Cela est dû à la valeur attribuée arbitrairement à la modalité non reprise dans la liste des coefficients. Minitab considère que ce coefficient est nul alors que SAS considère que ce coefficient est tel que la somme des coefficients pour toutes les modalités est nulle. Si on développe les équations pour l'Ardenne et pour le Condroz, on obtient les mêmes résultats. En effet, pour Minitab on a, respectivement pour l'Ardenne et le Condroz (figure 2) :

$$\text{logit}\,(\pi_i) = -0{,}8082 + 1{,}2269 + 0{,}00497\ \mathit{Altitude}$$

et $$\text{logit}\,(\pi_i) = -0{,}8082 + 0{,}00497\ \mathit{Altitude}.$$

Pour SAS on a (figure 3) :

$$\text{logit}\,(\pi_i) = -0{,}1947 + 0{,}6135 + 0{,}00497\ \mathit{Altitude}$$

et $$\text{logit}\,(\pi_i) = -0{,}1947 - 0{,}6135 + 0{,}00497\ \mathit{Altitude}.$$

Soit, dans les deux cas :

$$\text{logit}\,(\pi_i) = 0{,}4187 + 0{,}00497\ \mathit{Altitude}$$

et $$\text{logit}\,(\pi_i) = -0{,}8082 + 0{,}00497\ \mathit{Altitude}.$$

En plus des informations relatives aux coefficients, les deux logiciels fournissent des estimations et les limites de confiance des *odds ratios*. Pour une variable quantitative, l'*odds ratio* est le rapport suivant :

$$\frac{\pi(x)/[1-\pi(x)}{\pi(x+1)/[1-\pi(x+1)]}.$$

Il s'agit du rapport des rapports probabilité de succès/probabilité d'échec pour une augmentation d'une unité de la variable explicative. Cet *odds ratio* est directement lié au coefficient de régression $\hat{\beta}_j$ de la variable explicative en question : l'*odds ratio* est égal à $e^{\hat{\beta}_j}$.

Pour l'altitude, on a :

$$e^{0{,}00497} = 1{,}005.$$

Les limites de confiance de ce rapport sont obtenues en remplaçant $\hat{\beta}_j$ dans la formule ci-dessus successivement par la limite de confiance inférieure et supérieure de $\hat{\beta}_j$.

Pour une variable qualitative, l'*odds ratio* pour une modalité donnée est le rapport des rapports probabilité de succès/probabilité d'échec pour la modalité considérée et pour la modalité de référence. Il est égal à $e^{(\hat{\beta_k}-\hat{\beta_r})}$, $\hat{\beta}_k$ et $\hat{\beta}_r$ étant les coefficients pour la modalité k et la modalité de référence de la variable explicative en question.

Pour l'Ardenne on a, avec Minitab et avec SAS :

$$e^{1,2269-0} = e^{0,6135-(-0,6135)} = 3{,}411.$$

Le rapport de probabilité de dépérissement faible à très fort/probabilité de dépérissement très faible est donc 3,41 fois plus grand en Ardenne que dans le Condroz, à égalité d'altitude.

5.3. Test de signification d'un sous-ensemble de variables

La nullité simultanée des coefficients d'un groupe de k variables repose sur le calcul de :

$$\chi^2_{obs} = -2\log_e\left[\frac{L(\hat{\alpha},\hat{\boldsymbol{\beta}}_{réd})}{L(\hat{\alpha},\hat{\boldsymbol{\beta}})}\right] = 2\left[\log_e L(\hat{\alpha},\hat{\boldsymbol{\beta}}) - \log_e L(\hat{\alpha},\hat{\boldsymbol{\beta}}_{réd})\right],$$

$L(\hat{\alpha},\hat{\boldsymbol{\beta}}_{réd})$ et $L(\hat{\alpha},\boldsymbol{\beta})$ étant respectivement la valeur de la fonction de vraisemblance pour le modèle estimé en l'absence des k variables et pour le modèle complet. Si l'hypothèse nulle est vraie, la quantité χ^2_{obs} suit approximativement une distribution χ^2 à k degrés de liberté. On rejette donc l'hypothèse de nullité simultanée des k coefficients de régression si χ^2_{obs} est supérieur au percentile $1-\alpha$ de la variable χ^2 à k degrés de liberté.

Pratiquement, le calcul de G peut se faire par l'ajustement du modèle complet et du modèle avec omission des k variables. Soit G et $G_{réd}$ les statistiques associées au test global de signification de tous les coefficients de ces deux modèles (paragraphe 5.1), la valeur χ^2_{obs} est égale à la différence entre ces deux statistiques. En effet :

$$\begin{aligned} G - G_{réd} &= 2\left[\log_e L(\hat{\alpha},\hat{\boldsymbol{\beta}}) - \log_e L(\hat{\alpha})\right] - 2\left[\log_e L(\hat{\alpha},\hat{\boldsymbol{\beta}}_{réd}) - \log_e L(\hat{\alpha})\right] \\ &= 2\left[\log_e L(\hat{\alpha},\hat{\boldsymbol{\beta}}) - \log_e L(\hat{\alpha},\hat{\boldsymbol{\beta}}_{réd})\right]. \end{aligned}$$

Ce test peut notamment être utilisé pour tester la nullité simultanée des $q-1$ coefficients de régression relatifs à une variable qualitative à q modalités, c'est-à-dire pour tester la signification de cette variable qualitative.

6. CRITÈRES GLOBAUX D'AJUSTEMENT

6.1. La déviance et les résidus au sens de la déviance

Différents critères permettent de comparer plusieurs modèles relatifs à une même variable à expliquer. Ils sont liés à la fonction de vraisemblance et plus

précisément aux trois valeurs particulières du logarithme de la fonction de vraisemblance : valeur du maximum, valeur pour le modèle nul et valeur pour le modèle saturé.

Le premier de ces critères est la déviance D, définie par la relation suivante :

$$D = -2 \log_e \left[\frac{L(\hat{\alpha}, \hat{\boldsymbol{\beta}})}{L(s)} \right] = 2 \left[\log_e L(s) - \log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \right].$$

La vraisemblance pour le modèle saturé $L(s)$ étant toujours supérieure ou égale à la vraisemblance pour le modèle ajusté $L(\hat{\alpha}, \hat{\boldsymbol{\beta}})$, la déviance sera toujours nulle ou positive. Elle joue, en régression logistique, le même rôle que la somme des carrés des écarts résiduelle en régression classique : elle est d'autant plus faible que l'ajustement est bon, c'est-à-dire que l'ajustement conduit à des probabilités estimées proches des proportions observées.

Lorsque les données ne comportent qu'un nombre limité de profils différents et définis *a priori*, par exemple lorsque les variables explicatives sont qualitatives, la déviance possède une distribution qui tend vers une variable χ^2 lorsque le nombre d'observations tend vers l'infini. Le nombre de degrés de liberté associé à cette variable χ^2 est égal à $J - p$, p étant le nombre de paramètres dans le modèle. Il est alors possible de tester l'adéquation du modèle, en comparant la déviance au percentile $1 - \alpha$ de la distribution χ^2 à $J - p$ degrés de liberté : on considère que le modèle est inadéquat si la déviance est supérieure à $\chi^2_{1-\alpha}$. Il s'agit en fait d'un test par la méthode du rapport de vraisemblance de la conformité du modèle ajusté au modèle saturé (paragraphe 4.3).

La déviance est également liée aux résidus au sens de la déviance, qui se définissent comme suit, pour des données groupées :

$$d_j = \pm \left\{ 2 \left[y_j \log_e \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \log_e \left(\frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)} \right) \right] \right\}^{1/2},$$

le signe étant le signe de $y_j - m_j \hat{\pi}_j$. Pour $y_j = 0$ et $y_j = m_j$, la relation se simplifie :

$$d_j = -\sqrt{2 m_j |\log_e (1 - \hat{\pi}_j)|} \quad \text{si } y_j = 0$$

et

$$d_j = \sqrt{2 m_j |\log_e \hat{\pi}_j|} \qquad \text{si } y_j = m_j.$$

La somme des carrés de ces résidus donne la déviance :

$$D = \sum_{j=1}^{J} d_j^2.$$

La formule générale donnant d_j, pour un profil j donné, montre que :

$$d_j^2 = 2\left[O_{j1}\log_e(O_{j1}/E_{j1}) + O_{j2}\log_e(O_{j2}/E_{j2})\right],$$

O_{j1} et O_{j2} étant respectivement les fréquences observées de succès et d'échec pour le profil j, E_{j1} et E_{j2} étant des fréquences attendues correspondantes. La déviance, qui est la somme des d_j^2 sur les J profils, peut donc s'écrire, de manière synthétique :

$$D = 2\sum_{j=1}^{J}\sum_{k=1}^{2} O_{jk}\log_e(O_{jk}/E_{jk}),$$

la somme étant étendue aux $2J$ cellules du tableau croisant les profils et les deux modalités de la variable à expliquer et O_{jk} et E_{jk} étant respectivement les fréquences observées et attendues de chacune de ces $2J$ cellules.

Pour les données individuelles, le logarithme de la vraisemblance du modèle saturé est toujours égal à zéro (paragraphe 4.3) et on a :

$$D = -2\log L(\hat{\alpha}, \hat{\boldsymbol{\beta}}).$$

Les résidus au sens de la déviance s'écrivent, dans ce cas :

$$d_i = -\sqrt{2|\log_e(1-\hat{\pi}_j)|} \quad \text{si } y_i = 0$$

et

$$d_i = \sqrt{2|\log_e(\hat{\pi}_j)|} \quad \text{si } y_i = 1,$$

et on a :

$$D = \sum_{i=1}^{n} d_i^2.$$

Pour des données individuelles ou pour des données groupées sur la base de variables explicatives continues ou à peu près continues, la déviance ne possède pas asymptotiquement une distribution χ^2. Le test χ^2 ne peut donc pas être réalisé, mais la déviance reste un paramètre utile pour comparer différents modèles construits pour une même variable à expliquer.

La figure 5, extraite des résultats fournis par Minitab, donne la déviance ainsi que deux autres paramètres qui seront présentés aux paragraphes 6.2 et 6.3. La valeur de la déviance est de 73,3, le nombre de degrés de liberté est de 54 et la probabilité correspondante est de 0,041. Cette probabilité n'a cependant aucune signification pratique, car la déviance n'a pas une distribution χ^2 à 54 degrés de liberté. Le modèle contient, en effet, une variable quantitative - l'altitude - qui présente des valeurs identiques pour de nombreux individus. Le nombre de profils différents est de 57 alors qu'on dispose de 230 observations. Celui-ci

```
Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             66.6504  54  0.116
Deviance            73.3001  54  0.041
Hosmer-Lemeshow     15.9078   8  0.044
```

Figure 5 – Critères globaux d'ajustement pour le modèle avec *Altitude* et *Région* comme variables explicatives (Minitab).

n'était cependant pas fixé *a priori* avant la collecte des données et la déviance ne possède pas une distribution pouvant être approchée par une variable χ^2.

Pour le modèle ne faisant intervenir que la variable région, Minitab ne donnerait pas la déviance mais signalerait que le modèle utilise tous les degrés de liberté. En effet, la variable explicative donne lieu uniquement à deux profils différents et le modèle possède deux paramètres. Il en résulte que la vraisemblance liée au modèle se confond avec la vraisemblance du modèle saturé et que la déviance est nulle.

Pour des données groupées selon des variables qualitatives, une option de SAS permet d'obtenir la déviance qui, pour le modèle avec la région comme seule variable explicative est bien égale à zéro.

Lorsque les données sont présentées sous la forme de données individuelles, mais qu'elles présentent des profils multiples, comme dans le cas du dépérissement des chênes, les résidus calculés par Minitab et par SAS diffèrent. En effet, SAS calcule les n résidus en considérant les données individuelles et des profils identiques donnent des résidus identiques. Par contre, Minitab regroupe automatiquement les données à profils identiques et détermine un seul résidu par profil. Ce résidu est repris en regard du premier individu présentant ce profil et des données manquantes sont indiquées pour les autres individus ayant le même profil. Pour les données relatives au dépérissement, SAS donne donc 230 résidus mais Minitab ne calcule que 57 résidus et donne 173 données manquantes.

Cette façon de procéder est d'ailleurs la même pour les résidus de PEARSON dont il est question ci-dessous.

6.2. La somme des carrés des résidus au sens de PEARSON

Pour des données groupées, ces résidus sont définis par la relation suivante :

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}.$$

Il s'agit donc des écarts entre les fréquences observées de succès et les fréquences attendues, ces écarts étant divisés par l'écart-type des fréquences observées.

Pour des données et des profils fixés *a priori*, la somme des carrés des résidus de PEARSON tend vers une distribution χ^2 de PEARSON à $J-p$ degrés de liberté. Comme la déviance, cette somme de carrés des résidus peut servir à tester l'adéquation du modèle : on considère que le modèle est inadéquat si cette somme est supérieure à $\chi^2_{1-\alpha}$. On peut montrer que la somme des carrés des résidus est égale à la statistique χ^2 calculée à partir des différences entre les fréquences observées et les fréquences attendues d'un tableau à deux entrées, croisant les J profils et les deux modalités de la variable à expliquer :

$$\sum_{j=1}^{J} r_j^2 = \chi_P^2 = \sum_{j=1}^{J} \left[\frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j} + \frac{[(m_j - y_j) - m_j(1-\hat{\pi}_j)]^2}{m_j(1-\hat{\pi}_j)} \right],$$

ou encore, de façon plus synthétique :

$$\chi_P^2 = \sum_{j=1}^{J}\sum_{k=1}^{2}(O_{jk} - E_{jk})^2/E_{jk},$$

O_{jk} étant la fréquence observée et E_{jk} la fréquence attendue de la cellule jk.

Pour l'exemple numérique, la somme des carrés des résidus de PEARSON donnée par Minitab (figure 5) est égale à 66,7, mais, comme pour la déviance, cette valeur ne peut pas être comparée à un pourcentile théorique d'une variable χ^2 à 54 degrés de liberté, le nombre de profils différents n'étant pas fixé *a priori*.

6.3. Le test de HOSMER et LEMESHOW

Ce test basé sur le regroupement des observations en un nombre limité g de groupes en fonction de la probabilité estimée $\hat{\pi}_i$ ou $\hat{\pi}_j$. Le nombre de groupes est typiquement de l'ordre de la dizaine et les groupes ont, autant que possible, des effectifs identiques. Des variations dans les effectifs peuvent cependant se présenter, d'une part, lorsque le nombre total d'observations n'est pas un multiple de g et d'autre part, du fait que des observations ayant même profil et donc même probabilité estimée sont affectées à un même groupe.

Pour chaque groupe, on détermine la moyenne des probabilités estimées $\bar{\pi}_k$ des observations du groupe et on calcule les fréquences attendues pour le succès et l'échec. Ensuite, on établit le tableau à deux entrées croisant les groupes et les deux modalités de la variable à expliquer. On calcule enfin la statistique :

$$\chi_{HL}^2 = \sum_{k=1}^{g}\sum_{l=1}^{2}\left(O_{kl} - E_{kl}\right)^2/E_{kl}.$$

Dans cette relation, O_{kl} est la fréquence du groupe kl et E_{kl} est la fréquence attendue correspondante. Si le modèle est adéquat, la statistique χ^2_{HL}

suit approximativement une distribution χ^2 à $g-2$ degrés de liberté. On rejette par conséquent l'hypothèse d'adéquation du modèle si $\chi^2_{HL} > \chi^2_{1-\alpha}$.

La statistique de HOSMER et LEMESHOW présente clairement une analogie avec la somme des carrés des résidus de PEARSON puisque dans les deux cas on compare des fréquences observées et des fréquences attendues d'un tableau de fréquences à deux entrées. L'intérêt du test de HOSMER et LEMESHOW est qu'il peut être utilisé dans les situations où la somme des carrés des résidus de PEARSON ne suit pas une distribution χ^2, comme dans le cas de données individuelles.

La valeur donnée par Minitab pour ce test est de 15,9, le nombre de classes étant égal à dix donnant lieu à 8 degrés de liberté (figure 5).On conclut que le modèle n'est pas adéquat, les écarts entre les fréquences observées et les fréquences attendues étant trop importants.

La figure 6 donne le résultat obtenu avec SAS. Le nombre de groupes est égal à 9 et la valeur χ^2_{HL}, égale à 15,9 avec 7 degrés de liberté, conduit aussi à la conclusion que le modèle n'est pas adéquat. La figure 6 reprend également le tableau des fréquences observées et des fréquences attendues des groupes. Un tableau similaire est aussi donné par Minitab, mais n'est pas repris ici.

Partition for the Hosmer and Lemeshow Test

		Dep_binaire = F_Fort		Dep_binaire = T_faible	
Group	Total	Observed	Expected	Observed	Expected
1	24	16	12.66	8	11.34
2	23	13	13.59	10	9.41
3	27	14	17.37	13	9.63
4	23	21	17.79	2	5.21
5	26	20	22.62	6	3.38
6	25	20	22.47	5	2.53
7	26	26	23.81	0	2.19
8	25	22	23.10	3	1.90
9	31	31	29.58	0	1.42

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
15.9053	7	0.0260

Figure 6 – Test de HOSMER et LEMESHOW pour le modèle avec *Altitude* et *Région* comme variables explicatives (SAS).

6.4. Le critère d'AKAIKE et le critère de SCHWARTZ

Le critère AIC d'AKAIKE [6] et le critère SC de SCHWARTZ [7] sont donnés par les relations suivantes :

$$\begin{aligned} AIC &= -2\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) + 2k \\ SC &= -2\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) + k\log_e n, \end{aligned}$$

k étant le nombre de paramètres dans le modèle et n le nombre d'observations.

Ces deux critères pénalisent la vraisemblance si plus de paramètres sont estimés, la pénalisation étant plus sévère pour le critère de SCHWARTZ. Ils sont utiles pour la comparaison de modèles construits pour la même variable à expliquer, les valeurs les plus faibles correspondant à des modèles préférables. Ils sont donnés par SAS (figure 7) mais pas par Minitab.

6.5. Les coefficients R^2

En régression ordinaire, le coefficient de détermination multiple est couramment utilisé pour quantifier l'aptitude qu'ont les variables explicatives à prédire la variable à expliquer. Différentes mesures ont également été proposées dans ce but pour la régression logistique.

```
 Model Fit Statistics

                              Intercept
                  Intercept         and
Criterion            Only  Covariates

AIC                234.931     205.609
SC                 238.369     215.923
-2 Log L           232.931     199.609


R-Square    0.1349    Max-rescaled R-Square    0.2118
```

Figure 7 – Critères d'AKAIKE et de SCHWARTZ et valeurs R^2 pour le modèle avec *Altitude* et *Région* comme variables explicatives (SAS).

6. En anglais : AKAIKE*'s information criterion.*
7. En anglais : SCHWARTZ *criterion, Bayesian information criterion, BIC.*

Un premier coefficient est donné par le rapport suivant :

$$R^2 = \frac{\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) - \log_e L(\hat{\alpha})}{\log_e L(s) - \log_e L(\hat{\alpha})}.$$

Il s'agit de la mesure de l'amélioration de l'ajustement par la prise en compte des variables en proportion de l'amélioration maximum possible. Le coefficient varie par conséquent de 0 à 1. Il vaut 0 lorsque les variables utilisées n'apportent aucune information et il vaut 1 lorsque la prédiction est parfaite.

Dans le cas de données individuelles, $\log_e L(s)$ est nul et le coefficient s'écrit :

$$R^2 = \frac{\log_e L(\hat{\alpha}) - \log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}})}{\log_e L(\hat{\alpha})} = 1 - \frac{\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}})}{\log_e L(\hat{\alpha})}.$$

Une autre solution est basée sur la statistique G, calculée lors du test global de signification de l'ensemble des coefficients de régression du modèle (paragraphe 5.1) :

$$G = -2 \log_e \left[\frac{L(\hat{\alpha})}{L(\hat{\alpha}, \hat{\boldsymbol{\beta}})} \right].$$

Le coefficient R^2 généralisé s'écrit en effet :

$$R^2_G = 1 - \exp(-G/n) = 1 - \left[\frac{L(\hat{\alpha})}{L(\hat{\alpha}, \hat{\boldsymbol{\beta}})} \right]^{2/n}.$$

La justification est qu'en régression ordinaire, la valeur G est liée au coefficient de détermination multiple par la relation ci-dessus.

Un inconvénient de ce paramètre est que la borne supérieur du domaine de variation est égale à :

$$R^2_{G\,max} = 1 - L(\hat{\alpha})^{2/n},$$

puisqu'un ajustement parfait correspond au modèle saturé pour lequel on a, pour des données individuelles :

$$\log_e L(s) = 0 \quad \text{ou} \quad L(s) = 1.$$

C'est la raison pour laquelle un coefficient corrigé [8] a été proposé :

$$R^2_{G\,aj} = \frac{R^2_G}{1 - L(\hat{\alpha})^{2/n}}.$$

8. En anglais : *Maximum rescaled R square.*

Une discussion de ces différents coefficients est donnée par MENARD [2000].

Pour les données relatives au dépérissement, on a (paragraphe 5.1 et figure 2) :

$$\log_e L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = -99{,}804, \quad \log_e L(\hat{\alpha}) = -116{,}47 \quad \text{et} \quad G = 33{,}322.$$

On en déduit :

$$\begin{aligned} R^2 &= 1 - \frac{-99{,}804}{-116{,}465} = 0{,}1431, \\ R^2_G &= 1 - \exp\left(-\frac{33{,}322}{230}\right) = 0{,}1349 \\ R^2_{G\,aj} &= \frac{0{,}1349}{1 - [\exp(-116{,}465)]^{2/230}} = 0{,}2118. \end{aligned}$$

Auncun de ces coefficients n'est donné par Minitab mais les deux derniers peuvent être obtenus avec SAS (figure 7).

6.6. Les mesures d'association

Différents paramètres sont définis afin de quantifier l'association entre les probabilités estimées et l'appartenance aux deux catégories.

Considérons que les n observations sont divisées en deux groupes selon la variable à expliquer. Les observations du premier groupe correspondent aux individus caractérisés par le succès et sont identifiées par leur numéro d'ordre désigné par $i (i = 1, \ldots, n_1)$; les observations du deuxième groupe correspondent aux individus caractérisés par l'échec et sont identifiées par le numéro d'ordre désigné par $j (j = 1, \ldots, n_0)$.

Soit un couple particulier ij et soit $\hat{\pi}_i$ et $\hat{\pi}_j$ les probabilités estimées d'appartenir au premier groupe, respectivement pour l'individu i et j. Pour ce couple d'observations, trois situations peuvent se présenter. Il y a :

$$\begin{aligned} \text{concordance} \quad &\text{si} \quad \hat{\pi}_i > \hat{\pi}_j, \\ \text{discordance} \quad &\text{si} \quad \hat{\pi}_i < \hat{\pi}_j, \\ \text{ex-aequo} \quad &\text{si} \quad \hat{\pi}_i = \hat{\pi}_j. \end{aligned}$$

Au total, on peut définir $n_1 n_0$ couples ij et déterminer le nombre n_c de concordances, le nombre n_d de discordances et le nombre n_e d'ex-aequos. A partir de ces nombres, on définit les mesures d'association suivantes :

```
 Association of Predicted Probabilities and Observed Responses

Percent Concordant     75.0    Somers' D    0.519
Percent Discordant     23.1    Gamma        0.530
Percent Tied            2.0    Tau-a        0.170
Pairs                  8601    c            0.760
```

Figure 8 – Mesures d'association pour le modèle avec *Altitude* et *Région* comme variables explicatives (SAS).

le coefficient D de SOMERS : $D = (n_c - n_d) / (n_1 n_0)$,

le coefficient γ de GOODMAN et KRUSKAL : $\gamma = (n_c - n_d) / (n_c + n_d)$,

le coefficient τ de KENDALL : $\tau = (n_c - n_d) / [0{,}5n(n-1)]$,

le coefficient c : $c = (n_c + 0{,}5n_e)/n_1 n_0$.

Les quatre coefficients sont toujours compris entre 0 et 1 et sont d'autant plus grands que l'association entre valeurs observées et valeurs prédites est forte.

Les trois premiers coefficients sont fournis à la fois par Minitab et SAS mais le dernier n'est donné que par SAS. De très légères différences dans les résultats peuvent s'observer, l'algorithme utilisé par SAS procédant, par défaut, à un regroupement des observations en classes de probabilités estimées. Cette différence est cependant sans importance pratique.

Pour les données relatives au dépérissement, les résultats donnés par SAS sont repris à la figure 8.

7. INFORMATIONS COMPLÉMENTAIRES

Dans cette note, nous avons présenté divers outils permettant la critique d'un modèle de régression logistique binaire. Nous avons ainsi discuté des tests de signification d'une ou de plusieurs variables et de divers paramètres globaux quantifiant la qualité de l'ajustement. Les différentes statistiques proposées sont évidemment largement redondantes mais peuvent aussi, dans certains cas, conduire à des conclusions opposées.

Ainsi, un coefficient de régression pourrait être significatif lorsque le test est basé sur la méthode de WALD, mais être non significatif pour le test du rapport de vraisemblance, ou inversément. Rappelons que dans une telle situation, la préférence doit être donnée au test du rapport de vraisemblance.

De même, les tests χ^2 basés sur la déviance, sur les résidus de PEARSON et sur l'approche de HOSMER et LEMESHOW peuvent donner des résultats contradictoires. La première question à se poser à ce sujet concerne la validité de ces tests. Nous avons vu en effet que les tests basés sur la déviance et sur les résidus de PEARSON ne sont valables que si le nombre de profils différents générés par

les variables explicatives est fixé *a priori* et limité de manière à disposer d'un nombre suffisant d'observations par profil. Les statistiques n'ont aucune valeur inférentielle en présence de données individuelles ou de données groupées sur la base d'une variable quantitative observée avec une faible précision, comme c'est le cas pour l'altitude dans l'exemple examiné. Dans de telles situations, la préférence sera donnée au test de HOSMER et LEMESHOW.

Quant aux autres paramètres présentés (coefficients d'association et coefficients R^2), ils mesurent en réalité des caractéristiques différentes et ne peuvent donc pas être directement comparés. Ils peuvent cependant être utiles pour la comparaison de différents modèles de régression établis pour une même variable à expliquer. Dans ce cas, on comparera évidemment les valeurs obtenues d'un modèle à l'autre pour une même statistique.

La non-invariance de certains paramètres au groupement des données peut également perturber l'utilisateur. C'est le cas par exemple pour la déviance. Ainsi, pour les données relatives au chêne, il n'est pas possible de comparer le modèle ayant comme variable explicative l'altitude et le modèle ayant comme variable explicative la région. Le deuxième modèle est en effet un modèle saturé et conduit à une déviance nulle, alors que le premier modèle est un modèle non saturé, dont la déviance n'est pas nulle. Nous avons constaté au paragraphe 5.2 que le seconde modèle est effectivement préférable, non pas parce que sa déviance est plus faible, mais parce que la probabilité associée au test de conformité du coefficient de régression de la variable région est plus faible que la probabilité associée au test de conformité de la variable altitude (figures 2 et 3).

Enfin, rappelons encore une fois que des différences existent entre les résultats fournis par Minitab et SAS, d'une part en ce qui concerne les coefficients de régression en présence de variables explicatives qualitatives. Les deux logiciels utilisent en effet une paramétrisation différente, mais équivalente, comme nous l'avons vu au paragraphe 5.2.

De même, en présence de profils multiples, les deux logiciels donnent des résidus, au sens de la déviance et au sens de PEARSON qui sont différents, SAS calculant un résidu par observation et Minitab un résidu par profil.

BIBLIOGRAPHIE

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

ALLISON P. D. [1999]. *Logistic regression using SAS system : theory and application.* Cary, NC, SAS Institute, 302 p.

COLLETT D. [1999]. *Modelling binary data.* London, Chapman & Hall/CRC, 369 p.

DUYME F., CLAUSTRIAUX J. J. [2003]. La régression logistique binaire. *Notes Stat. Inform.* (Gembloux) 2004/6, 26 p.

GILLET A. [2005]. *Influences stationnelle, sylvicole et spécifique sur le dépérissement des chênes indigènes* (Quercus robur L. *et* Quercus petraea [*Matt.*]

Liebl.) *en Région wallonne.* Travail de fin d'étude. Gembloux, Faculté universitaire des Sciences agronomiques de Gembloux (Belgique), 75 p. + annexes.

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.

Menard S. [2000]. Coefficients of determination for multiple logistic regression analysis. *Amer. Stat.* 54, 17-24.

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

Palm R. et Brostaux Y. [2011]. La régression logistique avec Minitab. *Notes Stat. Inform.* (Gembloux) 2011/4, 15 p.

SAS Institute Inc. [2010]. *SAS/STAT 9.22 user's guide.* Cary, NC, SAS Institute Inc. 8460 p.

ANNEXE

Commandes Minitab et procédure SAS utilisées

Commandes Minitab

```
# Figures 2 et 5
Blogistic 'Dep_binaire' = Altitude Region;
  Factors 'Region';
  Logit;
  Reference 'Dep_binaire' 'Faible a tres fort' Region 'Condroz';
  Brief 2.

# Figure 4
Blogistic 'Dep_binaire' =  Region;
  Factors 'Region';
  Logit;
  Reference 'Dep_binaire' 'Faible a tres fort' Region 'Condroz';
  Brief 2.
```

Procedure SAS

```
* Figures 3,6,7 et 8
PROC LOGISTIC  DATA=lsas.deperissement;
 CLASS Region;
 MODEL Dep_binaire (event='F_Fort')= Altitude Region/
        CLPARM=Both Clodds=Both
        LACKFIT RSQ;
 run;
```

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 :

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.

PALM R., BROSTAUX Y. [2011]. La régression logistique avec Minitab. Notes Stat. Inform. (Gembloux) 2011/4, 15 p.

D/2011/2371/5