Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 141-149
Statistique Canada

Analyse longitudinale de données de l'enquête suisse sur la population active par régression logistique multidimensionnelle

PAUL-ANDRÉ SALAMIN[1]

RÉSUMÉ

Dans le cadre d'enquêtes longitudinales, de simples estimations de variations, comme des différences de pourcentages, peuvent ne pas être toujours suffisamment efficaces pour déceler des variations ayant de l' importance sur le plan pratique, notamment dans le cas de sous-populations. Le recours à des modèles, qui peuvent représenter la structure de dépendance de l'enquête longitudinale, peut aider à résoudre ce problème. Une des principales caractéristiques observées dans le cadre de l'Enquête suisse sur la population active (ESPA) est la situation d'activité. Étant donné que cette enquête comporte un plan d'échantillonnage à renouvellement de panel, les données qui en sont tirées sont des données nominales multidimensionnelles, et une grande partie des profils de réponse est absente conformément à ce plan. Le modèle logistique multidimensionnel, qui a été introduit par Glonek et McCullagh (1995) comme généralisation de la régression logistique présente de l'intérêt dans ce contexte parce qu'il autorise des observations répétées dépendantes et des profiles de réponse incomplets. Nous montrons qu'en utilisant la régression logistique multidimensionnelle, on peut représenter la structure de dépendance complexe de l'ESPA à l'aide d'un petit nombre de paramètres et obtenir des estimations plus efficaces d'une variation.

MOTS CLÉS: Données binaires longitudinales; modèle logistique multidimensionnel; enquête sur la population active.

1. INTRODUCTION

Un des principaux objectifs de l'Enquête suisse sur la population active (ESPA) est la production d'estimations des variations des pourcentages de population relatifs à diverses situations d'activité. Normalement, de simples estimations d'une variation, comme la différence entre les pourcentages observés d'une année à une autre au sujet des personnes qui occupent un emploi, sont calculées pour l'ensemble de la population, ainsi que pour un nombre important de sous-populations. En général, ce résultat est insatisfaisant, car les estimations relatives aux sous-populations peuvent ne pas être toujours suffisamment efficaces pour déceler des variations qui ont une importance sur le plan pratique. Le travail que nous présentons ici répond à la question de savoir si le recours à des modèles qui permettent de représenter la structure de dépendance de l'enquête précitée pouvait contribuer à résoudre ce problème.

Étant donné que l'ESPA est fondée sur un plan d'échantillonnage à renouvellement de panel, nous avons affaire à des données longitudinales nominales dont une grande partie des profils de réponse sont laissés incomplets conformément au plan. Le point central de notre étude est la modélisation des probabilités marginales, et notamment des probabilités d'être dans une situation d'activité donnée, en tant que fonction du temps et d'autres covariables qui définissent des sous-populations. Si les observations répétées de la situation d'activité étaient indépendantes, une façon normale de procéder consisterait à utiliser la régression logistique. Le modèle logistique multidimensionnel introduit par Glonek et McCullagh (1995) comme généralisation de la régression logistique est intéressant dans ce contexte car il autorise des observations répétées dépendantes et des profils de réponse incomplets.

Le but du présent article est de montrer que le recours à la régression logistique multidimensionnelle pour modéliser la structure de dépendance complexe de l'ESPA permet d'obtenir des estimateurs de variation plus efficaces. Bien que nous présentions la méthode en l'appliquant à des données de l'ESPA, il est clair qu'elle peut être appliquée à un éventail de données plus large.

De nombreuses questions importantes ne sont pas abordées dans le présent article. Parce que les données de l'ESPA sont tirées d'une enquête complexe, on peut soutenir que toute analyse devrait prendre en considération les poids d'échantillonnage (Pfeffermann 1993). Ici, nous utilisons uniquement les données non pondérées. Cependant, on peut montrer, à l'aide de la méthode de pseudo-vraisemblance de Binder (1983), que la régression logistique multidimensionnelle peut être élargie au cas en question (Salamin 1998). La non-réponse est toujours une préoccupation importante dans le cas des enquêtes sur échantillon. Ici, nous considérons uniquement les profils de réponse incomplets qui résultent du renouvellement du panel. Dans ce cas, l'hypothèse de valeurs manquantes complètement aléatoire est raisonnable. Il est à noter, cependant, que la régression logistique multidimensionnelle présente suffisamment de souplesse pour pouvoir intégrer des paramètres supplémentaires en considération des profils incomplets résultant de la perte de personnes faisant partie du panel. Ainsi, les personnes qui ont quitté le panel

[1] Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland.

auraient pu être incluses également dans l'analyse. Enfin, il est bien connu que des erreurs de classification peuvent introduire des biais importants dans les probabilités relatives aux profils de réponse observées (voir, par exemple, Pfeffermann, Skinner et Keith 1998). Il serait certainement souhaitable d'étudier l'incidence de ces biais sur les estimations de paramètres de la régression logistique multidimensionnelle, qui comportent des interprétations ayant trait à des moments marginaux.

Les modèles log-linéaires et les modèles marginaux sont étroitement liés à la régression logistique multidimensionnelle; nous y revenons dans la section 3. Nous présentons brièvement ci-après les modèles de transition, les modèles d'effets aléatoires et l'analyse de survie, dans le contexte de l'ESPA. Dans le cas d'un modèle de transition (voir, par ex., Diggle, Liang et Zeger 1994, chap. 10, ou Zeger et Liang 1992), les observations répétées de la situation d'activité sont corrélées parce que les situations d'activité passées ont une incidence sur la situation d'activité actuelle. Le point d'intérêt réside dans les probabilités de transition entre les diverses situations d'emploi; par exemple, la probabilité d'occuper un emploi peut dépendre du fait d'avoir été sans emploi dans le passé. Dans le contexte de la régression, les réponses antérieures sont traitées comme des variables explicatives supplémentaires. Un aspect important est la détermination du nombre de réponses antérieures à inclure comme variables prédictives. Si le modèle relatif aux probabilités de transition est correctement spécifié, on peut traiter des transitions répétées relatives à une personne en particulier comme des événements indépendants et utiliser des méthodes statistiques courantes, comme la régression logistique. Dans le cadre d'un modèle d'effets aléatoires (voir, par ex., Diggle et coll. 1994, chap. 9), la probabilité d'être dans une situation d'activité donnée est une fonction de variables explicatives et les coefficients de régression varient d'un individu à l'autre. Cette variabilité des coefficients de régression reflète l'hétérogénéité naturelle des individus qui est due à des facteurs non mesurés. Compte tenu des coefficients de régression, on présuppose que les observations répétées de la situation d'activité sont indépendantes. La corrélation entre les observations répétées survient uniquement parce que nous ne sommes pas en mesure d'observer les coefficients de régression réels. Cette façon de procéder présente la plus grande utilité lorsque le point d'intérêt réside dans l'inférence relative à des individus plutôt que dans des moyennes de population. Dans le cas de l'analyse de survie, que l'on désigne également par le terme «analyse des antécédents» dans les textes qui traitent d'économétrie (Lancaster 1990), le point d'intérêt est la modélisation des transitions entre les diverses situations d'activité au fil du temps, considérées comme une fonction de variables explicatives. Dans ce cas, le moment précis où a lieu la transition est important. Dans l'ESPA, la situation d'activité est observée une fois par an. Les changements qui ont eu lieu durant l'année qui a précédé l'interview peuvent être reconstitués. Cependant, étant donné que cette reconstitution est fondée sur l'autoévaluation des répondants, il peut y avoir des imprécisions en ce qui a trait aux situations antérieures et aux moments où les changement ont eu lieu. On peut trouver une analyse des données de l'ESPA fondée sur cette méthode dans Gerfin (1996).

Le présent article est organisé de la manière décrite ci-après. Dans la section 2, nous décrivons les données, un sous-ensemble d'environ 5000 individus tiré de l'ESPA, qui sont utilisées dans les exemples présentés dans les sections 4 et 5. Dans la section 3, nous présentons la régression logistique multidimensionnelle en la comparant avec les modèles log-linéaires et marginaux. Dans la section 4, nous montrons comment on peut utiliser la régression logistique multidimensionnelle pour représenter la structure de dépendance complexe des données de l'ESPA au moyen d'un faible nombre de paramètres. Dans la section 5, nous considérons l'estimation de la variation en comparant la régression logistique multidimensionnelle avec un simple estimateur de variation. Nous montrons que l'utilisation de la régression logistique multidimensionnelle aboutit à un gain d'efficacité. Enfin, dans la section 6, nous présentons nos conclusions et des orientations pour des études ultérieures.

2. DONNÉES DE L'ENQUÊTE SUISSE SUR LA POPULATION ACTIVE

On peut trouver une description détaillée du plan d'échantillonnage et de la méthode de pondération de l'ESPA dans Hulliger, Ries, Comment et Bender (1997). Nous rappelons ici uniquement certains des aspects importants de cette enquête. L'ESPA recueille des renseignements sur la situation d'activité de résidents de la Suisse âgés de quinze ans et plus. Depuis le deuxième trimestre de 1991, un échantillon d'environ 16 000 personnes fait l'objet d'interviews annuelles. Cette enquête est fondée sur un plan d'échantillonnage à renouvellement de panel, avec une durée de présence dans l'échantillon de cinq ans. Durant la phase de démarrage, c'est-à-dire entre 1992 et 1996, environ un cinquième de l'échantillon d'origine a été remplacé par un échantillon de renouvellement chaque année. Les unités de l'échantillon de renouvellement vont rester dans le panel pour une période complète de cinq ans.

Dans les exemples que nous donnons dans les sections 4 et 5, nous utilisons les observations relatives à la situation d'activité pour les années 1992 à 1995 qui ont été obtenues auprès des personnes faisant partie de l'échantillon du canton de Vaud. Le tableau 1 indique la structure des données ainsi que la taille des échantillons longitudinal et transversal. En raison du plan d'échantillonnage qui a été choisi, certains des profils de réponse sont incomplets. Par exemple, dans le cas des personnes qui ont été choisies en

1991 et qui ont été remplacées en 1994, la période d'observation, (notée (1)234, va de 1991 à 1994. Nous utilisons la notation (1)234 pour indiquer le fait que nous n'avons pas eu recours aux observations qui ont été effectuées en 1991.

Tableau 1
Structure des données, taille des échantillons longitudinal et transversal, canton de Vaud, 1992-1995

Première année dans l'échantillon	Époques d'observation des diverses parties de l'échantillon				Période d'observation	
91	92				(1)2	622
	92	93			(1)23	412
	92	93	94		(1)234	527
	92	93	94	95	(1)2345	481
92	92	93	94	95	2345	612
93		93	94	95	345	722
94			94	95	45	728
95				95	5	877
	2 654	2 754	3 070	3 420		4 981

La situation d'activité est une variable nominale qui comprend les trois catégories suivantes: «personne occupée», «sans emploi» et «hors de la population active». Dans les exemples que nous donnons dans les sections 4 et 5, nous utilisons une variable binaire qui prend la valeur 1 si la personne sondée occupe un emploi, et la valeur 2 si elle est sans emploi ou hors de la population active. Nous procédons de cette façon uniquement pour simplifier la présentation des modèles logistiques multidimensionnels. Étant donné que la méthode que nous utilisons permet de traiter un nombre arbitraire de catégories, il serait préférable de ne pas regrouper les situations d'activité dans le cas d'une analyse réelle. Dans le cas où il serait néanmoins nécessaire de regrouper certaines des situations d'activité, il faut faire preuve de prudence, étant donné que l'hétérogénéité des situations peut introduire des biais.

3. MODÈLES LOGISTIQUES MULTIDIMENSIONNELS

Le modèle logistique multidimensionnel qui a été introduit par Glonek et McCullagh (1995) permet de traiter des réponses multidimensionnelles de type nominal ou ordinal ainsi que des variables explicatives discrètes ou continues. Nous allons considérer ici uniquement des réponses binaires multidimensionnelles et des variables prédictives discrètes. Le modèle logistique multidimensionnel est un exemple de modèle linéaire généralisé (voir McCullagh et Nelder 1989). Sa fonction de liaison, appelée également transformation logistique multidimensionnelle, exprime la distribution conjointe des profils de réponse sous forme de moments marginaux par ordre croissant, les deux premiers moments étant des logits marginaux et des risques relatifs logarithmiques marginaux. La fonction de liaison a une caractéristique, désignée par le terme «reproductibilité», qui fait que le modèle logistique multidimensionnel s'applique à tout sous-ensemble du vecteur de réponse. Cette caractéristique assure l'égalité des interprétations des paramètres, quel que soit le nombre de variables de réponse, et que des paramètres d'un ordre supérieur soient inclus ou non. Cette propriété rend la régression logistique multidimensionnelle particulièrement intéressante pour l'analyse de données longitudinales, où les observations répétées d'un résultat ont lieu sur un pied d'égalité et où le nombre de ces observations répétées peut varier d'un individu à l'autre. La reproductibilité est également l'élément clé qui permet d'utiliser le modèle pour traiter des réponses incomplètes. Cependant, il est à noter qu'il faut présupposer que l'absence complète de données est aléatoire, si les mêmes paramètres doivent être utilisés pour modéliser aussi bien les réponses complètes que les réponses incomplètes. Les estimations de paramètres sont déterminées selon la méthode du maximum de vraisemblance. Une opération clé est l'inversion de la transformation logistique multidimensionnelle. Dans le cas de plus de trois réponses, il se peut que cette opération ne puisse pas toujours être effectuée, car dans un tel cas il existe des contraintes qui touchent les paramètres (Glonek et McCullagh 1995, Liang, Zeger et Qaqish 1992). En outre, la présence de cellules vides peut limiter l'ordre des paramètres qui peuvent être ajustés.

Le modèle log-linéaire est largement utilisé pour modéliser des données binaires multidimensionnelles. Dans le modèle log-linéaire saturé (voir, par ex., Liang et coll. 1992), le paramètre canonique associé à un sous-ensemble des variables a une interprétation sous forme de probabilités conditionnelles, compte tenu des autres variables; par exemple, les paramètres de premier et deuxième ordre sont des logits et des risques relatifs logarithmiques qui dépendent de toutes les autres réponses. Il s'ensuit que le modèle log-linéaire n'est pas reproductible, ce qui en fait un choix moins intéressant que la régression logistique multidimensionnelle pour l'analyse des données longitudinales. Il est néanmoins possible d'élaborer des modèles log-linéaires qui, à l'instar du modèle logistique multidimensionnel, comportent des logits marginaux comme paramètres. Cela mène aux modèles marginaux (Diggle et coll, 1994, chap. 8). Dans ces modèles, la dépendance des probabilités marginales à l'égard de variables explicatives est modélisée de manière distincte de la corrélation existant au sein des unités. Avec cette approche, les paramètres ne sont pas estimés par la méthode du maximum de vraisemblance; on spécifie plutôt uniquement la structure de la corrélation existant entre les observations répétées d'un résultat, et les paramètres sont estimés par la résolution d'équations d'estimation généralisées (EEG), un analogue multidimensionnel de la quasi-vraisemblance (McCullagh et Nelder 1989). Un certain nombre de spécifications de la structure de corrélation ont été proposées; par exemple, Liang et coll. (1992) ont utilisé les risques relatifs marginaux comme

Glonek et McCullagh (1995). Nous avons effectué certaines comparaisons entre la régression logistique multidimensionnelle et le progiciel PROC GENMOD de SAS (version 6.12). Ce progiciel permet d'ajuster des modèles de réponses corrélées d'après la méthode des EEG. Nous avons obtenu des estimations des logits marginaux très semblables. La méthode des EEG s'est avérée légèrement moins efficace que la régression logistique multidimensionnelle. Un des inconvénients de cette méthode est qu'elle ne permet pas d'obtenir des estimations de toutes les probabilités de profils de réponse, mais uniquement les probabilités marginales. Le modèle logistique multidimensionnel ne présente pas cet inconvénient parce que ses paramètres sont estimés selon la méthode du maximum de vraisemblance.

En suivant Glonek et McCullagh (1995), nous décrivons, dans la section 3.1, la transformation logistique multidimensionnelle, et nous donnons, dans la section 3.2, l'algorithme relatif au maximum de vraisemblance.

3.1 Transformation logistique multidimensionnelle

Soient $Y_1, Y_2, ..., Y_d$ un nombre d d'observations répétées de la même variable binaire, effectuées aux époques $t_1 < t_2 < ... < t_d$, et soit

$$\pi_{i_1 i_2 ... i_d} = P(Y_1 = i_1, Y_2 = i_2, ..., Y_d = i_d),$$

où $i_1, i_2, ..., i_d$ ont comme valeur 1 ou 2, les probabilités conjointes des variables aléatoires $Y_1, Y_2, ..., Y_d$. Dans le modèle logistique multidimensionnel, les probabilités conjointes de $Y_1, Y_2, ..., Y_d$ sont paramétrées sous forme de logits marginaux, de risques relatifs logarithmiques marginaux et de contrastes de risques relatifs logarithmiques marginaux. Ce paramétrage peut être exprimé de la manière suivante: $\eta = C^T \log(L\pi)$, où π est le vecteur de dimension $q = 2^d$

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

et où les matrices L et C sont les produits tensoriels, respectivement, d'un indicateur marginal et de matrices de contraste appropriés. Les matrices L et C, qui dépendent de la durée d de la période d'observation, sont définies dans une suite récurrente, en commençant par $L_0 = C_0 = 1$, avec

$$L_d = \begin{bmatrix} L_{d-1} \otimes 1_2^T \\ L_{d-1} \otimes \tilde{L} \end{bmatrix}$$

et

$$C_d = \begin{bmatrix} C_{d-1} & 0 \\ 0 & C_{d-1} \otimes \tilde{C} \end{bmatrix},$$

où $1_2^T = (1, 1)$, $\tilde{L}$ la matrice unité deux par deux et $\tilde{C} = (1, -1)^T$ (Glonek et McCullagh 1994).

À titre d'exemple, nous considérons des périodes d'observation d'une durée $d = 1, 2, 3, 4$. Dans le cas de $d = 1$, $\pi = (\pi_1, \pi_2)^T$ et $\eta = (\eta_0, \eta_1)^T = (\log \pi_+, \text{logit } Y_1)^T$, où l'indice positif indique la sommation et où logit Y_1 est défini comme suit:

$$\text{logit } Y_1 = \log \frac{P(Y_1 = 1)}{P(Y_1 = 2)} = \log \frac{\pi_1}{1 - \pi_1} = \log \frac{\pi_1}{\pi_2}.$$

Dans ce cas, la transformation logistique multidimensionnelle est équivalente à la transformation logistique habituelle. Il est à noter que même si le paramètre $\eta_0 = \log \pi_+ = 0$ est tout à fait superflu, il est utile de le conserver comme moyen permettant de s'assurer que l'application $\pi \to \eta$ est de plein niveau et exprime également l'exigence selon laquelle $\pi_+ = 1$.

Lorsque $d = 2$, $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12})^T =$$

$$(\log \pi_{++}, \text{logit } Y_1, \text{logit } Y_2, \log OR(Y_1, Y_2))^T$$

où

$$RR(Y_1, Y_2) =$$

$$\frac{P(Y_1 = 1, Y_2 = 1) P(Y_1 = 2, Y_2 = 2)}{P(Y_1 = 1, Y_2 = 2) P(Y_1 = 2, Y_2 = 1)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

est le risque relatif, une grandeur qui mesure l'association entre les variables Y_1 et Y_2. Les paramètres η_1 et η_2 sont les logits marginaux aux époques t_1 et t_2 ; par exemple:

$$\eta_1 = \text{logit } Y_1 = \log \frac{\pi_{1+}}{(1 - \pi_{1+})}.$$

Lorsque $d = 3$, $\pi = (\pi_{111}, \pi_{112}, ..., \pi_{221}, \pi_{222})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123})^T.$$

Les paramètres η_1, η_2 et η_3 sont les logits marginaux aux époques t_1, t_2 et t_3. Les paramètres η_{12}, η_{13} et η_{23} sont les risques relatifs logarithmiques des tableaux marginaux bidimensionnels correspondants; par exemple:

$$\eta_{23} = \log RR(Y_2, Y_3) = \log \frac{\pi_{+11}\,\pi_{+22}}{\pi_{+12}\,\pi_{+21}}.$$

Le paramètre η_{123} est un contraste des risques relatifs qui est donné par

$$\eta_{123} = \log RR(Y_1, Y_2 \mid Y_3 = 1) - \log RR(Y_1, Y_2 \mid Y_3 = 2)$$
$$= \log \frac{\pi_{111}\pi_{221}}{\pi_{121}\pi_{211}} - \log \frac{\pi_{112}\pi_{222}}{\pi_{122}\pi_{212}}.$$

Lorsque $d = 4$, $\pi = (\pi_{1111}, \pi_{1112}, ..., \pi_{2221}, \pi_{2222})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123},$$
$$\eta_4, \eta_{14}, \eta_{24}, \eta_{124}, \eta_{34}, \eta_{134}, \eta_{234}, \eta_{1234})^T.$$

Les paramètres η_i, η_{ij} et η_{ijk}, où $1 \le i < j < k \le 4$, sont définis comme ci-dessus à l'aide des tableaux marginaux appropriés. Le paramètre η_{1234} est un contraste des risques relatifs logarithmiques qui est donné par

$$\eta_{1234} = \log RR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 1)$$
$$- \log RR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 2)$$
$$- \log RR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 1)$$
$$+ \log RR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 2).$$

Une opération clé de l'estimation du maximum de vraisemblance est le calcul de l'inverse de la transformation logistique multidimensionnelle. Pour faire en sorte que $\pi > 0$, nous utilisons $\pi = \exp \nu$, c'est-à-dire que nous cherchons à trouver ν dans l'équation $\eta = C^T \log(L \exp \nu)$. En général, on ne dispose pas d'une solution explicite, de sorte qu'il faut utiliser une méthode itérative. On peut notamment appliquer les itérations de Newton-Raphson, comme nous l'expliquons ci-après. Pour plus de clarté, nous définissons les deux fonctions $\varphi(\pi) = C^T \log(L\pi)$ et $\psi(\nu) = \varphi(\exp \nu)$.

i) Commencer avec une approximation initiale ν_0.
ii) Prendre ensuite $\nu_{k+1} = \nu_k - [D\psi(\nu_k)]^{-1}(\varphi(\exp \nu_k) - \eta)$, où $D\psi(\nu)$ est la matrice jacobienne de la fonction $\psi(\nu)$, et itérer jusqu'à convergence.

Les matrices jacobiennes des fonctions $\varphi(\pi)$ et $\psi(\nu)$ sont données respectivement par $D\varphi(\pi) = C^T (\operatorname{diag} L\pi)^{-1} L$ et $D\psi(\nu) = (\exp \nu) \cdot \operatorname{diag}(\exp \nu)$.

3.2 Estimation du maximum de vraisemblance

Dans le cas d'une variable de réponse binaire observée à d époques il y a $q = 2^d$ profils de réponse possibles, où $i_1, i_2, ..., i_d$ ont comme valeur soit 1, soit 2. Dans le cas de chaque profil $i = (i_1, ..., i_d)$, nous définissons la variable indicatrice $Y_{i_1 ... i_d}$, qui est égale à 1 si l'on a observé le profil i, et à 0 dans le cas contraire. On obtient alors

$$P(Y_{i_1 ... i_d} = 1) = P(Y_1 = i_1, ..., Y_d = i_d) = \pi_{i_1 ... i_d}.$$

Les vecteurs de grandeur q étant définis comme suit:

$$Y = (Y_{11...11}, Y_{11...12}, ..., Y_{22...21}, Y_{22...22})^T$$

et

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

nous pouvons écrire $Y \sim M(1, \pi)$, c'est-à-dire que Y est un vecteur multinomial avec $q = 2^d$ catégories dont les probabilités sont données par le vecteur π.

Les modèles de régression logistique multidimensionnelle sont donc définis pour être ceux de la forme $\eta = X\beta$, où X est une matrice $q \times p$ de variables explicatives, β un vecteur de grandeur p de paramètres inconnus, et où $\eta = C^T \log(L\pi) = \varphi(\pi)$.

Si nous prenons y comme observation du vecteur aléatoire Y, nous pouvons alors exprimer le noyau de la fonction de vraisemblance logarithmique comme suit: $l(\beta; y) = y^T \log \pi(\beta)$, où, en utilisant l'inverse de la transformation logistique multidimensionnelle, nous pouvons exprimer les probabilités conjointes π comme une fonction du paramètre inconnu β, sous la forme $\pi(\beta) = \varphi^{-1}(X\beta)$. Le vecteur de valeur numérique est donné par

$$s(\beta) = s(\beta, y, X) = D\pi(\beta)^T (\operatorname{diag} \pi(\beta))^{-1} y,$$

où $D\pi(\beta)$, c'est-à-dire la matrice jacobienne de la fonction $\pi(\beta)$ qui lie le paramètre β au vecteur de probabilités π, est donnée par $D\pi(\beta) = [D\varphi(\varphi^{-1}(X\beta))]^{-1} X$, et où $D\varphi(\pi) = C^T (\operatorname{diag} L\pi)^{-1} L$ est la matrice jacobienne de la fonction de liaison. La matrice d'information est définie comme suit: $\Im(\beta) = Es(\beta)s(\beta)^T$. Il s'ensuit, d'après la présupposition relative à la distribution de Y, que $E(YY^T) = \operatorname{diag} \pi$, d'où nous pouvons déduire que

$$\Im(\beta) = \Im(\beta, X) = D\pi(\beta)^T (\operatorname{diag} \pi(\beta))^{-1} D\pi(\beta).$$

Si nous avons un nombre n d'observations indépendantes $y_k \sim M(1, \pi_k)$, $k = 1, ..., n$, où $\eta_k = C^T \log(L\pi_k) = X_k\beta$, le vecteur de valeur numérique et la matrice d'information sont donnés alors par $s(\beta) = \sum_{k=1}^n s(\beta, y_k, X_k)$ et par $\Im(\beta) = \sum_{k=1}^n \Im(\beta, X_k)$.

L'estimateur du maximum de vraisemblance de β est la solution de $s(\beta) = 0$, qui peut être trouvée à l'aide de l'algorithme de cotation de Fisher qui, à partir d'une quelconque valeur initiale β_0, itère la séquence $\beta_{m+1} + \beta_m + \Im_m^{-1}(\beta_m) s(\beta_m)$ jusqu'à convergence.

Des profils de réponse incomplets peuvent être intégrés facilement à l'analyse. En particulier, lorsqu'un quelconque sous-ensemble des variables de réponse $Y_1, Y_2, ..., Y_c$ est enregistré pour une unité donnée, la distribution des probabilités sur ce tableau marginal de grandeurs c est multinomiale et, en raison de la reproductibilité de la transformation logistique multidimensionnelle, un modèle de régression logistique multidimensionnelle s'applique au tableau des probabilités. En outre, la matrice de plan

d'échantillonnage qui lie les probabilités marginales à β est établie en choisissant les rangs appropriée de la matrice de plan d'échantillonnage complète qui serait utilisée si l'on disposait de données complètes concernant l'unité en question.

4. MODÈLES POUR LA DÉPENDANCE LONGITUDINALE

Dans la présente section, nous montrons, à l'aide des données de l'ESPA indiquées dans la section 2, comment on peut appliquer la régression logistique multidimensionnelle pour décrire la dépendance existant entre les observations répétées de la situation d'activité. Nous n'entendons pas effectuer une recherche exhaustive pour trouver un modèle optimal; nous allons plutôt montrer que la régression logistique multidimensionnelle permet de représenter une structure de dépendance complexe à l'aide d'un petit nombre de paramètres.

Nous considérons six modèles à complexité décroissante (voir tableau 2). Dans le cas de tous les modèles, nous avons un paramètre pour chacun des logits marginaux qui correspond à une époque d'observation donnée. Ce paramètre est désigné par la notation $\eta_i = \beta_i$. Étant donné que les époques d'observation sont les deuxièmes trimestres des années 1992 à 1995, nous prenons $i = 2, 3, 4, 5$. Ainsi, β_3, par exemple, correspond au logit de la probabilité d'occuper un emploi en 1993. De façon analogue, les indices pour les paramètres d'ordre supérieur vont de 2 à 5. Dans le cas du modèle 1, nous prenons un modèle saturé pour la dépendance longitudinale, c'est-à-dire que nous avons un paramètre pour chacune des interactions d'ordre 2, 3 ou 4 à l'intérieur de chaque période d'observation. Dans le cas des modèles 2 à 5, nous présupposons que les interactions d'ordre 3 et 4 sont toutes égales à zéro. La dépendance longitudinale est alors décrite seulement sous forme de risques relatifs logarithmiques. Dans le cas du modèle 2, nous avons recours à un modèle saturé pour les risques relatifs logarithmiques. Dans le modèle 3, nous laissons tomber la covariable période d'observation, c'est-à-dire que nous présupposons que les risques relatifs logarithmiques sont les mêmes pour toutes les périodes d'observation. Dans le modèle 4, nous utilisons des risques relatifs logarithmiques stationnaires, c'est-à-dire des risques relatifs logarithmiques qui dépendent uniquement de la différence entre les époques d'observation. Il est à noter que dans le modèle 4, le paramètre γ_1 correspond à la contrainte $\beta_{23} = \beta_{34} = \beta_{45}$ à laquelle sont assujettis les paramètres du modèle 3, et la même chose vaut pour γ_2 et γ_3. Dans le cas du modèle 5, on présuppose un modèle linéaire pour les risques relatifs logarithmiques stationnaires. Enfin, dans le modèle 6, nous présupposons que les observations effectuées à des époques différentes sont indépendantes. Il est à noter que dans ce cas, la régression logistique multidimensionnelle équivaut à la régression logistique ordinaire.

Tableau 2
Six modèles pour la dépendance longitudinale

	Paramètres		
Modèle	Logits marginaux	Risques relatifs logarithmiques	Paramètres de 3ième et 4ième ordre
1	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij,\text{période}}$	$\eta_{ijk} = \beta_{ijk,\text{période}}$, $\eta_{ijkl} = \beta_{ijkl,\text{période}}$
2	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij,\text{période}}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
3	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
4	$\eta_i = \beta_i$	$\eta_{ij} = \gamma_{\lvert i-j \rvert}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
5	$\eta_i = \beta_i$	$\eta_{ij} = \delta + \gamma . \lvert i-j \rvert$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
6	$\eta_i = \beta_i$	$\eta_{ij} = 0$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$

Les estimations de paramètres pour les modèles 2 à 6 sont indiquées dans le tableau 3. Le nombre de paramètres et les valeurs de la fonction de vraisemblance logarithmique dans le cas des estimations du maximum de vraisemblance sont indiqués dans le tableau 4, où nous avons inclus également, à des fins de comparaison, la vraisemblance logarithmique pour le modèle entièrement saturé.

Dans l'ensemble, nous constatons que la forme présupposée de la dépendance longitudinale semble avoir peu d'incidence sur les estimations des logits marginaux. Il s'agit là d'une caractéristique souhaitable, étant donné que les logits marginaux seraient normalement les paramètres auxquels on s'intéresse. Les erreurs-types des logits marginaux sont presque les mêmes dans le cas des modèles qui tiennent compte de la dépendance longitudinale, mais ils sont gonflés dans une mesure d'environ 15 % dans le cas de la régression logistique ordinaire (modèle 6). On peut également montrer que les estimations des logits marginaux présentent entre elles une corrélation positive dans le cas de modèles qui présupposent une dépendance longitudinale, et aucune corrélation dans le cas de la régression logistique ordinaire. Dans l'exemple qui nous intéresse, nous avons constaté que la corrélation se situe entre 0,4 et 0,8. Ainsi, la modélisation de la dépendance longitudinale permet également d'obtenir des estimations plus efficaces de la différence existant entre les logits marginaux.

D'après l'ajustement du modèle 1, on peut voir que les paramètres d'interaction d'ordre 3 et 4 ne sont pas très différents de 0. Cela laisse penser que la dépendance longitudinale peut être décrite uniquement par les risques relatifs logarithmiques. Cette hypothèse est corroborée par l'écart incrémental du modèle 2 par rapport au modèle 1, qui est de 7,9 pour 12 degrés de liberté. En outre, tous les paramètres du modèle 2 sont différents de zéro de manière importante, et un examen des résidus normalisés dans le cas des probabilités ajustées relatives aux profils de réponse ne révèle aucune anomalie. Dans le cas d'applications relatives à des statistiques officielles, le modèle 2 serait le modèle privilégié, étant donné qu'il est fondé sur un nombre de présuppositions aussi faible que possible tout en permettant une réduction importante du nombre de paramètres, ce qui

diminue le risque d'avoir des tableaux clairsemés lorsqu'on considère des périodes d'observation prolongées comportant un plus grand nombre de covariables.

Les modèles 3, 4 et 5 montrent qu'il serait néanmoins possible de simplifier grandement la description de la dépendance longitudinale sans perdre trop d'informations. En passant du modèle 2 au modèle 5, nous constatons que l'écart par rapport au modèle entièrement saturé n'augmente pas beaucoup (voir tableau 4). En outre, un examen des résidus montre que les modèles 3, 4 et 5 correspondent aux données presque aussi bien que le modèle 2. D'autre part, si le modèle 2 requiert vingt paramètres pour décrire la dépendance longitudinale, le modèle 5 n'en nécessite que deux. Ce fait doit être opposé au modèle 6, qui présuppose l'indépendance entre les observations effectuées à des époques différentes; la vraisemblance logarithmique est beaucoup plus faible que dans le cas du modèle entièrement saturé (voir tableau 4) et l'ajustement aux données laisse à désirer.

Tableau 3
Estimations de paramètres et erreurs-types

Paramètres	Période	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6
logit 92		0,6348 (0,0350)	0,6360 (0,0352)	0,6348 (0,0352)	0,6347 (0,0352)	0,6471 (0,0409)
logit 93		0,5555 (0,0335)	0,5570 (0,0338)	0,5597 (0,0335)	0,5601 (0,0335)	0,5509 (0,0396)
logit 94		0,5440 (0,0324)	0,5407 (0,0325)	0,5402 (0,0326)	0,5397 (0,0325)	0,5377 (0,0374)
logit 95		0,4699 (0,0317)	0,4711 (0,0320)	0,4710 (0,0320)	0,4712 (0,0320)	0,4705 (0,0351)
β_{23}	(1)23	4,2563 (0,3311)	4,2579 (0,1465)			
	(1)234	4,2003 (0,2894)				
	(1)2345	4,0859 (0,2954)				
	2345	4,4830 (0,2841)				
β_{34}	(1)234	4,0894 (0,2794)	4,1111 (0,1310)			
	(1)2345	3,9611 (0,2840)				
	2345	4,0989 (0,2600)				
	345	4,2490 (0,2468)				
β_{45}	(1)2345	5,3992 (0,3854)	4,5561 (0,1389)			
	2345	3,9779 (0,2544)				
	345	4,7288 (0,2735)				
	45	4,5069 (0,2600)				
β_{24}	(1)234	3,7168 (0,2641)	3,8371 (0,1442)			
	(1)2345	4,2560 (0,3059)				
	2345	3,5330 (0,2370)				
β_{35}	(1)2345	4,4000 (0,3098)	3,7913 (0,1334)			
	2345	3,6493 (0,2396)				
	345	3,6116 (0,2192)				
β_{25}	(1)2345	4,3984 (0,3173)	3,5774 (0,1530)			
	2345	3,2209 (0,2256)				
γ_1				4,3260 (0,0928)		
γ_2				3,8519 (0,1050)		
γ_3				3,5340 (0,1495)		
δ					4,7341 (0,1266)	
γ					-0,4191 (0,0653)	

Tableau 4
Nombre de paramètres et valeur de la fonction de vraisemblance logarithmique dans le cas des estimations du maximum de vraisemblance

Modèle	Nombre de paramètres d'ordre				Total	Vraisemblance logarithmique
	1	2	3	4		
Modèle complet	20	20	10	2	52	-5 342,7
1	4	20	10	2	36	-5 345,4
2	4	20	0	0	24	-5 349,4
3	4	6	0	0	10	-5 365,2
4	4	3	0	0	7	-5 368,9
5	4	2	0	0	6	-5 369,5
6	4	0	0	0	4	-7 815,3

5. COMPARAISON AVEC L'ESTIMATION SIMPLE DE LA VARIATION

Dans la présente section, nous nous concentrons sur l'estimation de la différence, observée d'une année donnée à une autre, dans les probabilités d'occuper un emploi. Nous montrons que des estimations fondées sur la régression logistique multidimensionnelle sont plus efficaces que des estimations simples définies comme étant la différence existant entre les proportions de personnes qui occupent un emploi.

Le modèle qui est considéré ici est le modèle 2 de la section 4, le sexe étant ajouté comme variable explicative supplémentaire. Dans le cas de chacun des sexes, nous avons un paramètre pour chacun des logits marginaux qui correspond à une année donnée. La dépendance longitudinale est prise en considération au moyen d'un modèle saturé pour les risques relatifs logarithmiques. Les paramètres de troisième et quatrième ordre sont établis à zéro. Ce modèle comporte donc 8 paramètres pour les logits marginaux et 40 paramètres pour les risques relatifs logarithmiques (2 sexes × 20 risques relatifs à l'intérieur des périodes d'observation; voir tableau 3). En inversant la transformation logistique multidimensionnelle, on peut également calculer les estimations des probabilités d'occuper un emploi et de la variation de ces probabilités d'une année à l'autre.

Un estimateur simple de variation est donné par la différence existant entre les proportions de personnes occupant un emploi qui est observée d'une année donnée à une autre. La variance, qui tient compte du chevauchement des deux échantillons, est donnée par l'expression suivante:

$$\frac{1}{n+r}\pi_{1+}(1-\pi_{1+}) + \frac{1}{n+c}\pi_{+1}(1-\pi_{+1}) - 2\frac{n}{(n+r)(n+c)}(\pi_{11}-\pi_{1+}\pi_{+1}),$$

où n est le nombre de cas au sujet desquels on dispose d'observations pour les deux années, r et c le nombre de cas au sujet desquels on dispose d'observations relatives à un an, π_{11} la probabilité d'occuper un emploi durant les deux ans, et π_{1+} et π_{+1} les probabilités marginales d'occuper un emploi.

Concernant les données de l'ESPA de la section 2, le tableau 5 indique les estimations de la différence de la probabilité d'occuper un emploi obtenues avec les deux méthodes. On remarquera que ces méthodes donnent des estimations de la variation similaires. Les erreurs-types des estimations simples sont, en moyenne, 30 % plus importants que dans le cas de la régression logistique multidimensionnelle. L'efficacité relative moyenne de la régression logistique multidimensionnelle par rapport aux estimations simples est de 1,7. Par comparaison, l'efficacité relative moyenne de la régression logistique multidimensionnelle par rapport à la régression logistique ordinaire est de 3,2.

Tableau 5
Variation de la probabilité d'occuper un emploi, canton de Vaud, 1992-1995

	Comparaison	Régression logistique multidimensionnelle	Estimation simple
Femme	92 *vs.* 93	0,0138 (0,0090)	0,0136 (0,0115)
	92 *vs.* 94	0,0184 (0,0102)	0,0168 (0,0134)
	92 *vs.* 95	0,0375 (0,0109)	0,0356 (0,0149)
	93 *vs.* 94	0,0047 (0,0087)	0,0031 (0,0107)
	93 *vs.* 95	0,0238 (0,0095)	0,0219 (0,0128)
	94 *vs.* 95	0,0191 (0,0076)	0,0188 (0,0100)
Homme	92 *vs.* 93	0,0220 (0,0095)	0,0283 (0,0116)
	92 *vs.* 94	0,0245 (0,0102)	0,0334 (0,0133)
	92 *vs.* 95	0,0387 (0,0106)	0,0452 (0,0144)
	93 *vs.* 94	0,0024 (0,0092)	0,0052 (0,0111)
	93 *vs.* 95	0,0167 (0,0098)	0,0169 (0,0130)
	94 *vs.* 95	0,0143 (0,0080)	0,0117 (0,0102)

6. CONCLUSIONS

Les analyses des données de l'ESPA que nous avons décrites dans le présent article montrent l'utilité de la régression logistique multidimensionnelle. La modélisation de la dépendance longitudinale est nécessaire pour obtenir un ajustement satisfaisant des probabilités relatives aux profils de réponse qui sont observés. En ne tenant pas compte de la dépendance longitudinale, nous obtenons quand même des estimations ponctuelles acceptables des logits marginaux, mais les informations sur la structure détaillée des données sont perdues. La modélisation de la dépendance longitudinale permet également d'obtenir des

estimations plus efficaces des paramètres marginaux et de la variation, comparativement à la régression logistique ordinaire et à l'estimation simple de la variation. Enfin, nous avons décrit la possibilité qu'offre la régression logistique multidimensionnelle de représenter une structure de dépendance complexe à l'aide d'un petit nombre de paramètres.

En utilisant les résultats obtenus par Glonek et McCullagh (1995), il est possible d'appliquer les exemples que nous avons présentés ici à des réponses multidimensionnelles de type nominal ou ordinal, avec des variables explicatives discrètes ou continues. La méthode que nous avons décrite permet également de tenir compte des poids d'échantillonnage (Salamin 1998). Dans le cas de l'ESPA, nous avons constaté que les poids d'échantillonnage ont peu d'incidence sur les estimations de paramètres du modèle logistique multidimensionnel. L'erreur-type des estimations de paramètres était gonflé d'environ 15 %. Cette augmentation modérée de la variabilité de ces estimations, qui est due aux poids d'échantillonnage, est plausible. En effet, étant donné qu'on choisit une seule personne par ménage dans le cadre de l'ESPA, nous ne nous attendions pas à un effet de grappe important.

Outre que pour les types d'analyse que nous avons présentés ici, la régression logistique multidimensionnelle peut être utilisée également pour modéliser des probabilités de non-réponse dans le cas d'études longitudinales. De tels modèles pourraient être utiles lorsqu'il faut corriger des poids d'échantillonnage pour tenir compte de la non-réponse. La possibilité qu'offre la régression logistique multidimensionnelle d'obtenir un modèle parcimonieux des données pourrait avoir de l'intérêt dans le cas de l'estimation relative à de petites région. Des estimateurs pour une région géographique donnée pourraient notamment être fondés sur des modèles relatifs à une région plus vaste choisie de manière appropriée.

Bien que nous n'ayons pas rencontré des difficultés importantes dans les exemples que nous avons présentés ici, des études ultérieures pourraient être nécessaires en ce qui a trait au problème des tableaux clairsemés. L'inversion de la transformation logistique multidimensionnelle est une opération importante lorsqu'il y a un grand nombre de cellules vides. La méthode proposée par Lang (1996), dans laquelle l'inversion de la fonction de liaison est évitée en spécifiant les modèles par l'entremise de contraintes, pourrait être intéressante dans un tel contexte. Un autre aspect à étudier est l'incidence des erreurs de classification sur les estimations de paramètres du modèle logistique multidimensionnel.

REMERCIEMENTS

Pour la rédaction du présent article, nous avons bénéficié d'entretiens avec des collègues de l'Office fédéral de la statistique, parmi lesquels méritent d'être mentionnés spécialement Beat Hulliger et Philippe Eichenberger. L'aide d'Ariane Bender, de la section responsable de l'Enquête suisse sur la population active, a été extrêmement utile. L'auteur remercie également le rédacteur et deux examinateurs pour leurs excellents conseils, qui ont permis d'améliorer cet article de manière importante.

BIBLIOGRAPHIE

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

DIGGLE, P.J., LIANG, K.-Y., et ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

GERFIN, M. (1996). *Entwicklung von ökonometrischen Modellen zur Analyse der Dynamik auf dem schweizerischen Arbeitsmarkt*. SLFS-News, Swiss Federal Statistical Office, Berne.

GLONEK, G.F.V., et McCULLAGH, P. (1994). Multivariate Logistic Models. Rapport technique 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide.

GLONEK, G.F.V., et McCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, B, 57, 533-546.

HULLIGER, B., RIES, A., COMMENT, T., et BENDER, A. (1997). Weighting the Swiss Labour Force Survey. (Éds. C. Malaguerra, S. Morgenthaler et E. Ronchetti). Dans *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switerland*, Basel: Birkhäuser Verlag.

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

LANG, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, 24, 726-752.

LIANG, K.-Y., ZEGER, S.L., et QAQISH, B. (1992). Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society*, B, 54, 3-40.

McCULLAGH, P., et NELDER, J.A. (1989). *Generalized Linear Models*, (2ième édition). London: Chapman and Hall.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

PFEFFERMANN, D., SKINNER, C., et KEITH, H. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, A, 161, Part 1, 13-32.

SALAMIN, P.-A. (1998). Multivariate logistic regression for data from complex surveys. À paraître *Recueil: Symposium '98, Analyse longitudinale pour les enquêtes complexes*, Statistique Canada, Mai 1998.

ZEGER, S.L., et LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.