Estimation structurée par pénalisation L_1 dans le modèle de régression logistique polytomique.

Vivian Viallon & Cédric Garcia

Univ. Lyon, Université Claude Bernard Lyon1, Ifsttar, UMRESTTE, UMR T_9405, F-69373, LYON.

Résumé. Nous nous intéressons à l'estimation paramétrique dans le modèle de régression logistique polytomique, ou multinomiale. Ce modèle est classique lorsque la variable réponse est catégorielle, en l'absence de relation d'ordre naturelle entre les $K+1$ catégories, $K \geq 2$. Etant donnés p régresseurs, ce modèle requiert l'estimation de K vecteurs de paramètres $\boldsymbol{\beta}_k \in \mathbb{R}^p$, $1 \leq k \leq K$. Nous proposons deux approches d'estimation, minimisant chacune un critère pénalisé par la norme ℓ_1 des paramètres, et visant à tirer profit de la parcimonie éventuelle au sein de chacun des $\boldsymbol{\beta}_k$ d'une part, et de l'homogénéité éventuelle entre les différents $\boldsymbol{\beta}_k$ d'autre part. Nos deux approches sont directement implémentables à l'aide de packages disponibles sous `R`. Nous les comparons empiriquement sur des jeux de données simulées. Nous proposons également une illustration pour étudier les facteurs de risque de différents sous-types de cancer du sein.

Mots-clés. Régression logistique multinomiale, régression logistique polytomique, régression logistique conditionnelle, Lasso, pénalité ℓ_1, analyse stratifiée.

Abstract. We consider parametric estimation under the polytomous, or multinomial, logistic regression model. This model is standard when the response variable is categorical, especially when the $K+1$, $K \geq 2$, categories can not be naturally ordered. Given p predictors, this model requires the estimation of K parameter vectors $\boldsymbol{\beta}_k \in \mathbb{R}^p$, $1 \leq k \leq K$. We propose two approaches based on ℓ_1-penalized criteria. Our aim is two-fold: to take advantage from (i) the potential sparsity among each of the $\boldsymbol{\beta}_k$ and (ii) the potential homogeneity between the various $\boldsymbol{\beta}_k$'s. Our two approaches can be implemented using available packages in `R`. We compare them empirically on synthetic data. We further illustrate these approaches on breast cancer data, where the objective is to study risk factors for various breast cancer histological sub-types.

Keywords. Multinomial logistic regression, polytomous logistic regression, conditional logistic regression, Lasso, ℓ_1-penalization, stratified analysis.

Estimation structurée par pénalisation L_1 dans le modèle de régression logistique polytomique.

Vivian Viallon & Cédric Garcia

Univ. Lyon, Université Claude Bernard Lyon1, Ifsttar, UMRESTTE, UMR T_9405, F- 69373, LYON.
vivian.viallon@univ-lyon.fr

1 Introduction : motivations et modèle

Ce travail se place dans le cadre de l'estimation et de la sélection de paramètres dans le modèle de régression logistique polytomique, ou multinomial. Plus précisément, on suppose disposer d'un échantillon $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, $n \geq 1$, où $\mathbf{X}_i \in \mathbb{R}^p$ décrit un ensemble de covariables, et $Y_i \in \{0, 1, \ldots, K\}$ est une variable réponse catégorielle à $K+1$ classes. Dans ce travail, nous nous plaçons dans le cadre général où aucune relation d'ordre naturelle n'existe entre les $K+1$ classes de la variable Y.

Un exemple d'application typique concerne le cancer du sein pour lequel il existe différents sous-types histologiques, définis notamment en fonction de la présence de certains récepteurs sur les cellules cancéreuses. Ainsi, dès lors que l'information sur le sous-type de cancer est disponible, la variable réponse Y n'est plus binaire, mais catégorielle : $Y = 0$ si le patient est sain, et $Y = k \in \{1, \ldots, K\}$ s'il est porteur du type k de cancer (sans relation d'ordre naturelle entre ces différents types).

Pour simplifier les notations, nous travaillons ici dans un modèle sans intercept. Le modèle de régression logistique polytomique suppose alors l'existence de K vecteurs de paramètres $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \mathbb{R}^p$ tels que

$$\log \left\{ \frac{\mathbb{P}(Y = k | \mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = 0 | \mathbf{X} = \mathbf{x})} \right\} = \boldsymbol{\beta}_k^T \mathbf{x}.$$

Notons que le choix de la catégorie de référence est arbitraire. Dans l'exemple du cancer du sein, il semble toutefois naturel de considérer la catégorie "sain", et donc l'évènement $\{Y = 0\}$, comme référence. Comme dans le cas de la régression logistique classique, les paramètres $\beta_{k,j}$ s'interprètent alors comme des (log)-odds-ratios ajustés, mesurant comment l'odds du type k de la maladie, $\mathbb{P}(Y = k | \mathbf{X} = \mathbf{x}) / \mathbb{P}(Y = 0 | \mathbf{X} = \mathbf{x})$, varie lorsque la j-ème covariable varie dans la population, les autres covariables étant fixées.

Les sous-types de cancer étant différents, on s'attend à ce que les vecteurs $\boldsymbol{\beta}_k$ soient globalement différents. Cependant, on peut également s'attendre à une certaine homogénéité dans le vecteur $\boldsymbol{\beta}^{(j)} = (\beta_{1,j}, \ldots, \beta_{K,j})^T$, pour certaines covariables. D'autre part, l'identification des hétérogénéités dans ce même vecteur revêt dans ce contexte un

intérêt tout particulier puisqu'elle peut suggérer des différences étiologiques entre les sous-types.

Notre objectif est donc double : (i) tirer profit de l'homogénéité attendue au sein des vecteurs $\boldsymbol{\beta}^{(j)}$, pour $j = 1, \ldots, p$, et ce afin d'améliorer la qualité de l'estimation, et (ii) identifier les hétérogénéités au sein de ces mêmes vecteurs. Notre proposition repose sur une idée analogue à celle proposée par Ollier et Viallon (2017) et Gross et Tibshirani (2016) pour l'estimation de modèles de régression linéaires généralisés sur données stratifiées. Plus précisément, elle repose sur la reparamétrisation $\beta_{k,j} = \tilde{\beta}_j + \delta_{k,j}$, où le paramètre $\tilde{\beta}_j$ correspond à l'effet global de la j-ème covariable, pour l'ensemble des sous-types de cancer, et $\delta_{k,j}$ correspond à la variation autour de cet effet global pour le sous-type k. Bien sûr, cette décomposition n'est pas unique. Mais l'utilisation de pénalité adaptée va nous permettre de spécifier une décomposition intéressante, et d'effectuer l'estimation sous cette réécriture sur-paramétrée du modèle. En particulier, la pénalisation des termes $|\delta_{k,j}|$ permet à notre approche de tirer profit de l'homogénéité attendue tout en identifiant les hétérogénéités.

2 Rappels : estimation des paramètres sous le modèle de régression logistique polytomique

Pour tout vecteur de paramètres $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{pK}$, soit $L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ la log-vraisemblance associée à l'échantillon $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. Avec $\boldsymbol{\beta}_0 = \mathbf{0}_p$ le vecteur nul de $\mathbb{R}^p$, il vient

$$\begin{aligned} L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) &= \sum_{i=1}^{n} \boldsymbol{\beta}_{Y_i}^T \mathbf{X}_i - \log \left\{ \sum_{k=0}^{K} \exp(\boldsymbol{\beta}_k^T \mathbf{X}_i) \right\} \\ &= \sum_{i:Y_i \neq 0} \boldsymbol{\beta}_{Y_i}^T \mathbf{X}_i - \sum_{i=1}^{n} \log \left\{ 1 + \sum_{k=1}^{K} \exp(\boldsymbol{\beta}_k^T \mathbf{X}_i) \right\}. \end{aligned}$$

Soit $\mathbb{1}[\cdot]$ la fonction indicatrice. Pour tout $1 \leq k \leq K$, soit $\mathcal{Y}^{(k)} \in \mathbb{R}^n$ tel que $\mathcal{Y}_i^{(k)} = \mathbb{1}[Y_i = k]$, pour tout $1 \leq i \leq n$. Considérons maintenant le vecteur de variables binaires $\mathcal{Y} \in \mathbb{R}^{nK}$ et la matrice $\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{nK \times Kp}$ définis par

$$\mathcal{Y} = \begin{pmatrix} \mathcal{Y}^{(1)} \\ \vdots \\ \mathcal{Y}^{(K)} \end{pmatrix} \quad \text{et} \quad \boldsymbol{\mathcal{X}} = \begin{pmatrix} \mathbf{X} & \ldots & \mathbf{0}_{n,p} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{n,p} & \ldots & \mathbf{X} \end{pmatrix},$$

où $\mathbf{X}$ est la matrice $n \times p$ contenant les n observations des régresseurs. Posons enfin $\mathbf{J} = (\mathbf{I}_n, \ldots, \mathbf{I}_n)$ la matrice de taille $n \times nK$, dont chacun des K blocs est la matrice identité d'ordre n. En posant $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, la log-vraisemblance L s'écrit alors,

$$L(\boldsymbol{\beta}) = L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \mathcal{Y}^T \boldsymbol{\mathcal{X}} \boldsymbol{\beta} - \left\{ \log \left(\mathbf{1}_n + \mathbf{J} \exp(\boldsymbol{\mathcal{X}} \boldsymbol{\beta}) \right) \right\}^T \mathbf{1}_n. \tag{1}$$

Dans cette expression, pour toute matrice A, $\exp(A)$ [resp. $\log(A)$] désigne la matrice de même dimension que A dont chaque élément vaut l'exponentiel [resp. le logarithme] de l'élément correspondant de A. Le vecteur $\mathbf{1}_n$ correspond quant à lui au vecteur de $\mathbb{R}^n$ donc chaque composante vaut 1.

Becg et Gray (1984) ont établi la consistance d'une approche alternative pour estimer les paramètres du modèle de régression polytomique. Elle consiste à réaliser K régressions logistiques (binaires) séparément, en considérant, pour chacune de ces régressions, uniquement les observations telles que $Y_i = k$ ou $Y_i = 0$. Pour tout $0 \leq k \leq K$, notons $\mathcal{I}_k = \{i : Y_i = k\}$ et $n_k = |\mathcal{I}_k|$ le cardinal de $\mathcal{I}_k$. Pour tout $1 \leq k \leq K$, on introduit alors le vecteur $\bar{\mathcal{Y}}^{(k)} = \mathcal{Y}^{(k)}_{\mathcal{I}_k \cup \mathcal{I}_0}$, c'est-à-dire le vecteur de taille $\bar{n}_k = |\mathcal{I}_k| + |\mathcal{I}_0|$ constitué des composantes de $\mathcal{Y}^{(k)}$ dont les indices sont dans $\mathcal{I}_k \cup \mathcal{I}_0$. Autrement dit, le vecteur $\bar{\mathcal{Y}}^{(k)}$ contient les valeurs $\mathbb{1}[Y_i = k]$ pour les seules observations i telles que $Y_i \in \{0, k\}$. De même, on définit pour tout $1 \leq k \leq K$, $\mathbf{X}^{(k)} = {}_{\mathcal{I}_k \cup \mathcal{I}_0}\mathbf{X}$, la matrice de taille $\bar{n}_k \times p$, constituée des lignes de la matrice $\mathbf{X}$ correspondantes aux observations i telles que $Y_i \in \{0, k\}$. Pour tout vecteur $\boldsymbol{\beta}_k \in \mathbb{R}^p$, la vraisemblance en jeu dans la k-ème régression logistique peut alors s'écrire

$$\begin{aligned} L_k(\boldsymbol{\beta}_k) &= \sum_{i:\bar{\mathcal{Y}}_i^{(k)}=1} \boldsymbol{\beta}_k^T \mathbf{X}_i^{(k)} - \sum_{i=1}^{\bar{n}_k} \log\{1 + \exp(\boldsymbol{\beta}_k^T \mathbf{X}_i^{(k)})\} \\ &= \sum_{i=1}^{\bar{n}_k} \left[\bar{\mathcal{Y}}_i^{(k)} \boldsymbol{\beta}_k^T \mathbf{X}_i^{(k)} - \log\{1 + \exp(\boldsymbol{\beta}_k^T \mathbf{X}_i^{(k)})\} \right] \\ &= (\bar{\mathcal{Y}}^{(k)})^T \mathbf{X}^{(k)} \boldsymbol{\beta}_k - [\log\{\mathbf{1}_{\bar{n}_k} + \exp(\mathbf{X}^{(k)} \boldsymbol{\beta}_k)\}]^T \mathbf{1}_{\bar{n}_k}. \end{aligned}$$

Soit $\bar{N} = \sum_{k=1}^K \bar{n}_k = Kn_0 + \sum_{k=1}^K n_k$. Introduisons maintenant le vecteur $\bar{\mathcal{Y}} \in \mathbb{R}^{\bar{N}}$ et la matrice $\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{\bar{N} \times p}$ définis par

$$\bar{\mathcal{Y}} = \begin{pmatrix} \bar{\mathcal{Y}}^{(1)} \\ \vdots \\ \bar{\mathcal{Y}}^{(K)} \end{pmatrix} \quad \text{et} \quad \bar{\boldsymbol{\mathcal{X}}} = \begin{pmatrix} \mathbf{X}^{(1)} & \dots & \mathbf{0}_{n,p} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{n,p} & \dots & \mathbf{X}^{(K)} \end{pmatrix}.$$

Observons que la maximisation séparée des K vraisemblances $L_1, \dots, L_K$ revient à la maximisation globale de la vraisemblance suivante, avec $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_K^T)^T$,

$$\bar{L}(\boldsymbol{\beta}) = \sum_{k=1}^K L_k(\boldsymbol{\beta}_k) = \bar{\mathcal{Y}}^T \bar{\boldsymbol{\mathcal{X}}} \boldsymbol{\beta} - \{\log(\mathbf{1}_{\bar{N}} + \exp(\bar{\boldsymbol{\mathcal{X}}} \boldsymbol{\beta}))\}^T \mathbf{1}_{\bar{N}}. \tag{2}$$

Estimation structurée par une approche pénalisée

Des versions pénalisées des vraisemblances L et $\bar{L}$ définies en (1) et (2), par exemple par la norme ℓ_1 du vecteur de paramètre $\boldsymbol{\beta} \in \mathbb{R}^{Kp}$, peuvent être envisagées dans un contexte de

grande dimension (*e.g.*, si $Kp \gg n$). Ces approches renvoient des estimations typiquement creuses pour chacun des vecteurs $\boldsymbol{\beta}_k$, mais ne tirent pas profit de l'homogénéité attendue entre ceux-ci. D'autre part, pour tout j fixé, les estimations $(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$ obtenues sont différentes par construction et les hétérogénéités ne peuvent donc pas être identifiées.

Notre proposition consiste à travailler sous la décomposition $\boldsymbol{\beta}_k = \tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}_k$ et à estimer les $(K+1)$ vecteurs $\tilde{\boldsymbol{\beta}}, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$, tout en pénalisant par une version pondérée de la norme ℓ_1 du vecteur résultant, de dimension $(K+1)p$. Par exemple, en partant de la vraisemblance sous le modèle polytomique, notre méthode renvoie des estimations $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K$ telles que $\widehat{\boldsymbol{\beta}}_k = \widehat{\tilde{\boldsymbol{\beta}}} + \widehat{\boldsymbol{\delta}}_k$, où $\widehat{\tilde{\boldsymbol{\beta}}}, \widehat{\boldsymbol{\delta}}_1, \ldots, \widehat{\boldsymbol{\delta}}_K$ sont les solutions, typiquement creuses, maximisant la fonction objectif suivante

$$L_\lambda(\tilde{\boldsymbol{\beta}}, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K) = \sum_{i:Y_i \neq 0} (\tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}_{Y_i})^T \mathbf{X}_i - \sum_{i=1}^{n} \log \left[1 + \sum_{k=1}^{K} \exp\{(\tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}_{Y_i})^T \mathbf{X}_i\} \right] - \lambda_1 \|\tilde{\boldsymbol{\beta}}\|_1 - \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{\delta}_k\|_1,$$

où $\lambda = (\lambda_1, \lambda_2)$. Dans le cas où l'on utilise l'approche reposant sur les K régressions logistiques séparées, notre approche renvoie des estimations $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K$ telles que $\widehat{\boldsymbol{\beta}}_k = \widehat{\tilde{\boldsymbol{\beta}}} + \widehat{\boldsymbol{\delta}}_k$, où $\widehat{\tilde{\boldsymbol{\beta}}}, \widehat{\boldsymbol{\delta}}_1, \ldots, \widehat{\boldsymbol{\delta}}_K$ sont les solutions, typiquement creuses, maximisant la fonction objectif suivante

$$\bar{L}_\lambda(\tilde{\boldsymbol{\beta}}, \boldsymbol{\delta}_1 \ldots, \boldsymbol{\delta}_K) = \sum_{k=1}^{K} \Big[\sum_{i:\bar{\mathcal{Y}}_i^{(k)}=1} (\tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}_k)^T \mathbf{X}_i^{(k)} - \sum_{i=1}^{\bar{n}_k} \log\{1 + \exp((\tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}_k)^T \mathbf{X}_i^{(k)})\} \Big] - \lambda_1 \|\tilde{\boldsymbol{\beta}}\|_1 - \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{\delta}_k\|_1,$$

Soit $\tau = \lambda_2/\lambda_1$. Introduisons les matrices

$$\boldsymbol{\mathcal{X}}_\tau = \begin{pmatrix} \mathbf{X} & \frac{\mathbf{X}}{\tau} & \ldots & \mathbf{0}_{n,p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X} & \mathbf{0}_{n,p} & \ldots & \frac{\mathbf{X}}{\tau} \end{pmatrix} \quad \text{et} \quad \bar{\boldsymbol{\mathcal{X}}}_\tau = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau} & \ldots & \mathbf{0}_{n,p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0}_{n,p} & \ldots & \frac{\mathbf{X}^{(K)}}{\tau} \end{pmatrix}.$$

En posant $\boldsymbol{\theta}_\tau = (\tilde{\boldsymbol{\beta}}, \tau\boldsymbol{\delta}_1, \ldots, \tau\boldsymbol{\delta}_K)$, de dimension $(K+1)p$, on a alors

$$L_\lambda(\tilde{\boldsymbol{\beta}}, \boldsymbol{\delta}_1 \ldots, \boldsymbol{\delta}_K) = L_\lambda(\boldsymbol{\theta}_\tau) = \mathcal{Y}^T \boldsymbol{\mathcal{X}} \boldsymbol{\theta}_\tau - \{\log(\mathbf{1}_n + \mathbf{J}\exp(\boldsymbol{\mathcal{X}}_\tau \boldsymbol{\theta}_\tau))\}^T \mathbf{1}_n - \lambda_1 \|\boldsymbol{\theta}_\tau\|_1 \quad (3)$$

$$\bar{L}_\lambda(\tilde{\boldsymbol{\beta}}, \boldsymbol{\delta}_1 \ldots, \boldsymbol{\delta}_K) = \bar{L}_\lambda(\boldsymbol{\theta}_\tau) = \bar{\mathcal{Y}}^T \bar{\boldsymbol{\mathcal{X}}}_\tau \boldsymbol{\theta}_\tau - \{\log(\mathbf{1}_{\bar{N}} + \exp(\bar{\boldsymbol{\mathcal{X}}}_\tau \boldsymbol{\theta}_\tau))\}^T \mathbf{1}_{\bar{N}} - \lambda_1 \|\boldsymbol{\theta}_\tau\|_1. \quad (4)$$

Ces dernières équations établissent que nos deux approches peuvent être directement implémentées à partir d'algorithmes de résolution du lasso, dans les modèles de régression logistique binaire (pour la vraisemblance pénalisée $\bar{L}_\lambda$), ou dans les modèles de régression logistique polytomique (pour la vraisemblance pénalisée L_λ).

Nous comparons ces deux approches, en terme de performances des estimations (erreur d'estimation, précision de l'identification du support et des hétérogénéités) mais aussi en terme de temps de calcul, sur des données simulées. Afin d'illustrer l'intérêt de la prise en compte de la structure attendue au sein des vecteurs $\boldsymbol{\beta}_k$, nous incluons également les résultats issus d'approche ne cherchant pas à tirer profit de cette structure attendue.

Bibliographie

[1] Becg, C. B. et Gray, R. (1984), Calculation of polychotomous logistic regression parameters using individualized regressions, *Biometrika*, 71, 11–16.
[2] Gross, S. et Tibshirani, R. (2016), Data Shared Lasso: A novel tool to discover uplift, *Computational Statistics and Data Analysis*,
[3] Ollier, E. et Viallon, V. (2017), Regression modelling on stratified data with the lasso, *Biometrika*, à paraître.