Par convention, les vecteurs sont des vecteurs colonne. Pour un vecteur u, on note u^T la transposée et $\|u\|$ sa norme euclidienne.

Résultat admis Soit A une matrice symétrique réelle $p \times p$. Nous notons par $\lambda_{\min}(A)$ la valeur propre minimale de la matrice A ; et par $|||A|||$ sa norme opérateur

$$|||A||| := \sup_{\|x\| \leq 1} \|Ax\| .$$

On admettra que si $\{A_n,\ n \in \mathbb{N}\}$ est une suite de matrices réelles symétriques telle que $\lim_{n\to\infty} |||A_n - A||| = 0$ où A est une matrice symétrique réelle, alors $\lim_{n\to\infty} \lambda_{\min}(A_n) = \lambda_{\min}(A)$.

Problème

Considérons la suite d'expériences statistiques

$$\left(\{0,1\}^n, \mathscr{P}(\{0,1\}^n), \left\{ \mathbb{P}_{n,\theta} = \bigotimes_{i=1}^{n} \mathrm{Ber}\left(\varphi(\theta^T \mathbf{x}_i)\right) \ :\ \theta \in \mathbb{R}^p \right\} \right)$$

où $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ sont des vecteurs de variables explicatives ($\mathbf{x}_i \in \mathbb{R}^p$) et

$$\varphi(t) := \frac{1}{1+\mathrm{e}^{-t}} = \frac{\mathrm{e}^t}{1+\mathrm{e}^t}.$$

On notera $\{Y_k, 1 \leq k \leq n\}$ les observations, et on introduit les notations matricielles

$$\mathbf{X}_n := \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \quad \mathbf{Y}_n := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \Phi_n(\theta) := \begin{bmatrix} \varphi(\theta^T \mathbf{x}_1) \\ \vdots \\ \varphi(\theta^T \mathbf{x}_n) \end{bmatrix}.$$

Enfin, on définit les fonctions $h : \mathbb{R} \to [0,1]$ et $\mathbf{F}_n : \mathbb{R}^p \to \mathbb{R}^{p\times p}$ par

$$h := \varphi(1-\varphi), \quad \mathbf{F}_n(\theta) := \sum_{i=1}^{n} h(\theta^T \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^T. \tag{1}$$

On suppose dans la suite que $\mathbf{X}_n$ est de rang p.

1. Montrer que le modèle est identifiable.
2. Montrer que pour tout $\theta \in \mathbb{R}^p$, $\mathbf{F}_n(\theta)$ est définie positive.
3. Montrer que h est 1-Lipschitzienne sur $\mathbb{R}$.
4. Déterminer les fonctions de vraisemblance et de log-vraisemblance des observations ; qu'on notera respectivement $\theta \mapsto \mathrm{L}_n(\theta)$ et $\theta \mapsto \ell_n(\theta)$. *Noter que l'on a toujours $\ell_n(\theta) \leq 0$ comme le logarithme d'un nombre compris entre 0 et 1.*
5. Montrer que, pour tout $\theta \in \mathbb{R}^p$,

$$\begin{aligned} &\nabla \ell_n(\theta) = \sum_{i=1}^{n} \{Y_i - \varphi(\theta^T \mathbf{x}_i)\} \mathbf{x}_i = \mathbf{X}_n^T \left\{ \mathbf{Y}_n - \Phi_n(\theta) \right\} , \\ &\nabla^2 \ell_n(\theta) = -\mathbf{F}_n(\theta) , \\ &\mathbb{E}_{n,\theta} \left[\nabla \ell_n(\theta) \nabla \ell_n(\theta)^T \right] = \mathbf{F}_n(\theta) . \end{aligned}$$

   En déduire que (presque surement) la fonction $\theta \mapsto \ell_n(\theta)$ est strictement concave.

► Nous allons maintenant étudier les conditions sous lesquelles l'estimateur du maximum de vraisemblance existe (au sens où il existe $\hat{\theta}_n^{\mathrm{MV}} \in \mathbb{R}^p$ tel que $\ell_n(\hat{\theta}_n^{\mathrm{MV}}) = \sup_{\theta\in\mathbb{R}^p} \ell_n(\theta)$). On distingue trois cas, qui, dans le cas où $p = 2$, peuvent être illustrés par la Figure 1.

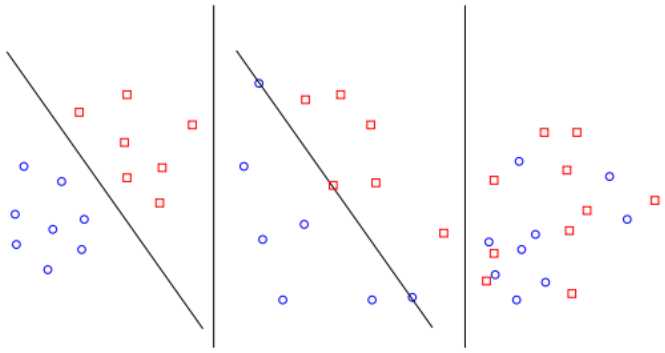

FIGURE 1 – De gauche à droite : séparabilité complète, séparabilité quasi-complète, et recouvrement. Les ronds et carrés représentent les n points $\mathbf{x}_k$ dans $\mathbb{R}^2$; on a représenté des ronds bleus lorsque l'observation Y_k associée vaut 0, et des carrés rouges lorsqu'elle vaut 1.

6. Supposons tout d'abord que les données sont *complètement séparables*, i.e. il existe $\theta_\star \neq 0_{\mathbb{R}^p}$ tel que pour tout $k \in \{1,\cdots,n\}$,

$$\begin{cases} \theta_\star^T \mathbf{x}_k > 0, & \text{si } Y_k = 1 \\ \theta_\star^T \mathbf{x}_k < 0, & \text{si } Y_k = 0. \end{cases}$$

   Montrer que $\lim_{\lambda\to+\infty} \ell_n(\lambda\theta_\star) = 0$ et en déduire que l'estimateur du maximum de vraisemblance n'existe pas dans ce cas.

7. Supposons maintenant que les données sont *quasi-séparables*, i.e. il existe $\theta_\star \neq 0_{\mathbb{R}^p}$ et $\mathscr{E} \subset \{1,\ldots,n\}$ tels que pour tout $k \in \{1,\cdots,n\}\setminus\mathscr{E}$,

$$\begin{cases} \theta_\star^T \mathbf{x}_k > 0, & \text{si } Y_k = 1 \\ \theta_\star^T \mathbf{x}_k < 0, & \text{si } Y_k = 0 \end{cases}$$

   et pour tout $k \in \mathscr{E}$, $\theta_\star^T \mathbf{x}_k = 0$. Montrer que le maximum de vraisemblance n'existe pas dans ce cas. *Indication : pour ce faire, on pourra évaluer la vraisemblance en* $\theta_\lambda := \lambda\theta_\star + (\bar{\theta} - \theta_\star)$ *pour* $\bar{\theta} \in \mathbb{R}^p$ *fixé, et montrer que la vraisemblance* L_n *est le produit d'un terme constant en* λ *et d'un terme strictement croissant quand* $\lambda \to \infty$.

8. Supposons maintenant l'existence d'un *recouvrement*, i.e. pour tout $\theta \in \mathbb{R}^p \setminus 0_{\mathbb{R}^p}$, il existe k_1, k_2 tels que $Y_{k_1} = Y_{k_2}$ et $\theta^T \mathbf{x}_{k_1} > 0$, $\theta^T \mathbf{x}_{k_2} < 0$. Montrer que l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$ existe et est unique.
   Indication : Pour θ *sur la sphère* $\mathscr{S}(0,1) := \{\theta : \|\theta\| = 1\}$, *on note* $k_{1,\theta} := \arg\max_{k\in\{1,\cdots,n\}} \theta^T \mathbf{x}_k$, $k_{2,\theta} := \arg\min_{k\in\{1,\cdots,n\}} \theta^T \mathbf{x}_k$ *(en cas d'égalité, on choisit l'indice le plus petit). Montrer qu'il existe* $\zeta > 0$ *tel que pour tout* $\theta \in \mathscr{S}(0,1)$, *on a* $\theta^T \mathbf{x}_{k_{1,\theta}} > \zeta$ *et* $\theta^T \mathbf{x}_{k_{2,\theta}} < -\zeta$. *En déduire que pour tout* $M > 0$, *il existe* λ_M *tel que pour tout* $\theta \in \mathscr{S}(0,1)$, *et tout* $\lambda > \lambda_M$, $\ell_n(\lambda\theta) \leq -M$; *puis la limite de* ℓ_n *en l'infini ; puis exploiter la stricte concavité de la fonction* ℓ_n.

Nous supposerons dans toute la suite que l'hypothèse de recouvrement est satisfaite pour tout $n \geq n_0$.

Méthode numérique pour le calcul de $\hat{\theta}_n^{\mathrm{MV}}$. $\hat{\theta}_n^{\mathrm{MV}}$ est l'unique racine de l'équation $\nabla \ell_n = 0$. Il est possible de résoudre cette équation en utilisant un algorithme de Newton-Raphson : partant d'une valeur θ^t, la mise à jour se fait par

$$\begin{aligned}\theta^{t+1} &= \theta^t - \{\nabla^2 \ell_n(\theta^t)\}^{-1} \nabla \ell_n(\theta^t) = \theta^t + \left(\mathbf{X}_n^T \mathbf{W}(\theta^t) \mathbf{X}_n\right)^{-1} \mathbf{X}_n^T \left(\mathbf{Y}_n - \Phi_n(\theta^t)\right), \\ &= (\mathbf{X}_n^T \mathbf{W}(\theta^t) \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{W}(\theta^t) \mathbf{z}_t \ ,\end{aligned}$$

en ayant posé

$$\mathbf{W}(\theta) := \mathrm{diag}\left(h(\theta^T \mathbf{x}_1), \cdots h(\theta^T \mathbf{x}_n)\right), \qquad \mathbf{z}_t := \mathbf{X}_n \theta^t + \{\mathbf{W}(\theta^t)\}^{-1} (\mathbf{Y}_n - \Phi_n(\theta^t)).$$

Par analogie avec la formule de l'estimateur des Moindres Carrés en régression linéaire, on voit que θ^{t+1} est solution du problème

$$\theta^{t+1} = \underset{\theta \in \mathbb{R}^p}{\arg\min} (\mathbf{z}_t - \mathbf{X}_n \theta)^T \mathbf{W}(\theta^t) (\mathbf{z}_t - \mathbf{X}_n \theta) \ .$$

On appelle parfois $\mathbf{z}_t$ la réponse ajustée. De plus, comme la mise à jour est solution d'un problème quadratique, cet algorithme est connu sous le nom de IRLS pour *Iteratively Reweighted least-Squares*.

input : θ^0, $t = 0$: initial point
$\mathbf{X}_n$, $\mathbf{Y}_n$: covariates, binary observations
$\rho > 0$
output: approximation of $\hat{\theta}_n^{\mathrm{MV}}$

Compute $\Phi_n(\theta^0)$ and the diagonal matrix $\mathbf{W}(\theta^0)$.
$t \leftarrow 0$.
while $\left\|\mathbf{X}_n^T (\mathbf{Y}_n - \Phi_n(\theta^t))\right\| \geq \rho$ **do**
  $\mathbf{z}_t = \mathbf{X}_n \theta^t + \{\mathbf{W}(\theta^t)\}^{-1} (\mathbf{Y}_n - \Phi_n(\theta^t))$
  $\theta^{t+1} = (\mathbf{X}_n^T \mathbf{W}(\theta^t) \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{W}(\theta^t) \mathbf{z}_t$
  Compute $\Phi_n(\theta^{t+1})$ and the diagonal matrix $\mathbf{W}(\theta^{t+1})$.
  $t \leftarrow t + 1$
end
Return θ^t

Algorithm 1: Algorithme IRLS pour la résolution numérique des équations $\nabla \ell_n = 0$.

Nous considérons les hypothèses suivantes :

H1. Pour tout $\theta \in \mathbb{R}^p$, il existe une matrice $Q(\theta)$ définie positive telle que

$$\lim_{n \to \infty} |||n^{-1} \mathbf{F}_n(\theta) - Q(\theta)||| = 0 \ .$$

H2. $\sup_{n \geq 0} n^{-1} \sum_{i=1}^n \|\mathbf{x}_i\|^3 < \infty$.

Nous admettons que sous ces hypothèses la suite $\{\hat{\theta}_n^{\mathrm{MV}}\}_{n=n_0}^{\infty}$ est consistante Nous allons établir la normalité asymptotique : pour tout $\theta \in \mathbb{R}^p$,

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{MV}} - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \{Q(\theta)\}^{-1}).$$

9. Montrer qu'il existe une constante C telle que pour tout $\theta, \vartheta \in \mathbb{R}^p$, et tout $n > 0$,

$$\|\mathbf{F}_n(\theta) - \mathbf{F}_n(\vartheta)\| \leq Cn\|\theta - \vartheta\|.$$

10. Montrer que pour tout $\theta \in \mathbb{R}^p$, $n \geq n_0$,

$$\frac{\nabla \ell_n(\hat{\theta}_n^{\mathrm{MV}}) - \nabla \ell_n(\theta)}{\sqrt{n}} = \left(\frac{-\mathbf{F}_n(\theta)}{n} + R_n\right) \sqrt{n}(\hat{\theta}_n - \theta)$$

où $R_n \overset{\mathbb{P}_{n,\theta}-\text{prob}}{\longrightarrow} 0$. *Pour la suite, on remarquera que* $\nabla \ell_n(\hat{\theta}_n^{\mathrm{MV}}) = 0$.

11. Montrer que $n^{-1/2}\nabla \ell_n(\theta)$ est asymptotiquement normal. *Indication : utiliser l'expression de* ℓ_n *donnée par la question 5 puis appliquer le théorème de Lindeberg-Feller (Théorème IV-2.41 du polycopié).*

12. Conclure que $\sqrt{n}(\hat{\theta}_n^{\mathrm{MV}} - \theta) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0, \{Q(\theta)\}^{-1})$.

▶ Enfin, nous mettons en oeuvre un test asymptotique sur les coefficients de régression. Soit $k \in \{1, \cdots, p\}$ et $\theta = (\theta_1, \cdots, \theta_p) \in \mathbb{R}^p$.

13. On note $\beta_{n,k}$ le $k$ème coefficient de la diagonale de $\{\mathbf{F}_n(\hat{\theta}_n^{\mathrm{MV}})/n\}^{-1}$. Montrer que

$$\sqrt{\frac{n}{\beta_{n,k}}}\,(\hat{\theta}_{n,k}^{\mathrm{MV}} - \theta_k) \overset{\mathbb{P}_{n,\theta}}{\Longrightarrow} \mathrm{N}(0,1)\,.$$

14. Construire un intervalle de confiance asymptotique pour le paramètre θ_k.
15. En déduire un test de niveau asymptotique α pour $H_0 \ : \ \theta_k = 0$ contre $H_1 \ : \ \theta_k \neq 0$.
16. Calculer la p-valeur asymptotique de ce test.

Application numérique

Le modèle de régression logistique fait partie d'une classe générale de modèles appelés modèles log-linéaires. Ces modèles sont particulièrement utiles pour l'étude des tableaux de contingence (tableaux de comptes). De tels tableaux se produisent lorsque les observations sont croisées à l'aide de plusieurs variables catégorielles (les tableaux de contingence sont parfois appelés classifications croisées). Le formalisme de la régression logistique est alors approprié si l'une des variables catégorielles prend deux valeurs et peut être considérée comme une variable cible. Par exemple, dans les essais cliniques, le fait de savoir si le patient vit ou meurt est une variable cible, et différentes variables catégorielles pourraient être des prédicteurs potentiels (par exemple, le sexe, l'appartenance au groupe de traitement ou de contrôle, la présence ou l'absence de certains symptômes, etc.

Voici un exemple de tableau de contingence de cette forme. L'une des catastrophes maritimes les plus célèbres s'est produite lors du voyage inaugural du paquebot Titanic, qui a heurté un iceberg dans l'Atlantique Nord et a coulé le 15 avril 1912. De nombreux articles, livres et films ont raconté l'histoire de ce désastre, mais une analyse statistique relativement simple raconte l'histoire d'une manière particulièrement évocatrice. Les données sont dans le fichier titanic.csv. Des packages existants implémentant la régression logistique et les tests sont `glm` sur R, et `statsmodels` sur python. Attention à la construction du modèle : l'intercept n'est pas inclus par défaut dans `statsmodels`, il faut donc penser à le rajouter au moment de la construction du modèle, ainsi que pour les nouvelles prédictions.

Chaque enregistrement de l'ensemble de données décrit un passager. Les attributs sont définis comme suit :

1. `PassengerId` : Identifiant du passager
2. `Survived` : indique si le passager est décédé ou a survécu (0 = décès, 1 = survivant)
3. `Pclass` : Classe du passager (1 = 1st ; 2 = 2nd ; 3 = 3rd)
4. `Name` : Nom du passager
5. `Sex` : Sexe du passager
6. `Age` : Age du passager
7. `SibSp` : Nombre de frères et sœurs et de conjoints à bord
8. `Parch` : Nombre de parents/enfants à bord
9. `Ticket` : Numéro du ticket
10. `Fare` : Tarif du ticket
11. `Cabin` : Cabine
12. `Embarked` : Port d'embarquement (C = Cherbourg ; Q = Queenstown ; S = Southampton)

On remarquera que l'âge du passager (ou le numéro de sa cabine) n'est pas toujours disponible. La présence de données manquantes est un problème que l'on rencontre très fréquemment en statistique. Nous éliminerons le numéro de cabine de l'analyse et remplacerons les valeurs de l'âge manquantes par la moyenne des âges des observations disponibles (en pratique, il faudrait faire des choses plus subtiles, mais nous n'entrerons pas dans ce niveau de détails). Avant de procéder à une analyse par régression logistique, nous allons tout d'abord procéder à une analyse préliminaire des données. On élimine de l'analyse les attributs `PassengerId`, `Ticket`, `Cabin` et `Embarked`. On traite `Sex` et `Pclass` comme des variables catégorielles (et on introduira donc des variables "muettes" ou dummy variables).

1. Déterminer le nombre et la proportion de passagers décédés.
2. Déterminer le pourcentage d'hommes et de femmes parmi les personnes décédées et les survivants. Qu'observe-t-on ?
3. Reprendre cette analyse pour les différentes classes. Qu'observe-t-on ?
4. Calculer la matrice de corrélation des covariables. Qu'observe-t-on ?

On procède maintenant à la régression logistique

5. On inclut d'abord tous les paramètres dans la régression. Calculer les intervalles de confiance à 95% pour l'ensemble des paramètres. Qu'observe-t-on ?
6. Déterminer la p-valeur du test

$$H_0 : \beta_{\texttt{Parch}} = 0, \quad \text{contre} \quad H_1 : \beta_{\texttt{Parch}} \neq 0$$

   Que peut-on conclure ?
7. On reprend l'analyse en éliminant la variable `Parch`. Calculer les intervalles de confiance à 95% pour l'ensemble des paramètres. Qu'observe-t-on ?
8. Déterminer la p-valeur du test

$$H_0 : \beta_{\texttt{Fare}} = 0, \quad \text{contre} \quad H_1 : \beta_{\texttt{Fare}} \neq 0$$

9. On reprend l'analyse en éliminant la variable `Parch` et `Fare`. Calculer les intervalles de confiance à 95 %. Que peut-on conclure ?
10. Dans ce modèle, quelles sont les probabilité de survie d'un homme dans un cas, et d'une femme dans un autre cas, chacun âgé de 22 ans sans famille et voyageant en 1ère classe ?