Autour des a prioris peu-informatifs dans les modèles bayésiens de régression logistique

Mickael Schaeffer [1] & François Lefebvre [2] & Erik Sauleau [3] & Nicolas Meyer [4]

[1] *Hôpital civil, Service de santé publique, Groupe méthode en recherche clinique, 1 place de l'hôpital, BP 426, 67091 Strasbourg, mickael.schaeffer@chru-strasbourg.fr*
[2] *Hôpital civil, Service de santé publique, Groupe méthode en recherche clinique, 1 place de l'hôpital, BP 426, 67091 Strasbourg, francois.lefebvre@chru-strasbourg.fr*
[3] *Laboratoire de Biostatistiques et d'informatique médicale, Université de Strasbourg, 4 rue Kirschleger - 67085 Strasbourg Cedex, ea.sauleau@unistra.fr*
[4] *Laboratoire de Biostatistiques et d'informatique médicale, Université de Strasbourg, 4 rue Kirschleger - 67085 Strasbourg Cedex, nmeyer@unistra.fr*

Résumé. L'estimation de la distribution associée au coefficient d'un modèle de régression logistique peut être effectuée par des méthodes bayésiennes. Dans ce cas, l'utilisation de lois a priori très peu informatives, comme les distributions Gaussiennes à variance large, peuvent parfois amener à des estimations biaisées ou à une surestimation de la variance a posteriori. Dans cet article, nous proposerons une variance à utiliser dans le cas d'un a priori peu-informatif, afin d'éviter toute surestimation a posteriori, et d'améliorer la qualité des prédictions. Nous établirons un lien entre la paramétrisation de deux distributions Beta a priori pour des proportions et la paramétrisation du coefficient associé à la comparaison dans un modèle de régression logistique. Cette paramétrisation est établie en utilisant une des propriétés de la régression logistique, à savoir l'égalité du coefficient de la régression avec le logarithme d'un rapport de cotes, c'est-à-dire un rapport de distribution *Beta*. Nous montrerons à l'aide de simulations que la distribution ainsi définie présente un gain en terme de variabilité estimée à posteriori.

Mots-clés. Régression logistique, analyses bayésiennes, a priori non-informatif, fonction de lien exponentielle, logit.

1 Introduction

La comparaison de deux proportions $p_i, i \in \{1; 2\}$ indépendantes est quotidiennement utilisée dans de nombreux contextes. Dans le cadre des analyses bayésiennes, il est d'usage de prendre pour chacune des proportions p_i des distributions a priori Beta de paramètres α_i et β_i traduisant les connaissances acquises précédemment ou un avis d'expert sous forme de pseudo-cas. On peut également utiliser la régression logistique qui est un modèle

permettant de comparer deux proportions : elle établit un lien direct avec le rapport de cotes associées aux deux proportions. La régression logistique est définie par :

$$\left\{\begin{array}{lcl} Y_i & \text{II} & \forall i \in \{1; N\} \\ Y & \sim & \mathcal{B}(1,p) \\ \mu & = & \mathbb{E}(g(p)) = \alpha + \beta \times Groupe \\ p & = & g^{-1}(\mu) = \frac{e^{\mu}}{1+e^{\mu}} \end{array}\right. \tag{1}$$

où N est le nombre d'observations de l'échantillon.

Nous nous intéressons ici au choix de la distribution a priori des coefficients d'une régression logistique simple bayésienne pour comparer deux proportions. En effet la combinaison de l'information a priori et de l'information contenue dans les données fournit, éventuellement au moyen de méthodes de Monte-Carlo par chaînes de Markov, une estimation de la distribution a posteriori de chacun des coefficients de la régression logistique. On rappelle ici le théorème du Bayes : en notant y la vraisemblance de l'échantillon, β le paramètre d'intérêt, le théorème de Bayes se présente de la façon suivante :

$$p(\beta|y) = \frac{p(\beta, y)}{p(y)} = \frac{p(y|\beta)p(\beta)}{p(y)} = \frac{p(y|\beta)p(\beta)}{\int p(y|\beta)p(\beta)d\beta} \propto p(y|\beta)p(\beta)$$

Sous cette modélisation logistique, le coefficient du prédicteur linéaire associé à la variable Groupe est un réel, défini sur $]-\infty; +\infty[$. La distribution a priori ainsi utilisée pour sa modélisation est habituellement une distribution $\pi(\beta) = \mathcal{N}(\mu, \sigma^2)$.
Nous avons souhaité établir un lien entre les paramètres de la distribution a priori de la régression logistique et les paramètres des distributions a priori Beta associées à ces proportions dans le cas de la modélisation de la différence des deux proportions. Notre objectif était donc de trouver la loi de probabilité a priori du coefficient d'une régression logistique permettant d'exprimer le même a priori que lors de l'utilisation de deux lois $\mathcal{B}(\alpha_i; \beta_i)$ pour comparer deux proportions.

2 Méthodes

Soit les proportions d'intérêts $p_i, i \in 1, 2$, telles que $p_1 \sim \mathcal{B}(a, b)$ et $p_2 \sim \mathcal{B}(c, d)$. Par ailleurs, les paramètres α et β de la régression logistique s'expriment de la façon suivante :

$$\left\{\begin{array}{lcl} \alpha & = & \log\left(\frac{p_2}{1-p_2}\right) \\ \beta & = & \log(OR) = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) \end{array}\right. \tag{2}$$

Ensuite, pour établir la correspondance entre les distributions a priori Beta des proportions p_i et la distribution du coefficient de la régression logistique, nous nous sommes intéressés à la distribution de $\log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$ lorsque $p_1 \sim \mathcal{B}(a, b)$ et $p_2 \sim \mathcal{B}(c, d)$.

L'étude de la distribution de $\beta = \log(OR)$ sous l'hypothèse de deux proportions issues de distributions Beta permet de montrer que :

$$f_\beta(h) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} e^{vd} \int_{\mathbb{R}} \frac{e^{u(a+c)}}{(1+e^u)^{(a+b)}(e^u+e^v)^{(c+d)}} du \tag{3}$$

qui, dans le cas de deux distributions uniformes a priori ($\mathcal{U}[0;1]$) sur les proportions ($a = b = c = d = 1$) s'écrit alors :

$$f_\beta(h) = e^h \int_{\mathbb{R}} \frac{e^{2u}}{(1+e^u)^2(e^u+e^h)^2} du \tag{4}$$

Cette distribution définie en (4) a une moyenne de 0 et une variance de 6,58 environ. On constate ainsi que la correspondance entre deux distributions uniformes sur $[0;1]$ (donc une situation d'équiprobabilité pour chaque point de l'intervalle) dans la paramétrisation Beta est une distribution de moyenne 0 et de variance 6,58 dans la modélisation logistique. Ainsi dans la situation d'un a priori vague, ou peu-informatif (uniforme), et dans le cas d'une modélisation logistique, on démontre qu'une distribution a priori doit être paramétrée avec une variance de moins de 6,58 points, afin d'éviter une sur-estimation de la variance a posteriori estimée par la modèle. Cela montre également que l'a priori usuellement utilisé par défaut pour le paramètre β de la régression logistique, à savoir une $\mathcal{N}(0, 1000)$ est trop peu informatif, ou en tout cas moins informatif que ce qu'exprime un couple de loi $\mathcal{B}(\alpha_i, \beta_i)$ uniformes pour les deux proportions.

Il est également possible de démontrer, sous l'hypothèse de distributions $Beta(a, b)$ et $Beta(c, d)$ sur p_1 et p_2, que le paramètre α de la régression logistique suit alors la distribution suivante :

$$f_\alpha(h) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} \frac{e^{h(a+c)}}{(1+e^h)^{(a+b)}} \int_{\mathbb{R}} \frac{e^{vd}}{(e^h+e^v)^{(c+d)}} dv \tag{5}$$

qui, dans le cas deux distributions uniformes ($a = b = c = d = 1$) :

$$f_\alpha(h) = \frac{e^{2h}}{(1+e^h)^2} \left(\frac{1}{e^h}\right) = \frac{e^h}{(1+e^h)^2} \tag{6}$$

On montre enfin que dans le cas de deux distributions uniformes sur p_1 et p_2, $Var(\beta) = 2 \times Var(\alpha)$. La variance de la distribution définie en (6) est alors d'environ 3,29.

3 Estimations a posteriori en utilisant $f_\alpha(h)$ et $f_\beta(h)$

Dans la section suivante nous notons $f_\alpha(h)$ et $f_\beta(h)$ les distributions associées aux logarithmes des rapports de lois uniformes définies en (4) et (6). Les estimations a posteriori

sont comparées avec celles d'une loi Normale utilisée habituellement en cas d'information a priori vague, i.e une $\mathcal{N}(0, 1000)$. On s'intéresse aux résultats en termes de biais et de variance estimée, pour différentes tailles d'échantillon. On a donc :

$$\left\{ \begin{array}{rcl} \pi_\beta(\beta|y) & \propto & \pi(y|\beta) f_\beta(h) \\ \pi_\alpha(\alpha|y) & \propto & \pi(y|\alpha) f_\alpha(h) \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{rcl} \pi_N(\beta|y) & \propto & \pi(y|\beta) \mathcal{N}(0, 10^2) \\ \pi_N(\alpha|y) & \propto & \pi(y|\alpha) \mathcal{N}(0, 10^2) \end{array} \right. \tag{7}$$

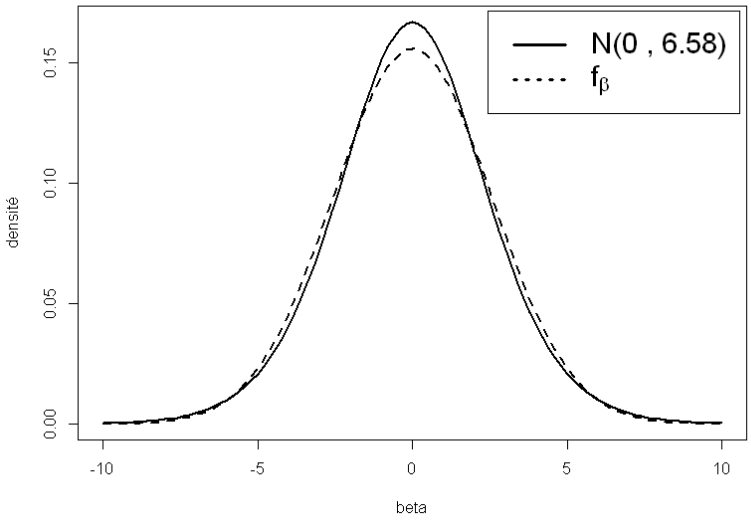
Les données ont été simulées pour des proportions p_1 et p_2 fixées mais des tailles d'échantillon différentes. Trois couples de proportions sont présentés dans le tableau ci-dessous : ($20\% - 40\%$, $20\% - 60\%$ et $40\% - 40\%$). Les distributions a posteriori ont été estimées en utilisant les méthodes de Monte-Carlo par chaînes de Markov. Des chaînes de 100 000 itérations ont été réalisées, avec une période de chauffe de 5000 itérations. Les résultats sont présentés dans le tableau (1). Les valeurs de α et de β sont calculées à partir de p_1 et p_2 comme mentionné dans la section précédente. Les estimations par maximum de vraisemblance des coefficients (le logarithme du rapport de cotes) sont également présentées dans la dernière colonne du tableau.

		Paramètre	$N = 5$		$N = 10$		$N = 20$		$N = 200$		Estimation
p_1	p_2	loi a priori	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	Max Vrais
0.2	0.4	α – $f_\alpha(h)$	-1.07	0.93	-1.22	0.72	-1.30	0.54	-1.38	0.18	-1.386
		β – $f_\beta(h)$	0.73	1.23	0.85	0.95	0.91	0.69	0.97	0.23	0.981
0.2	0.4	α – $N(0, 10^2)$	-1.80	1.35	-1.61	0.89	-1.52	0.59	-1.39	0.18	-1.386
		β – $N(0, 10^2)$	1.29	1.69	1.16	1.12	1.10	0.76	0.98	0.23	0.981
0.2	0.6	α – $f_\alpha(h)$	-1.09	0.92	-1.24	0.71	-1.29	0.53	-1.38	0.17	-1.386
		β – $f_\beta(h)$	1.40	1.24	1.60	0.94	1.67	0.69	1.78	0.23	1.792
0.2	0.6	α – $N(0, 10^2)$	-1.81	1.39	-1.56	0.87	-1.47	0.60	-1.40	0.18	-1.386
		β – $N(0, 10^2)$	2.31	1.72	2.00	1.11	1.89	0.76	1.81	0.23	1.792
0.4	0.4	α – $f_\alpha(h)$	-0.32	0.82	-0.37	0.61	-0.39	0.45	-0.40	0.14	-0.405
		β – $f_\beta(h)$	-0.02	1.16	-0.01	0.86	0.00	0.64	0.00	0.20	0
0.4	0.4	α – $N(0, 10^2)$	-0.49	1.01	-0.45	0.68	-0.42	0.47	-0.41	0.15	-0.405
		β – $N(0, 10^2)$	-0.01	1.44	0.00	0.97	-0.01	0.66	0.01	0.20	0

TABLE 1 – Comparaison des estimations a posteriori pour α et β en fonction de la distribution a priori choisie $(f_\alpha(h), f_\beta(h))$ ou $(N(0, 10^2), N(0, 10^2))$

On observe un biais par rapport à l'estimation par le maximum de vraisemblance, qui est du même ordre avec les deux distributions a priori. En revanche, la variance de l'estimation du paramètre est inférieure avec les distributions $f_\alpha(h)$ et $f_\beta(h)$ par rapport aux distributions normales à variance large. Les différences en termes de biais deviennent négligeables lorsque les effectifs sont grands. Pour les petits effectifs, comme par exemple lorsque $N = 5$ dans chaque groupe, la définition de la distribution a priori a donc un impact non négligeable, sur les estimations des paramètres.

On présente sur le graphique ci-contre la distribution $f_\beta(h)$ dans le cas où les paramètres a, b, c et d sont fixés à 1, ainsi que la courbe représentative d'une distribution Normale de moyenne 0 et de variance 6,58.
La distance entre les deux courbes est proche de 0, ainsi pour approximer la distribution issue du rapport de deux distributions $\mathcal{B}(\alpha_i, \beta_i)$, il est raisonnable de considérer la distribution Normale de paramètres correspondants.



4 Discussion

La distribution a priori peu-informative?

Nous avons montré que dans le cadre d'une information a priori vague, c'est à dire dans le cas d'une distribution a priori de variance large, voir une distribution plate, que la définition de la distribution a priori dans un contexte Bayésien peut être spécifiée de multiples façons. Les distributions utilisées habituellement sont des distributions Normales de variance importante (voir virtuellement infinie), à la fois sur l'intercept d'un modèle mais aussi sur le coefficient estimé. L'utilisation d'une telle distribution se justifie par analogie avec la plupart des modèles linéaires généralisés, où il est d'usage de prendre des distributions de variance très large lorsqu'aucune information n'est disponible. Nous avons vu que, dans le cas d'une régression logistique, plus particulièrement dans le cas d'une comparaison de deux proportions, une information a priori est toujours disponible, puisque les proportions sont toutes les deux bornées dans l'intervalle $[0; 1]$, ce qui se traduit en une distribution a priori pour le coefficient de la régression logistique (le logarithme du rapport de cotes) dont la variance est majorée. Nous avons également établi une correspondance entre les coefficients des distributions *Beta* a priori sur les proportions et les paramètres de la distribution a priori correspondant dans la régression logistique, que ce soit pour l'intercept ou pour le coefficient associé à la variable groupe.

Biais sur les estimations

Le choix de la distribution a priori a un rôle important dans l'utilisation des méthodes bayésiennes, tout particulièrement lorsque les effectifs sont faibles. Considérer une distribution a priori non adaptée peut amener à des estimations a posteriori biaisées, ou des variances surestimées. Le choix d'une distribution de variance trop large, particulièrement dans le cas d'un modèle de régression logistique peut être responsable d'un problème de non-convergence des chaînes MCMC ou d'une mauvaise estimation des coefficients et par conséquent du rapport de cotes.

5 Conclusion

Nous avons dérivé une loi de probabilité pour les paramètres α et β d'une régression logistique permettant d'exprimer le même a priori que lors de la comparaison de deux proportions via des lois $\mathcal{B}(\alpha_i, \beta_i)$, pour toutes valeurs de α_i et β_i. Notre étude suggère donc que lors de l'estimation des paramètres d'une régression logistique, bien que la distribution Normale soit une bonne approximation de la distribution du coefficient de la régression, l'utilisation des lois a priori non-informatives usuelles, spécifiant une variance très large, ne correspond pas au paramétrage utilisé habituellement pour comparer deux proportions, sous le même argument de non-informativité. Par ailleurs, on peut prendre par défaut une $\mathcal{N}(0, 6.58)$ dans un contexte peu-informatif.

Bibliographie

[1] Andrew Gelman, *Prior distributions for variance parameters in hierarchical models*, Columbia University, Bayesian Analysis 2006.

[2] Paul C. Lambert, *How vague is vague ? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS*, Leicester, U.K, Statistics in Medecine, 2005.

[3] Irony TZ, Singpurwalla ND. *Noninformative priors do not exist : a discussion with Jose M. Bernardo*, Journal of Statistical Infererence and Planning, 1997

[4] Walley P, Gurrin LC, Burton PR. *Analysis of clinical data using imprecise prior probabilities*, The Statistician, 1996