La régression logistique

Frédéric et Myriam Bertrand[1]

[1] IRMA, Université Louis Pasteur
Strasbourg, France

Master 2ème Année 27-11-2006

Ce cours se base sur l'ouvrage de Bruno Falissard *Comprendre et utiliser les statistiques dans les sciences de la vie*, Professeur des universités et praticien hospitalier à la faculté de médecine Paris-Sud, et le syllabus de *Biostatisque* de Philippe Lambert, Professeur, Université catholique de Louvain.

Exemple

Nombre de souris développant une tumeur au poumon après exposition à la fumée de cigarettes (Essenbergs, Science, 1952).

Groupe	Tumeur présente	Tumeur absente	Total
Contrôle	19	13	32
Traitement	21	2	23

Question : *Existe-t-il une corrélation entre le développement de la maladie et l'apparition du cancer ?*

- Pour **tester** l'existence de ce lien il serait possible de procéder à un test du khi-deux (étudié en L3) :

```
Les dénombrements attendus sont imprimés sous
les dénombrements observés
        Succès  Echec  Total
     1      21      2     23
         16,73   6,27

     2      19     13     32
         23,27   8,73

 Total      40     15     55
Khi deux = 1,091 + 2,910 + 0,784 + 2,092 =
6,878
DL = 1, P = 0,009
```

Ce test ne permet pas de déterminer la **nature** de ce lien, c'est-à-dire comment sont liées les variations des deux variables.

- **Pour parer à cet inconvénient :** On utilise *la régression logistique* qui permet de **modéliser** la probabilité de succès à l'aide des variables explicatives dont nous disposons. Ceci nous permettra de tester si ces changements sont significatifs à un niveau α donné.

De même que la régression linéaire (simple ou multiple) est un prolongement de l'étude du coefficient de corrélation linéaire de deux variables quantitatives, de même la régression logistique est une généralisation d'un coefficient servant à évaluer la corrélation de deux variables qualitatives : *le rapport des côtes* ou *odds-ratio*.

Définition

On appelle **côte du succès** *le rapport*

$$\exp(\theta) = \frac{\pi}{1-\pi}$$

où π est la probabilité de succès.

La probabilité de succés s'exprime à partir de la côte de succès de la manière suivante :

$$\pi = \frac{\exp(\theta)}{1 + \exp(\theta)}.$$

Pour fixer les idées voici quelques valeurs de la côte du succès en fonction la probabilité de succès. (Le logarithme de) cette côte :

- est $(< 0) < 1$ lorsque $\pi < 0.5$.
- est $(= 0) = 1$ lorsque $\pi = 0.5$.
- est $(> 0) > 1$ lorsque $\pi > 0.5$.
- $(\rightarrow -\infty) \rightarrow 0$ lorsque $\pi \rightarrow 0$.
- $(\rightarrow +\infty) \rightarrow +\infty$ lorsque $\pi \rightarrow 1$.

Exemple

La probabilité de succès (i.e. celle de développer une tumeur) observée est égale à :

$$\hat{\pi} = \frac{40}{55} = 0.73$$

$$\Downarrow$$

$$\exp(\hat{\theta}) = \frac{\hat{\pi}}{1 - \hat{\pi}} = 2.67$$

$$\Downarrow$$

$$\hat{\theta} = 0.98.$$

Le logarithme du rapport de côtes :

- On peut calculer la côte de succès dans différentes conditions. Le *rapport de côtes* Ψ permet alors d'évaluer l'infuence du facteur considéré :

  $$\Psi = \frac{\exp(\theta_2)}{\exp(\theta_1)} = \exp(\theta_2 - \theta_1).$$

  Lorsque Ψ est > 1 (< 1) le succès a une côte supérieure (inférieure) pour le deuxième niveau du facteur.
- Le *logarithme du rapport de côtes*, $\theta_2 - \theta_1$, est > 0 (< 0) lorsque le succès a une probabilité supérieure (inférieure) pour le deuxième niveau du facteur.

Exemple

La côte du succès (= "développer une tumeur") observée est égale à :

$$\begin{cases} \text{Côte(succès|exposé)} = \exp(\hat{\theta}_2) = \dfrac{21}{2} = 10.5 \\ \text{Côte(succès|contrôle)} = \exp(\hat{\theta}_1) = \dfrac{19}{13} = 1.46 \end{cases}$$

$$\text{d'où} \quad \hat{\Psi} = \frac{21 \cdot 13}{19 \cdot 2} = 7.18 > 1$$

$$\text{et} \quad \log(\hat{\Psi}) = \hat{\theta}_2 - \hat{\theta}_1 = 1.97 > 0.$$

La côte de succès de la tumeur est supérieure (multipliée par 7) lorsque les souris sont exposées à la fumée de cigarettes.

Intervalle de confiance

- Si pour chaque individu, la probabilité de succès est π, alors, le nombre Y de succès parmi n individus indépendants suit une loi binomiale $\mathcal{B}(n,\pi)$. Ainsi :

$$\mathbb{E}[Y] = n\pi \quad ; \quad \mathrm{Var}[Y] = n\pi(1-\pi)$$
$$\mathbb{E}\left[\hat{\pi} = \frac{Y}{n}\right] = \frac{1}{n}\mathbb{E}[Y] = \pi \quad ; \quad \mathrm{Var}\left[\hat{\pi}\right] = \frac{1}{n^2}\mathrm{Var}[Y] = \frac{\pi(1-\pi)}{n}.$$

- Un intervalle de confiance (dans le cadre d'application de l'approximation de la loi binomiale par une loi normale) à 95 % pour π est donné par :

$$\hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

- Dans notre exemple on souhaiterait comparer les probabilités π_1 et π_2 de développer une tumeur sous et sans exposition à la fumée de cigarettes et déterminer si elles sont significativement différentes. Cela reviendrait à déterminer s'il existe un lien entre le développement de la tumeur et le facteur risque considéré.
- On peut déjà répondre à cette question en construisant un intervalle de confiance à 95 % pour $\pi_1 - \pi_2$.

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm 1.96\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

Exemple

$$0 \notin (0.114, 0.524)$$

On en déduit que la différence $\pi_1 - \pi_2$ est significativement écartée de 0 au seuil $\alpha = 5\%$. Ainsi on sait non seulement la fumée de cigarettes a un effet significatif sur le nombre de cancer développés mais on a quantifié cet effet.

- Dans des situations plus complexes (plus de deux variables ou de deux niveaux du facteur) l'approche précédente est trop lourde. On travaille alors avec les côtes de succès.
- Si X est une variable explicative à K niveaux, le modèle logistique suppose que :

$$(Y|X = x_k) \sim \mathcal{B}(n_k, \pi_k)$$

avec

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \theta_k = \mu + \alpha_k; (\alpha_1 = 0)$$

$$\Rightarrow \pi_k = \frac{\exp(\mu + \alpha_k)}{1 + \exp(\mu + \alpha_k)}.$$

- Le logarithme de la côte de succès sous le premier niveau du facteur vaut μ.
- Le logarithme du rapport des côtes du succès sous les $k^{\text{ème}}$ et 1$^{\text{er}}$ niveau du facteur vaut $\theta_k - \theta_1 = \alpha_k$.
- Par conséquent une valeur de $\alpha_k > 0$ (< 0) indique que la côte du succès observée est plus grande (petite) sous le $k^{\text{ème}}$ niveau du facteur que sous le 1$^{\text{er}}$ niveau du facteur.

Estimation des α_k

- On estime les α_k à l'aide d'une méthode statistique appelée méthode du maximum de vraisemblance.
- Dans ce cas, on sait qu'asymptotiquement (lorsque la taille de l'échantillon tend vers l'infini) les estimateurs des α_k suivent une loi normale de paramètres μ et σ^2 à déterminer.
- De plus, ces estimateurs sont sans biais.

Par conséquent un intervalle de confiance à 95 % approximatif pour les α_k est donné par :

$$\hat{\alpha}_k \pm 1.96 \times \sigma(\hat{\alpha}_k).$$

Les différents modèles possibles sont :

- *Modèle* **1** *avec effet traitement :*

$$\operatorname{logit}(\pi_k) = \theta_k = \mu + \alpha_k.$$

- *Modèle* **2** *sans effet traitement ($\alpha_2 = 0$ ci-dessus) :*

$$\operatorname{logit}(\pi_k) = \theta_k = \mu.$$

On compare alors la probabilité de succès estimée dans le groupe k, notée $\tilde{\pi}_k$ et la proportion de succès observée notée $\hat{\pi}_k$.

Définition

La déviance D est alors définie ainsi :

$$\begin{aligned} D &= -2\sum_k \left\{ y_k \log\left(\frac{\tilde{\pi}_k}{\hat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{1-\tilde{\pi}_k}{1-\hat{\pi}_k}\right)\right\} \\ &= -2(l(\tilde{\pi}_k) - l(\hat{\pi}_k)). \end{aligned}$$

Cette quantité est à rapprocher de la somme des carrés à minimiser dans la régression linéaire simple ou multiple. Elle évalue globalement la qualité de l'ajustement obtenu.

Le deuxième modèle ne fait pas intervenir de variable explicative. Il peut servir à tester la nullité de toutes les pentes : l'équivalent du test de Fisher global dans le cadre de la régression logistique.

On calcule la statistique $G^2 = D_2 - D_1 = -2(l_2 - l_1)$ comparant la déviance des deux modèles.

Définition

Sous l'hypothèse nulle H_0 que les restrictions impliquées par le modèle **2** *au modèle* **1** *sont correctes,*

$$G \overset{H_0}{\sim} \chi^2_{ddl_2 - ddl_1}.$$

Exemple

Sous l'hypothèse nulle

$$H_0 : \quad \alpha_2 = 0$$

on a

$$G_2 = 7.635, \quad ddl_1 = 0, ddl_2 = 1, \quad et \quad p = 0.006.$$

Ce qui permet de décider que α_2 *est significativement différent de* 0 *au niveau* $\alpha = 5\%$. *On obtient également les informations suivantes :* $\hat{\mu} = 0.38$ *et* $\hat{\alpha}_2 = 1.97$. *Ceci permet de calculer les probabilités de succès :* 0.59 *et* 0.91. *Le rapport des côtes du groupe exposé contre le groupe de contrôle est estimé par* $\exp(\hat{\alpha}_2) = 7.24$ *soit une côte de succès plus de 7 fois plus grande pour le groupe des traités.*

On peut construire un intervalle de confiance (approximatif) $(1-\alpha)\cdot 100\%$ pour le logarithme du rapport de côtes (abrégé en LRC) du groupe k contre le groupe de référence α_k avec

$$\hat{\alpha}_k \pm 1.96 \times \sigma(\hat{\alpha}_k).$$

Exemple

Dans notre exemple, on obtient : $\alpha_2 \in (0.36; 3,58)$ *confirmant le rejet de l'hypothèse nulle* H_0 *(avec* $\alpha = 5\%$*) et l'augmentation significative de développer un cancer du poumon après exposition à la fumée de cigarettes. L'intervalle de confiance approximatif pour le rapport de côte est alors égal à* $(1.43, 36.0)$.

Exemple

Voici un second exemple que l'on va traiter avec Minitab. Relation entre les habitudes tabagiques d'étudiants en Arizona et les habitudes de leurs parents (Agresti, 1990, p. 124).

Nombre de parents fumeurs	Enfant fumeur	Enfant non fumeur	Total
Deux	400	1380	1780
Un seul	416	1823	2239
Aucun	188	1168	1358

On définit le succès comme étant le fait de fumer pour l'enfant, le modèle logistique précédent devient :

$$\text{logit}(\pi_k) = \theta_k = \mu + \alpha_k; (\alpha_1 = 0).$$

La catégorie de référence est par défaut "Aucun". On utilise

Minitab pour mener à bien l'analyse. On peut tester l'hypothèse null

$$H_0 : \alpha_2 = \alpha_3 = 0$$

en comparant la déviance de ce modèle avec celle du précedent. $G^2_{obs} = 38.37$ d'où une p-valeur de 0.000.
Conclusion du test : Association significative au niveau $\alpha = 5\%$ entre habitudes tabagiques des parents et des enfants.

Exemple

Effet de la cypermethrine à différentes doses (en μg) sur la survie de parasites. Pour chaque niveau de dose, 20 parasites sont exposés. La survie éventuelle de l'animal est évaluée après 72 heures. Les animaux peuvent être distingués par leur sexe (Collett, 1991, CRC, P. 75).

Dose Mâle	N morts Mâle	Dose Femelle	N morts Femelle
1	1	1	0
2	4	2	2
4	9	4	6
8	13	8	10
16	18	16	12
32	20	32	16

Variable explicative continue

Ignorons le sexe de l'animal en premier lieu.

Question : Existe-t-il un lien entre la mort d'une larve et la dose reçue ? Si oui quelle est la nature de cette relation ?

- On cherche donc à déterminer comment la probabilité de succès π change avec une ou plusieurs variables explicatives continues à partir des observations de y_i succès en n_i expériences indépendantes sous des valeurs de X observées égales à x_i, $(i = 1, \ldots, I)$.
- On souhaite utiliser une modélisation de la côte de succès sachant que $X = x$, c'est-à-dire :

$$(Y|X = x_i) \sim \mathcal{B}(n_i, \pi_i)$$

$$\text{logit}(\pi_i) = \theta_i = \theta_i(x_i).$$

Pour avoir une première idée de la relation entre la côte de succès et X, on examine le **logarithme de la côte empirique** contre x_i :

$$\tilde{\theta}_i = \log\left(\frac{y_i + 0.5}{n_i - y_i + 0.5}\right).$$

On s'aperçoit qu'une transformation logarithmique serait la bienvenue.

Le modèle suggéré est donc :

$$(Y|X = x_i) \sim \mathcal{B}(n_i, \pi_i)$$

avec

$$\text{logit}(\pi_i) = \theta_i = \alpha_0 + \beta_1 x_i$$

où

$$x_i = log(dose_i).$$

Régression logistique : variables explicatives mixtes

- Dans l'exemple précédent, on a ignoré l'influence potentielle du sexe sur la probabilité de succès. L'analyse précédente indique que la dose influe de manière significative sur la probabilité qu'une larve meurt.
- Considérons le cas simple où on a à la fois une variable continue X et une variable qualitative Z. Les données sont donc du type $(y_{ki}, n_{ki}, x_{ki}, z_{ki})$. Le modèle suggéré est donc :

$$(Y|X = x_{ki}, Z = z_{ki}) \sim \mathcal{B}(n_{ki}, \pi_{ki})$$

avec

$$\text{logit}(\pi_{ki}) = \theta_{ki}.$$

Nous avons donc 5 modèles à notre disposition :

- X+Z+X*Z, $(\alpha_0 + \alpha_k) + (\beta_1 + \tau_k)x_{ki}$.
- X+Z, $(\alpha_0 + \alpha_k) + \beta_1 x_{ki}$.
- X, $\alpha_0 + \beta_1 x_{ki}$.
- Z, $\alpha_0 + \alpha_k$.
- 1, α_0.

Reste à détecter les modèles convenables à l'aide du test du G^2. Pour cela, on utilise Minitab et le fichier de données disponible sur le site.