La régression logistique

Frédéric Bertrand[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1ère Année 18-03-2008

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes

Références

Ce chapitre se base sur

1. l'ouvrage *Comprendre et utiliser les statistiques dans les sciences de la vie*, écrit par Bruno Falissard, Professeur des universités et praticien hospitalier à la faculté de médecine Paris-Sud,
2. et le syllabus de *Biostatisque* de l'Université catholique de Louvain, écrit par Philippe Lambert, Professeur à l'Université de Liège.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes

Exemple

Nombre de souris développant une tumeur au poumon après exposition à la fumée de cigarettes

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

d'après Essenbergs, Science, 1952.

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

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes

Remarque

Lorsque nous disposons de deux variables qualitatives X et Y, les moyennes et les variances n'existent plus.
Par conséquent, les coefficients comme le coefficient de corrélation linéaire ou les rapports définis dans les autres chapitres n'ont plus lieu d'exister.
Il ne reste donc qu'un seul élément exploitable : **la loi conjointe du couple** (X, Y)**.**

À partir de cette information, trois questions semblent naturelles et pertinentes :

Q1. Les variables X et Y sont-elles indépendantes ?

Q2. Les distributions conditionnelles de Y sachant X (respectivement X sachant Y) sont-elles homogènes ?

Q3. La distribution du couple (X, Y) est-elle « proche » d'une distribution théorique ?

Indépendance entre deux variables

Une méthode pour répondre à la première question :

« Les variables X et Y sont-elles indépendantes ? »

consiste :

- à construire le tableau de contingence associé aux variables X et Y sous l'hypothèse d'indépendance, lequel est obtenu en effectuant le produit des fréquences marginales.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Exemple
Test d'indépendance du χ^2
Rapport des côtes
Intervalle de confiance

Suite

- Puis comparer la distribution empirique, c'est-à-dire celle contenue dans le tableau de contingence, avec la distribution théorique, c'est-à-dire celle obtenue par calcul. L'interprétation résultant de la comparaison de ces deux distributions est alors la suivante :
  1. Si les deux distributions sont identiques, les variables X et Y sont indépendantes.
  2. Si les deux distributions sont différentes, les variables X et Y ne sont pas indépendantes (elles peuvent être liées, corrélées ou non-corrélées).

Remarque

Pour autant, il est très rare en pratique, même dans le cas de variables réellement indépendantes, d'observer une égalité des distributions théoriques et empiriques et cela pour deux raisons :

1. du fait que nous observons un échantillon et non pas la population entière
2. à cause des erreurs de mesure.

Pour palier à ce problème, une idée consiste à évaluer la « distance » entre ces deux distributions bidimensionnelles désignées sous les notations

$$\{f_{ij} : 1 \leqslant i \leqslant I; 1 \leqslant j \leqslant J\} \quad \text{et} \quad \{f_{i.} \times f_{.j} : 1 \leqslant i \leqslant I; 1 \leqslant j \leqslant J\}.$$

Cette « distance » est mesurée via la quantité

$$\chi^2_{(I-1)(J-1)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{..} f_{i.} \times f_{.j})^2}{n_{..} f_{i.} \times f_{.j}}$$

appelée « distance du χ^2 » à $ddl = (I-1)(J-1)$ degrés de liberté.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Exemple
Test d'indépendance du χ^2
Rapport des côtes
Intervalle de confiance

Ainsi, pour répondre à la première question , nous effectuons un test, appelé le **test d'indépendance du χ^2**, qui permettra de prendre une décision quant à l'hypothèse d'indépendance. Cela revient à tester l'hypothèse d'indépendance :

$$\mathcal{H}_0 : f_{ij} = f_{i.} \times f_{.j}, \quad \forall 1 \leqslant i \leqslant I \quad \text{et} \quad \forall 1 \leqslant j \leqslant J,$$

ou encore

$$\mathcal{H}_0 : f_{ij} - f_{i.} \times f_{.j} = 0, \quad \forall 1 \leqslant i \leqslant I \quad \text{et} \quad \forall 1 \leqslant j \leqslant J,$$

l'**hypothèse nulle** contre l'hypothèse de non indépendance

$$\mathcal{H}_1 : f_{ij} \neq f_{i.} \times f_{.j}, \text{ pour au moins un couple } (i,j),$$

l'**hypothèse alternative**.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Exemple
Test d'indépendance du χ^2
Rapport des côtes
Intervalle de confiance

Principe du test d'indépendance du χ^2

Il consiste à rejeter l'hypothèse nulle $\mathcal{H}_0$ (c'est-à-dire l'indépendance des deux variables X et Y) dès que la quantité $\chi^2_{(I-1)(J-1)}$, calculée à partir des données empiriques, est « trop » grande. En d'autres mots, nous rejetons l'hypothèse nulle $\mathcal{H}_0$ dès que

$$\chi^2_{(I-1)(J-1)} \geqslant c,$$

où $c > 0$ est une valeur à déterminer (dans la table du χ^2) en fonction des degrés de liberté et de l'erreur α, que le statisticien s'autorise selon le contexte. Cette valeur a souvent été appelée **valeur critique** dans ce cours.
Dans le cas contraire, nous acceptons l'hypothèse nulle $\mathcal{H}_0$.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Exemple
Test d'indépendance du χ^2
Rapport des côtes
Intervalle de confiance

Remarque

Cela revient à dire que si la distance du χ^2 est « petite », la différence entre les deux distributions n'est pas significative et peut être imputée à des erreurs de mesure.
Dans ce cas, nous décidons d'accepter l'hypothèse nulle $\mathcal{H}_0$, c'est-à-dire que l'on accepte l'hypothèse d'indépendance entre les variables X et Y.
Dans le cas contraire, la différence est jugée significative, et nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$, c'est-à-dire que nous rejetons l'hypothèse d'indépendance des deux variables X et Y.

Condition d'application

Lorsque la taille de l'échantillon est suffisamment grande (de telle sorte que l'effectif théorique, sous l'hypothèse nulle $\mathcal{H}_0$, dans chaque colonne dépasse 5, c'est-à-dire $n_{..}f_{i.} \times f_{.j} \geqslant 5$), la constante c s'obtient en consultant une table dite de la **loi du χ^2**. Cette constante varie selon deux paramètres :

1. l'erreur α et
2. les degrés de liberté *ddl*.

Cette table est une table à double entrée.

Rappel

Nous rappelons que l'erreur α, choisie généralement par le statisticien et aussi par le domaine dans lequel sont extraites les données, est appelée l'**erreur de première espèce**. Elle représente la probabilité de rejeter l'hypothèse nulle $\mathcal{H}_0$ à tort.
Il est important de garder à l'esprit que nous pouvons accepter l'hypothèse nulle $\mathcal{H}_0$ alors que les deux variables ne sont pas indépendantes.
Mais, faute de « preuves » suffisantes, nous acceptons l'hypothèse d'indépendance $\mathcal{H}_0$.

Schéma pratique de la réalisation du test d'indépendance du χ^2

1. Nous nous donnons une erreur α (généralement α vaut $1\%, 5\%$ ou 10% en fonction du domaine dans lequel les données sont traitées).
2. Nous déterminons le $ddl = (I-1)(J-1)$.
3. Nous déterminons la valeur critique c via la table du χ^2.
4. Nous calculons le $\chi^2_{(I-1)(J-1)}$.
5. Nous comparons le $\chi^2_{(I-1)(J-1)}$ à la valeur critique c et nous concluons grâce à cette comparaison soit à l'acceptation de l'hypothèse nulle $\mathcal{H}_0$ soit au rejet de l'hypothèse nulle $\mathcal{H}_0$.

Retour à l'exemple

Nombre de souris développant une tumeur au poumon après exposition à la fumée de cigarettes

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

d'après Essenbergs, Science, 1952.

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

Réponse

Pour **tester** l'existence de ce lien, il est possible de procéder à un test du χ^2 :

```
Les dénombrements attendus sont notés 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, c = 3,841
```

Avec le logiciel R

```
>
Exemple1<-matrix(c(21,2,19,13),byrow=T,nrow=2,
+ dimnames=list(c("Traitement","Contrôle"),
+c("Tumeur présente","Tumeur absente")))
> Exemple1
Tumeur présente Tumeur absente
Traitement 21 2
Contrôle 19 13
> chisq.test(Exemple1,correct=FALSE)
Pearson's Chi-squared test
data : Exemple1
X-squared = 6.8781, df = 1, p-value = 0.008726
```


Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Exemple
Test d'indépendance du χ^2
Rapport des côtes
Intervalle de confiance

Suite de l'exemple avec le logiciel R

La ligne de commande sous R pour faire apparaître les valeurs théoriques pour vérifier que l'effectif dans chaque case est supérieur à 5.

```
>chisq.test(Exemple1,correct=FALSE)$expected
Tumeur présente Tumeur absente
Traitement 16.72727 6.272727
Contrôle 23.27273 8.727273
```

Remarque

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 :

Nous utilisons *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.

Remarque

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

Nous appelons **côte du succès** le rapport

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

où π est la probabilité de succès.

Définition

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

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

Quelques exemples de côte

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 $\pi > 0.5$.
- $(\to -\infty) \to 0$ lorsque $\pi \to 0$.
- $(\to +\infty) \to +\infty$ lorsque $\pi \to 1$.

Exemple

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

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

$$\Downarrow$$

$$\exp(\hat{\theta}) = \frac{\hat{\pi}}{1-\hat{\pi}} = \frac{0.73}{0.27} = 2.67$$

$$\Downarrow$$

$$\hat{\theta} = \ln(2.67) = 0.98.$$

Le logarithme du rapport de côtes de succès :

1. Nous pouvons calculer la côte de succès dans différentes conditions.
2. Définition : Le *rapport de côtes* Ψ permet alors d'évaluer l'infuence du facteur considéré :

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

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

Exemple

La côte de succès (= « développer une tumeur ») observée est égale à :

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

$$\text{D'où} \quad \hat{\Psi} = \frac{10.5}{1.46} = 7.18 > 1$$

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

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, nous avons :

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

Définition

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 :

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

Retour à l'exemple

Dans l'exemple, nous souhaiterions 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.

Nous pouvons déjà répondre à cette question en construisant un intervalle de confiance à 95 % pour $\pi_1 - \pi_2$:

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

Retour à l'exemple : application numérique

Calculons l'intervalle de confiance :

$$21/23 - 19/32$$
$$\pm 1.96 \sqrt{\frac{21/23 \times (1 - 21/23)}{23} + \frac{19/32 \times (1 - 19/32)}{32}}.$$

Donc, nous remarquons $0 \notin (0.114, 0.524)$. Nous en déduisons que la différence $\pi_1 - \pi_2$ est significativement écartée de 0 au seuil $\alpha = 5\%$.

Ainsi nous savons non seulement que la fumée de cigarettes a un effet significatif sur le nombre de cancers développés mais surtout nous avons quantifié cet effet.

Remarque

Dans des situations plus complexes, à savoir par exemple dans des cas où il y a plus que deux variables qualitatives ou plus que deux niveaux du facteur qui est joué par la variable qualitative (nous rappelons que nous parlons de facteur lorsque nous avons à faire à des variables qualitatives), l'approche précédente est trop lourde.

$\Rightarrow$ Nous travaillons alors avec les côtes de succès que nous allons définir.

Définition

Si X est une variable explicative à I niveaux, le modèle logistique suppose que :

$$(Y|X = x_i) \sim \mathcal{B}(n_i, \pi_i), \quad \text{où} \quad i = 1, \ldots, I$$

avec

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_i = \mu + \alpha_i; (\alpha_1 = 0)$$

$$\Rightarrow \pi_i = \frac{\exp(\mu + \alpha_i)}{1 + \exp(\mu + \alpha_i)}.$$

Définition

Le logarithme de la côte de succès sous le premier niveau du facteur vaut μ.

Définition

Le logarithme du rapport des côtes du succès sous les $i^{\text{ème}}$ et 1^{er} niveau du facteur vaut $\theta_i - \theta_1 = \alpha_i$.

Remarque

Par conséquent une valeur de $\alpha_i > 0$ (< 0) indique que la côte du succès observée est plus grande (petite) sous le $i^{\text{ème}}$ niveau du facteur que sous le 1^{er} niveau du facteur.

Estimation des α_i

- Nous estimons les α_i à l'aide d'une méthode statistique appelée méthode du maximum de vraisemblance.
- Dans ce cas, nous savons qu'asymptotiquement (lorsque la taille de l'échantillon tend vers l'infini) les estimateurs des α_i suivent une loi normale de moyenne α_k et de variance $\mathrm{Var}\,[\hat{\alpha}_i]$.
- De plus, ces estimateurs sont sans biais.
- Par conséquent un intervalle de confiance à 95 % approximatif pour les α_i est donné par :

$$\hat{\alpha}_i \pm 1.96 \times \sqrt{\mathrm{Var}\,[\hat{\alpha}_i]}.$$

Les différents modèles possibles pour l'exemple

- *Modèle* **1** *avec « effet du traitement » :*

$$\mathrm{logit}(\pi_i) = \theta_i = \mu + \alpha_i \quad \text{où} \quad i = 1 \quad \text{ou} \quad 2.$$

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

$$\mathrm{logit}(\pi_i) = \theta_i = \mu \quad \text{où} \quad i = 1 \quad \text{ou} \quad 2.$$

Nous comparons 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 :

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

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.

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

Définition

Sous l'hypothèse nulle $\mathcal{H}_0$ que les restrictions impliquées par le modèle **2** au modèle **1** sont correctes,

$$G^2 \overset{\mathcal{H}_0}{\sim} \chi^2_{ddl_2 - ddl_1}.$$

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Définitions
Les modèles
Déviance

Exemple

Sous l'hypothèse nulle

$$\mathcal{H}_0 : \quad \alpha_2 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \quad \alpha_2 \neq 0$$

nous trouvons, grâce à `R` :
`G2= Null Deviance = 7.635, ddl = 1, et c = 3.8415 (qchisq(0.95,1)).`
Ce qui permet de décider que α_2 est significativement différent de 0 au niveau $\alpha = 5\%$.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Définitions
Les modèles
Déviance

Exemple

Nous obtenons également les informations suivantes, à l'aide de `R` :

```
Coefficients :
(Intercept) GroupeT
0.3795 1.9719
```

Ce qui nous permet d'écrire que $\hat{\mu} = 0.3795$ et $\hat{\alpha}_2 = 1.9719$.
Ce qui nous permet de calculer les probabilités de succès :

- $\hat{\pi}_2 = \frac{\exp(0.3795)}{1+\exp(0.3795)} = 0.5937$ et
- $\hat{\pi}_1 = \frac{\exp(2.3514)}{1+\exp(2.3514)} = 0.9130 \quad (2.3514 = 0.3795 + 1.9719).$

Le rapport des côtes du groupe exposé contre le groupe de contrôle est estimé par $\exp(\hat{\alpha}_2) = 7.1843$ soit une côte de succès plus de 7 fois plus grande pour le groupe des traités.

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes
Définitions
Les modèles
Déviance

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

$$\hat{\alpha}_k \pm 1.96 \times \sqrt{\mathrm{Var}[\hat{\alpha}_k]}.$$

Exemple

Dans notre exemple, nous obtenons avec le logiciel `R` : α_2=`GroupeT∈(0.3590202;3.584751)` confirmant le rejet de l'hypothèse nulle $\mathcal{H}_0$ (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 `(1.431926;36.044384)`.

Lignes de commande de R pour obtenir les résultats précédents

```
>Exemple1bis<- read.table(file.choose(),
+dec=".",sep=";",header=T)
model<-glm(cbind(Tumeur,Total-Tumeur)~Groupe,
+data=Exemple1bis,family=binomial(link=logit))
model
summary(model)
confint.default(model)
exp(confint.default(model))
anova(model)
```


Exemple

Voici un second exemple que nous allons traiter.
Existe-t-il une relation entre les habitudes tabagiques d'étudiants en Arizona et les habitudes de leurs parents ?

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

D'après Agresti, 1990, p. 124.

Résultats

Nous définissons 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 ».
Nous utilisons R pour mener à bien l'analyse. Nous testons l'hypothèse nulle

$$\mathcal{H}_0 : \alpha_2 = \alpha_3 = 0$$

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 les habitudes tabagiques des parents et celles des enfants.

Un troisième 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	N morts	Dose	N morts
Mâle		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.

Questions :

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

- Nous cherchons 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, $(i = 1, \ldots, I)$.
- Nous souhaitons 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, nous examinons le **logarithme de la côte empirique** contre x_i :

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

Nous nous apercevons 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, nous avons 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 nous avons 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

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

Introduction
Régression logistique : variable explicative qualitative
Régression logistique : variable explicative continue
Régression logistique : variables explicatives mixtes

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, vous utiliserez le logiciel R et le fichier de données disponible sur le site.