Régression logistique ou modèle binomial

Résumé

Introduction au modèle linéaire et modèle linéaire général : la régression logistique ou modèle binomial. Retour au plan du cours.

1 Introduction

Dans ce chapitre, nous définissons le contexte pratique de la *régression logistique* qui s'intéressent plus particulièrement à la description ou l'explication d'observations constitués d'effectifs comme, par exemple, le nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais. Contrairement aux modèles du chapitre précédent basés sur l'hypothèse de normalité des observations, les lois concernées sont discrètes et associées à des dénombrements : binomiale, multinomiale. Néanmoins, ce modèle appartient à la famille du *modèle linéaire général* (annexe) et partagent à ce titre beaucoup d'aspects (estimation par maximum de vraisemblance, tests, diagnostics) et dont la stratégie de mise en œuvre, similaire au cas gaussien, n'est pas reprise.

Une première section définit quelques notions relatives à l'étude de la liaison entre variables qualitatives. Elles sont couramment utilisées dans l'interprétation des modèles de régression logistique.

2 Odds et odds ratio

Une variable

Soit Y une variable qualitative à J modalités. On désigne la chance (ou *odds* [1] de voir se réaliser la $j$ème modalité plutôt que la $k$ème par le rapport

$$\Omega_{jk} = \frac{\pi_j}{\pi_k}$$

où π_j est la probabilité d'apparition de la $j$ème modalité. Cette quantité est estimée par le rapport n_j/n_k des effectifs observés sur un échantillon. Lorsque la variable est binaire et suit une loi de Bernouilli de paramètre π, l'odds est le rapport $\pi/(1-\pi)$ qui exprime une cote ou chance de gain.

Par exemple, si la probabilité d'un succès est 0.8, celle d'un échec est 0.2. L'odds du succès est 0.8/0.2=4 tandis que l'odds de l'échec est 0.2/0.8=0.25. On dit encore que la chance de succès est de 4 contre 1 tandis que celle d'échec est de 1 contre 4.

Table de contingence

On considère maintenant une table de contingence 2×2 croisant deux variables qualitatives binaires X^1 et X^2. les paramètres de la loi conjointe se mettent dans une matrice :

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

où $\pi_{ij} = P[\{X^1 = i\} \text{ et } \{X^2 = j\}]$ est la probabilité d'occurence de chaque combinaison.

- Dans la ligne 1, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :
$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}}.$$
- Dans la ligne 2, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :
$$\Omega_2 = \frac{\pi_{21}}{\pi_{22}}.$$

On appelle *odds ratio* (rapport de cote) le rapport

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Ce rapport prend la valeur 1 si les variables sont indépendantes, il est supérieur à 1 si les sujets de la ligne 1 ont plus de chances de prendre la première colonne que les sujets de la ligne 2 et inférieur à 1 sinon.

Exemple : supposons qu'à l'entrée dans une école d'ingénieurs, 7 garçons sur 10 sont reçus tandis que seulement 4 filles sur 10 le sont. L'odds des garçons est alors de 0.7/0.3=2.33 tandis que celle des filles est de 0.4/0.6=0.67.

1. Il n'existe pas, même en Québécois, de traduction consensuelle de "odds" qui utilise néanmoins souvent le terme "cote".

L'odds ratio est de 2.33/0.67=3.5. La chance d'être reçu est 3.5 plus grande pour les garçons que pour les filles.

L'odds ratio est également défini pour deux lignes (a, b) et deux colonnes (c, d) quelconques d'une table de contingence croisant deux variables à J et K modalités. L'odds ratio est le rapport

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}} \quad \text{estimé par l'odds ratio empirique} \quad \widehat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}.$$

3 Régression logistique

3.1 Type de données

Cette section décrit la modélisation d'une variable qualitative Z à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de maladie, panne d'un équipement, faillite d'une entreprise, bon ou mauvais client.... Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\boldsymbol{\beta}$ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités

$$\pi = P(Z = 1) \quad \text{ou} \quad 1 - \pi = P(Z = 0),$$

ou plutôt une transformation de celles-ci, par l'observation conjointe des variables explicatives. L'idée est en effet de faire intervenir une fonction réelle monotone g opérant de $[0, 1]$ dans $\mathbb{R}$ et donc de chercher un modèle linéaire de la forme :

$$g(\pi_i) = \mathbf{x}_i'\beta.$$

Il existe de nombreuses fonctions, dont le graphe présente une forme sigmoïdale et qui sont candidates pour remplir ce rôle, trois sont pratiquement disponibles dans les logiciels :

probit : g est alors la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.

log-log avec g définie par

$$g(\pi) = \ln[-\ln(1 - \pi)]$$

mais cette fonction est dissymétrique.

logit est définie par

$$g(\pi) = \mathit{logit}(\pi) = \ln\frac{\pi}{1-\pi} \quad \text{avec} \quad g^{-1}(x) = \frac{e^x}{1+e^x}.$$

Plusieurs raisons, tant théoriques que pratiques, font préférer cette dernière solution. Le rapport $\pi/(1 - \pi)$, qui exprime une "cote", est l'*odds* et la *régression logistique* s'interprète donc comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio" c'est-à-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.

Cette section se limite à la description de l'usage élémentaire de la régression logistique. Des compléments concernant l'explication d'une variable qualitative ordinale (plusieurs modalités), l'intervention de variables explicatives avec effet aléatoire, l'utilisation de mesures répétées donc dépendantes, sont à rechercher dans la bibliographie.

3.2 Modèle binomial

On considère, pour $i = 1, \ldots, I$, différentes valeurs *fixées* $x_i^1, \ldots, x_i^q$ des variables explicatives $X^1, \ldots, X^q$. Ces dernières pouvant être des variables quantitatives ou encore des variables qualitatives, c'est-à-dire des facteurs issus d'une planification expérimentale.

Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations ($n = \sum_{i=1}^{I} n_i$) de la variable Z qui se mettent sous la forme $y_1/n_1, \ldots, y_I/n_I$ où y_i désigne le nombre de "succès" observés lors des n_i essais. On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, la probabilité π_i de succès est constante. Alors, la variable Y_i sachant n_i et d'espérance $E(Y_i) = n_i\pi_i$ suit une loi *binomiale* $\mathcal{B}(n_i, \pi_i)$ dont la fonction de densité s'écrit :

$$P(Y = y_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1 - \pi_i)^{(n_i - y_i)}.$$

On suppose que le vecteur des fonctions *logit* des probabilités π_i appartient au sous-espace vect$\{X^1, \ldots, X^q\}$ engendré par les variables explicatives :

$$\mathit{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, I$$

ce qui s'écrit encore

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}} \quad i = 1, \ldots, I.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.

L'optimisation fournit une estimation $\mathbf{b}$ de $\boldsymbol{\beta}$, il est alors facile d'en déduire les estimations ou prévisions des probabilités π_i :

$$\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1+e^{\mathbf{x}_i'\mathbf{b}}}$$

et ainsi celles des effectifs

$$\widehat{y}_i = n_i \widehat{\pi}_i.$$

Remarques

1. La matrice $\mathbf{X}$ issue de la planification expérimentale est construite avec les mêmes règles que celles utilisées dans le cadre de l'analyse de covariance mixant variables explicatives quantitatives et qualitatives. Ainsi, les logiciels gèrent avec plus ou moins de clarté le choix des variables indicatrices et donc des paramètres estimables ou contrastes associés.
2. La situation décrite précédemment correspond à l'observation de données *groupées*. Dans de nombreuses situations concrètes et souvent dès qu'il y a des variables explicatives quantitatives, les observations $\mathbf{x}_i$ sont toutes distinctes. Ceci revient donc à fixer $n_i = 1; i = 1, \ldots, I$ dans les expressions précédentes et la loi de Bernouilli remplace la loi binomiale. Certaines méthodes ne sont alors plus applicables et les comportements asymptotiques des distributions des statistiques de test ne sont plus valides, le nombre de paramètres tendant vers l'infini.
3. Dans le cas d'une variable explicative X dichotomique, un logiciel comme SAS fournit, en plus de l'estimation d'un paramètre b, celle des odds ratios ; b est alors le log odds ratio ou encore, e^b est l'odds ratio. Ceci s'interprète en disant que Y a e^b fois plus de chance de succès (ou de maladie comme par un exemple un cancer du poumon) quand $X = 1$ (par exemple pour un fumeur).

3.3 Régressions logistiques polytomique et ordinale

La régression logistique adaptée à la modélisation d'une variable dichotomique se généralise au cas d'une variable Y à plusieurs modalités ou polytomique. Si ces modalités sont ordonnés, on dit que la variable est qualitative ordinale. Ces types de modélisation sont très souvent utilisés en épidémiologie et permettent d'évaluer ou comparer des risques par exemples sanitaires. Des estimations d'odds ratio ou rapports de cotes sont ainsi utilisés pour évaluer et interpréter les facteurs de risques associés à différents types (régression polytomique) ou seuils de gravité (régression ordinale) d'une maladie ou, en marketing, cela s'applique à l'explication, par exemple, d'un niveau de satisfaction d'un client. Il s'agit de comparer entre elles des estimations de fonctions logit.

Dans une situation de *data mining* ou fouille de données, ce type d'approche se trouve lourdement pénalisé lorsque, à l'intérieur d'un même modèle polytomique ou ordinal, plusieurs types de modèles sont en concurrence pour chaque fonction logit associée à différentes modalités. Différents choix de variables, différents niveaux d'interaction rendent trop complexe et inefficace cette approche. Elle est à privilégier uniquement dans le cas d'un nombre restreint de variables explicatives avec un objectif explicatif ou interprétatif.

À titre illustratif, explicitons le cas simple d'une variable Y à k modalités ordonnées expliquée par une seule variable dichotomique X. Notons $\pi_j(X) = P(Y = j|X)$ avec $\sum_{j=1}^{k} \pi_j(X) = 1$. Pour une variable Y à k modalités, il faut, en toute rigueur, estimer $k-1$ prédicteurs linéaires :

$$g_j(X) = \alpha_j + \beta_j X \quad \text{pour} j = 1, \ldots, k-1$$

et, dans le cas d'une variable ordinale, la fonction lien logit utilisée doit tenir compte de cette situation particulière.

Dans la littérature, trois types de fonction sont considérées dépendant de l'échelle des rapports de cote adoptée :

- échelle basée sur la comparaison des catégories adjacentes deux à deux,
- sur la comparaison des catégories adjacentes supérieures cumulées,
- et enfin sur la comparaison des catégories adjacentes cumulées.

Pour $k = 2$, on retrouve les trois situations se ramènent à la même d'une variable dichotomique. C'est le dernier cas qui est le plus souvent adopté ; il conduit à définir les fonctions des "logits cumulatifs" de la forme :

$$\log \frac{\pi_{j+1} + \cdots + \pi_k}{\pi_1 + \cdots + \pi_j} \quad \text{pour} j = 1, \ldots, k-1.$$

Pour un seuil donné sur Y, les catégories inférieures à ce seuil, cumulées, sont comparées aux catégories supérieures cumulées. Les fonctions logit définies sur cette échelle dépendent chacune de tous les effectifs, ce qui peut conduire à une plus grande stabilité des mesures qui en découlent.

Si les variables indépendantes sont nombreuses dans le modèle ou si la variable réponse Y comporte un nombre élevé de niveaux, la description des fonctions logit devient fastidieuse. La pratique consiste plutôt à déterminer un coefficient global b (mesure d'effet) qui soit la somme pondérée des coefficients b_j. Ceci revient à faire l'hypothèse que les coefficients sont homogènes (idéalement tous égaux), c'est-à-dire à supposée que les rapports de cotes sont proportionnels. C'est ce que calcule implicitement la procédure LOGISTIC de SAS appliquée à une variable réponse Y ordinale en estimant un seul paramètre b mais $k-1$ termes constants correspondant à des translations de la fonctions logit. La procédure LOGISTIC fournit le résultat du test du score sur l'hypothèse H_0 de l'homogénéité des coefficients β_j.

Le coefficient b mesure donc l'association du facteur X avec la gravité de la maladie et peut s'interpréter comme suit : pour tout seuil de gravité choisi sur Y, la cote des risques d'avoir une gravité supérieure à ce seuil est e^b fois plus grande chez les exposés ($X = 1$) que chez les non exposés ($X = 0$).

Attention dans SAS, la procédure LOGISTIC adopte une paramétrisation $(-1, 1)$ analogue à celle de la procédure CATMOD mais différente de celle de GENMOD ou SAS/Insight $(0, 1)$. Ceci explique les différences observées dans l'estimation des paramètre d'une procédure à l'autre mais les modèles sont identiques.

4 Choix de modèle

4.1 Recherche pas à pas

Principalement deux critères (test du rapport de vraisemblance et test de Wald, cf. bibliographie), sont utilisés de façon analogue au test de Fisher du modèle linéaire gaussien. Ils permettent de comparer un modèle avec un sous-modèle et d'évaluer l'intérêt de la présence des termes complémentaires. On suit ainsi une stratégie descendante à partir du modèle complet. L'idée est de supprimer, un terme à la fois, la composante d'interaction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald. Les tests présentent une structure hiérarchisée. SAS facilite cette recherche en produisant une décomposition (Type III) de ces indices permettant de comparer chacun des sous-modèles excluant un des termes avec le modèle les incluant tous.

Attention, du fait de l'utilisation d'une transformation non linéaire (logit), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction.

4.2 Critère

L'approche précédente favorise la qualité d'ajustement du modèle. Dans un but prédictif, certains logiciels, comme Splus/R ou Enterpirse Miner, proposent d'autres critères de choix (AIC, BIC). Une estimation de l'erreur de prévision par validation croisée est aussi opportune dans une démarche de choix de modèle.

5 Illustration élémentaire

5.1 Les données

On étudie l'influence du débit et du volume d'air inspiré sur l'occurence (codée 1) de la dilatation des vaisseaux sanguins superficiels des membres inférieurs. Un graphique élémentaire représentant les modalités de Y dans les coordonnées de $X^1 \times X^2$ est toujours instructif. Il montre une séparation raisonnable et de bon augure des deux nuages de points. Dans le cas de nombreuses variables explicatives quantitatives, une analyse en composantes prin-

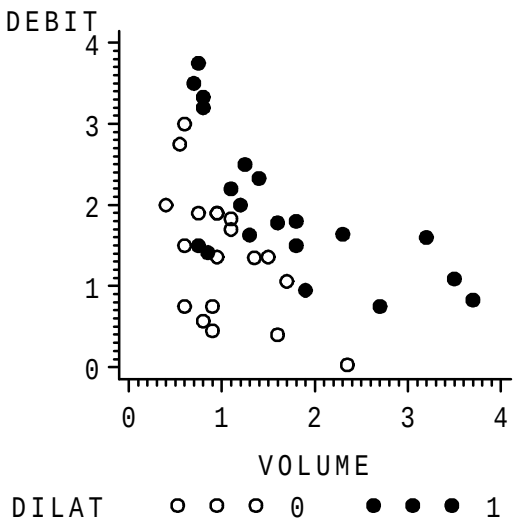

FIGURE 1 – Dilatation : Nuage des modalités de Y dans les coordonnées des variables explicatives.

cipales s'impose. Les formes des nuages représentés, ainsi que l'allure des distributions (étudiées préalablement), incitent dans ce cas à considérer par la suite les logarithmes des variables. Une variable `un` ne contenant que des "1" dénombrant le nombre d'essais est nécessaire dans la syntaxe de `genmod`. Les données sont en effet non groupées.

```
proc logistic data=sasuser.debvol;
model dilat=l_debit l_volume;
run;
proc genmod data=sasuser.debvol;
model dilat/un=l_debit l_volume/d=bin;
run;
```

```
The LOGISTIC Procedure

                              Intercept
                   Intercept      and
 Criterion           Only     Covariates    Chi-Square for Covariates
 AIC                56.040       35.216          .
 SC                 57.703       40.206          .
 -2 LOG L           54.040       29.216(1)    24.824 with 2 DF (p=0.0001)
 Score                 .            .         16.635 with 2 DF (p=0.0002)

            Parameter(2) Standard    Wald(3)     Pr >    Standardized  Odds
Variable DF  Estimate    Error   Chi-Square  Chi-Square  Estimate     Ratio
INTERCPT 1     2.8782    1.3214     4.7443     0.0294          .         .
L_DEBIT  1    -4.5649    1.8384     6.1653     0.0130    -2.085068    0.010
L_VOLUME 1    -5.1796    1.8653     7.7105     0.0055    -1.535372    0.006
```

Cette procédure fournit des critères de choix de modèle dont la déviance (1), le vecteur **b** des paramètres (2) et les statistiques des tests (3) comparant le modèle excluant un terme par rapport au modèle complet tel qu'il est décrit dans la commande.

```
        Criteria For Assessing Goodness Of Fit
     Criterion              DF         Value      Value/DF
     Deviance               36       29.2156        0.8115 (1)
     Scaled Deviance        36       29.2156        0.8115 (2)
     Pearson Chi-Square     36       34.2516        0.9514 (3)
     Scaled Pearson X2      36       34.2516        0.9514
     Log Likelihood          .      -14.6078             .
              Analysis Of Parameter Estimates
  Parameter     DF     Estimate (4)  Std Err    ChiSquare (5) Pr>Chi
  INTERCEPT      1      -2.8782      1.3214       4.7443       0.0294
  L_DEBIT        1       4.5649      1.8384       6.1653       0.0130
  L_VOLUME       1       5.1796      1.8653       7.7105       0.0055
  SCALE (6)      0       1.0000      0.0000            .            .
```

(1)	Déviance du modèle par rapport au modèle saturé.
(2)	Déviance pondérée si le paramètre d'échelle est différent de 1 en cas de sur-dispersion.
(3)	Statistique de Pearson, voisine de la déviance, comparant le modèle au modèle saturé .
(4)	Paramètres du modèle.
(5)	Statistique des tests comparant le modèle excluant un terme par rapport au modèle complet.
(6)	Estimation du paramètre d'échelle si la quasi-vraisemblance est utilisée.

5.2 Régression logistique ordinale

On étudie les résultats d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province de l'Alberta à Edmonton au Canada (Jobson, 1991). Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table croisant :

1. Etat du conducteur : Normal ou Alcoolisé
2. Sexe du conducteur
3. Port de la ceinture : Oui Non
4. Gravité des blessures : 0 : rien à 3 : fatales

Les modalités de la variable à expliquer concernant la gravité de l'accident sont ordonnées.

```
/* régression ordinale */
proc logistic data=sasuser.ceinture;
class sexe alcool ceinture;
model gravite=sexe alcool ceinture ;
weight effectif;
run;
```

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	Gr0	1	1.8699	0.0236	6264.9373	<.0001
Intercept	Gr1	1	2.8080	0.0269	10914.3437	<.0001
Intercept	Gr2	1	5.1222	0.0576	7917.0908	<.0001
sexe	Sfem	1	-0.3118	0.0121	664.3353	<.0001
alcool	A_bu	1	-0.5017	0.0190	697.0173	<.0001
ceinture	Cnon	1	-0.1110	0.0174	40.6681	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
sexe	Sfem vs Shom	0.536	0.511	0.562
alcool	A_bu vs Ajeu	0.367	0.340	0.395
ceinture	Cnon vs Coui	0.801	0.748	0.858