6.2 Les principaux modèles GLM

Dans cette section, nous décrions les principaux modèles GLM utilisés. Il en existe de nombreuses variantes que nous ne pouvons pas toutes mentionner ici. L'objectif est donc de comprendre les rouages de ces modèles afin de pouvoir en cas de besoin reporter ces connaissances sur des modèles plus spécifiques. Pour faciliter la lecture de cette section, nous vous proposons une carte d'identité de chacun des modèles présentés. Elles contiennent l'ensemble des informations pertinentes à retenir pour chaque modèle.

6.2.1 Les modèles GLM pour des variables qualitatives

Nous abordons en premier les principaux GLM utilisés pour modéliser des variables binaires, multinomiales et ordinales. Prenez bien le temps de saisir le fonctionnement du modèle logistique binomial car il sert de base pour les trois autres modèles présentés.

6.2.1.1 Le modèle logistique binomial

Le modèle logistique binomial est une généralisation du modèle de Bernoulli que nous avons présenté dans l'introduction de cette section. Le modèle logistique binomiale couvre donc deux cas de figure :

1. La variable observée est binaire (0 ou 1). Dans ce cas, le modèle logistique binomiale devient un simple modèle de Bernoulli.
2. La variable observée est un comptage (nombre de réussites) et on dispose d'une autre variable avec le nombre de réplications de l'expérience. Par exemple, pour chaque intersection d'un réseau routier, nous pourrions avoir le nombre de décès à vélo (variable Y de comptage) et le nombre de collisions vélo / automobile (variable quantifiant le nombre d'expérience, chaque collision étant une expérience). Spécifiquement, on tente de prédire le paramètre p de la distribution binomiale à l'aide de notre équation de régression et la fonction logistique comme fonction de lien.

Tableau 6.6: Carte d'identité du modèle logistique binomial

Type de variable dépendante	Variable binaire (0 ou 1) ou comptage de réussite à une expérience (ex : 3 réussites sur 5 expériences)
Distribution utilisée	Binomiale
Formulation	$Y \sim Binomial(p)$ $g(p) = \beta_0 + \beta X$ $g(x) = log(\frac{x}{1-x})$
Fonction de lien	logistique
Paramètre modélisé	p
Paramètres à estimer	β_0, β
Conditions d'application	Non-séparation complète, Absence de surdispersion ou sousdispersion

6.2.1.1.1 Interprétation des paramètres

Les seuls paramètres à estimer du modèle sont les coefficients β_0 et β. La fonction de lien logistique transforme la valeur de ces coefficients, en conséquence, ils ne peuvent plus être interprétés simplement. β_0 et β sont des logarithmes de rapports de cote (*log odd ratio*). Le rapport de cote est relativement facile à interpréter. Pour l'obtenir, il suffit d'utiliser la fonction exponentielle (l'inverse de la fonction logarithmique) pour passer des log rapport de cote à de simples rapport de cote. Donc si $exp(\beta)$ est inférieur à 1, il réduit la probabilité d'observer l'évènement et inversement si $exp(\beta)$ est supérieur à 1.

Par exemple, admettons que nous ayons eu un coefficient β_1 de 1,2 pour une variable X_1 dans une régression logistique. Il est nécessaire d'utiliser son exponentiel pour l'interpréter de façon intuitive. $exp(1,2) = 3,32$, ce qui signifie que lorsque l'on augmente X_1 d'une unité, on multiplie les chances par 3,32 d'observer 1 plutôt que 0 comme valeur de *Y*. Admettons maintenant que β_1 vaille -1,2, on calcule donc $exp(-1,2) = 0,30$, ce qui signifie qu'a chaque augmentation d'une unité de X_1, on multiplie les chances par 0,30 d'observer 1 plutôt que 0

comme valeur de *Y*. En d'autres termes, on divise par 3,33 ($1/0,30 = 3,33$) les chances d'observer 1 plutôt que 0, soit une diminution de 70% ($1 - 0,3 = 0,7$) des chances d'observer 1 plutôt que 0.

Les rapports de cotes

Le rapport de cote ou rapport des chances est une mesure utilisée pour exprimer l'effet d'un facteur sur une probabilité très utilisé dans le domaine de la santé, mais aussi des paris. Prenons un exemple concret avec le port du casque à vélo. Si sur 100 accidents impliquant des cyclistes portant un casque on observe seulement 3 cas de blessures graves à la tête, contre 15 dans un second groupe de 100 cyclistes ne portant pas de casques, on peut calculer le rapport de cote suivant :

$$\frac{p(1-q)}{q(1-p)} = \frac{0,15 * (1-0,03)}{0,03 * (1-0,15)} = 5,71$$

avec *p* la probabilité d'observer le phénomène (ici la blessure grave à la tête) dans le groupe 1 (ici les cyclistes sans casque) et *q* la probabilité d'observer le phénomène dans le groupe 2 (ici les cyclistes avec un casque). Ce rapport de cote indique que les cyclistes sans casques ont 5,71 fois plus de chances de se blesser gravement à la tête lors d'un accident que ceux portant un casque.

6.2.1.1.2 Les conditions d'application

La non-séparation complète signifie qu'aucune des variables *X* n'est, à elle seule, capable de parfaitement distinguer les deux catégories 0 et 1 de la variable *Y*. Dans un tel cas de figure, les algorithmes d'ajustement utilisés pour estimer les paramètres des modèles sont incapables de converger. Notez aussi l'absurdité de créer un modèle pour prédire une variable *Y* si une variable X est capable à elle seule de la prédire à coup sûr. Ce problème est appelé un effet de Hauck-Donner, il est assez facile de le repérer, car la plupart du temps les fonctions de R signalent ce problème (message d'erreur sur la convergence). Sinon, des valeurs extrêmement élevées ou faibles pour certains rapports de cote peuvent aussi indiquer un effet de Hauck-Donner.

La surdispersion est un problème spécifique aux distributions n'ayant pas de paramètre de dispersion (binomiale, poisson, exponentielle, etc.), pour lesquelles la variance dépend directement de la moyenne. On parle de surdispersion lorsque dans un modèle les résidus (ou erreurs) sont plus dispersés de ce que suppose la distribution utilisée. À l'inverse, il est aussi possible (mais rare) d'observer de cas de sous-dispersion (lorsque la dispersion des résidus est plus petite que ce que suppose la distribution choisie). Ce cas de figure se produit généralement lorsque le modèle parvient à réaliser une prédiction trop précise pour être fiable. Si vous rencontrez une forte sous-dispersion, cela signifie souvent que l'un de vos prédicteurs provoque une séparation complète. La meilleure option dans ce cas est de supprimer le prédicteur en question du modèle. La variance attendue d'une distribution binomiale est $nb * p * (1-p)$, soit