Cours1_GLM

January 19, 2024

1 Vraisemblance et Modèles linéaires généralisés

(Generalized linear models : GLM ou GLIM)

1.1 Comment ajuster un modèle linéaire lorsque les aléas sont non-gaussien ?

1.1.1 ex : Données de comptage

Si la probabilité qu'un événement se produise dans un intervalle $T = 1$ est λ, le nombre d'événements attendu (l'espérance) dans l'intervalle $T = t$ est $E(N) = \lambda \times t$, et la variance du nombre d'événements est également $var(N)\lambda \times t$. Enfin, la distribution des probabilités des nombre d'événements est la **loi de Poisson** dont la formule est

$$P(N = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

Lorsque λ est grand (≥ 30), cette loi peut être approximée par une loi normale. Donc si la variable expliquée correspond à des comptages et que, pour tous les cas de figures étudiés, on a $\hat{y}$ (i.e. $\bar{y}$) ≥ 30, on peut utiliser modèles linéaire classique ◡̈. Mais ce cas de figure est rare...

Solutions

Transformations (Cette section s'applique à tous les modèles avec transformation, et pas seulement à l'analyse de donnée de comptage.) * Utilisez un modèle log-linéaire (ou au besoin, un modèle log-log lien). D'une façon générale, la distribution log-normale est utilisée pour modéliser des données positives où les observations diffèrent de plusieurs ordres de grandeur. * Utilisez la transformation box-cox (liens 1 et 2).

Que ce soit avec le log ou le box-cox, des comptages à 0 sont problématiques ($\log(0) = -\infty$). La stratégie classiquement utilisée consiste à ajouter une constante (δ) à la variable expliquée : le modèle devient $\log(y_i + \delta) = ax + b + \cdots + e_i$ ou box-cox$(y_i + \delta) = ax + b + \cdots + e_i$. La valeur de δ peut-être optimisée pour mieux normaliser les données ou fixée arbitrairement à 1 par commodité.

/!\ : Ce type d'approche a été largement utilisé par le passé, mais elle peut complexifier substantiellement l'interprétation des résultats, et ainsi conduire à des erreurs systématiques dans les conclusions : * Data Transformations for Inference with Linear Regression: Clarifications and Recommendations * Do not log-transform count data * Comment on 'Do not log-transform count data'

Generalized linear models (GLM) Des modèles linéaires basés sur d'autres lois que la loi normale ont étés développés, il s'agit principalement des GLM basés sur la loi de poisson et sur la loi binomial. Ces modèles sont relativement bien adaptés aux données de comptage, mais avant de les décrire plus en détail, il est nécessaire d'introduire la notion de ***vraissemblance***, sur laquelle ils reposent.

La vraisemblance est probablement le concept le plus important des statistiques inférentielles après la *p*.value. Toutes les statistiques Bayésiennes sont basées dessus, mais aussi tous les modèles plus complexes que les modèles linéaires (ANOVA et régréssions), et un très grand nombre de modèles très spécifiques à certains domaines scientifiques (phylogénies, alignement de séquences, généalogies, prédiction de structures protéiques, ...).

1.2 Vraisemblance (Likelihood)

En statistique, la vraisemblance d'un modèle est la probabilité d'observer nos données si le modèle est vrai. Autrement dit, la vraisemblance du modèle θ étant donné notre variable expliquée y est :

$$\mathcal{L}(\theta|y) = P(y|\theta)$$

(Notez que par *convention*, on inverse les termes de part et d'autre du | . Parfois, $\mathcal{L}(\theta|y)$ est indiqué $\mathcal{L}(\theta; y)$)

Pour calculer la probabilité de notre jeu de donnée si le modèle est vrai, on calcule la probabilité de chacune des observations de notre jeu de donnée si le modèle est vrai, puis on prend le produit de toutes ces probabilités. Cette dernière étape fait donc l'hypothèse que les observations sont indépendantes (lien). Autrement dit, on a :

$$\mathcal{L}(\theta|y) = P(y|\theta) =$$
$$P(x_1|\theta) \times P(x_2|\theta) \times P(x_{...}|\theta) \times P(x_n|\theta) =$$
$$\prod_{i=1}^{n} P(x_i|\theta)$$

En particulier pour les grands jeux de données, calculer la vraisemblance implique donc de réaliser le produit d'un très grand nombre de valeurs proche de zéro, ce qui, informatiquement, est compliqué à faire de façon précise. Ainsi, on préfère calculer le log de la vraisemblance, car $\log(a \times b) = \log(a) + \log(b)$:

$$\log(\mathcal{L}(\theta|y)) = \log\left(\prod_{i=1}^{n} P(x_i|\theta)\right) = \sum_{i=1}^{n} \log\left(P\left(x_i|\theta\right)\right)$$

En pratique, on utilise ce concept, en premier lieu, pour estimer les paramètres d'un modèle, grâce à la **fonction de vraisemblance**. Il s'agit de la fonction qui lie les paramètres du modèle et à sa vraisemblance. Par exemple, pour un modèle M qui a p paramètres θ qui sont ajustés au jeu de données D, la fonction de vraisemblance est donnée par :

$$f\left(\{\theta_1, \theta_2, \theta_{...}, \theta_p\}; y\right) = \mathcal{L}(M_{\{\theta_1, \theta_2, \theta_{...}, \theta_p\}}|y)$$

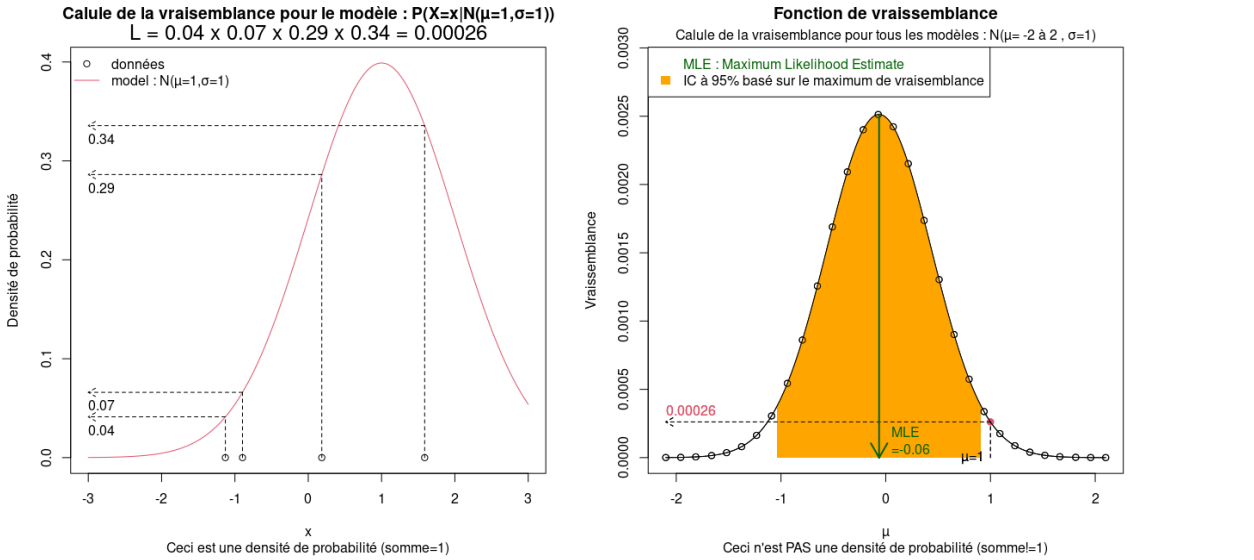
Souvent, il est possible d'obtenir la formule mathématique de la fonction de vraisemblance, ce qui permet de connaitre la combinaison de paramètres qui maximise la fonction de vraisemblance. Lorsque ce n'est pas possible, on utilise des algorithmes de tâtonnement.

En second lieu, lorsque l'on a des modèles de structure différente, dont les paramètres ont été ajustés aux données (généralement par maximum de vraisemblance), on utilise leur vraisemblance pour les comparer, soit par le test du rapport de vraisemblance, soit par l'AIC (liens 1, 2, 3, 4 et 4), soit par d'autres indices (généralement basés sur la vraisemblance).

En exécutant la cellule ci-dessous, vous ferez apparaitre une animation visant à expliquer le calcule de d'une **vraisemblance**, la **fonction de vraisemblance**, et l'estimation de paramètre par le **maximum de vraisemblance**.

[1]:
```
# options(repr.plot.width=14, repr.plot.height=7, repr.plot.res = 150)
speed = 0.5 # vitesse de l'animation

source("./Figures/explainLikelihood.R")
```



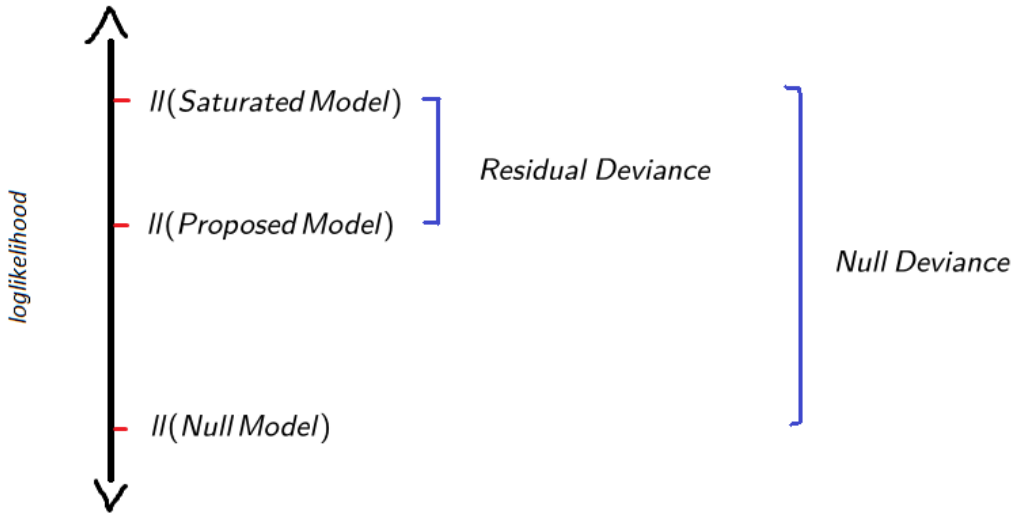
Question pour voir si vous suivez : *Quelle est la similitude entre la* p.*value et la vraisemblance ?*

Réponse : »» » > La *p*.value correspond à la probabilité, si H_0 est vrai, d'observer des données au moins aussi éloignées de l'attendue sous H_0 que ce que le sont nos données, $P\left(|X| \geq |X_{obs}| \; |H_0\right)$. > > La vraisemblance correspond à la probabilité de nos données sur le modèle, θ, est vrai, c.-à-d. $P\left(D|\theta\right)$. > > La principale différence réside donc dans le fait que l'on utilise généralement la *p*.value lorsque l'on souhaite rejeter H_0, alors que l'on utilise la vraisemblance lorsque l'on souhaiterait plutôt accepter le modèle. Ensuite, ces deux concepts utilisent la densité de probabilité qui leur est associée d'une façon de légèrement différente. Mais ces deux concepts reposent lourdement sur la densité de probabilité décrivant la probabilité des données (ou d'une statistique résumant les données), si une hypothèse (H_0) ou un modèle (θ) est vrai.

Pour une autre explication de la vraisemblance, et pour aller plus loin : Un très bon cours

1.2.1 Déviance

Une notion connexe de la vraisemblance est la déviance. La déviance mesure la déviation entre le GLM étudié, et un modèle dit saturé qui s'ajusterait parfaitement aux données. Ce modèle saturé peut, par exemple, être une régression qui a été ajusté à deux observations. Autrement, il s'agit d'une mesure de la déviation entre les prédictions du GLM et les données elle-mêmes, mais en standardisant par la vraisemblance maximale du modèle. En tant que mesure de la déviation entre les prédictions du modèle et les données, elle est l'équivalant de la somme du carré des écarts ($\sum_{i=1}^{n} e_i^2$) pour les modèles linéaires classiques. La déviance est donc un résumé de la variabilité des résidus, elle est donc parfois appelée **déviance résiduelle**. A contrario, la déviance nulle correspond à la déviance d'un modèle qui ne contiendrait qu'un seul paramètre, l'intercept, dont la valeur serait alors $\hat{y}$:



La formule de la déviance est donc $\mathcal{D} = 2\left(\log \mathcal{L}\left(\theta_s|y\right) - \log \mathcal{L}\left(\theta_f|y\right)\right)$, où θ_s représente le modèle saturé, et θ_f le modèle focal, dont on cherche à calculer la déviance. La déviance étant généralement utilisée pour différents modèles ajustés aux-même données, on simplifie parfois sa formule en $\mathcal{D} \rightarrow 2\left(-\log \mathcal{L}\left(\theta_f|y\right)\right)$ afin de s'affranchir du calcul de $\log \mathcal{L}\left(\theta_s|y\right)$. De façon abusive, on parle encore de deviance pour cette formule.

1.3 Modèles linéaires généralisés (GLM)

Les modèles linéaires généralisés permettent d'ajuster des modèles linéaires, y compris lorsque les aléas affectant la variable expliquée ne sont pas normalement distribués.

À cette fin, deux modifications des modèles linéaires classiques sont réalisées : 1. Une **fonction de lien** est appliquée aux prédictions du modèle qui se présente donc comme suit : $> y_i = \widehat{y_i} + e_i$ $f(\widehat{y_i}) = ax + b + ... >$ Ce qui peut être ré-écrit en
$> y_i = f^{-1}\left(ax + b + ...\right) + e_i$, $>$ où f^{-1} est la fonction réciproque de la fonction de lien. 3. Les résidus ne sont plus supposés suivre une loi normale, mais une autre loi de probabilité (Poisson, Négative-Binomiale, Binomiale ou autre, selon les données analysées et le choix de l'utilisateurice).

En comparaison aux modèles linéaires classiques (LM), ces modèles font les hypothèses suivantes :

Comme pour les modèles linéaires classiques, * **les résidus doivent être i.i.d.** * Comme pour les

LM, **les variables explicatives peuvent être des transformations non-linéaires** de certaines variables originales.

À la différence des modèles linéaires classiques, * un GLM ne suppose PAS de relation linéaire entre la variable réponse et les variables explicatives, mais **il suppose une relation linéaire entre la réponse attendue transformée par fonction de lien et les variables explicatives** (voire équation ci-dessus). * **la variable expliquée n'a pas besoin d'être normalement distribuée**, mais il est généralement supposé qu'elle suit une loi de la famille des exponentielles (ex : binomiale, de Poisson, multinomiale, normale, etc.) * **Il n'est PAS nécessaire que l'homogénéité de la variance soit satisfaite**. En fait, dans de nombreux cas, le modèle *suppose* une hétérogénéité de la variance d'une forme particulière. Le non-respect de la forme de cette hétérogénéité de la variance, ce qui peut conduire à de la **sousdispersion**, ou de la **surdispersion** (under ou overdispersion). Le non-respect de la forme de la variance supposée par un GLM est généralement bien plus grave que le non-respect de l'homogénéité des variances dans un modèle linéaire classique. **Cela peut conduire à de conclusions entièrement déconnectées de la réalité des données.**

L'estimation des paramètres utilise l'estimation du maximum de vraisemblance (MLE) plutôt que les moindres carrés ordinaires (OLS).

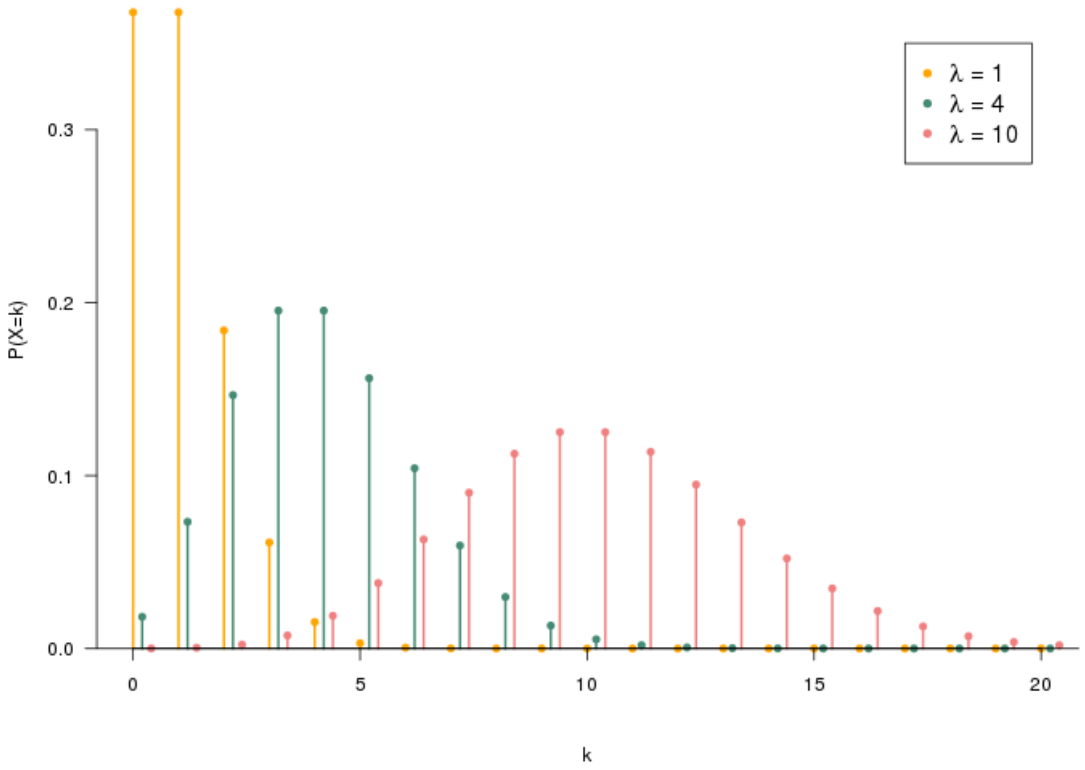
1.3.1 Types de données

Focus sur les lois Poisson et Binomiale, et les modèles du même nom

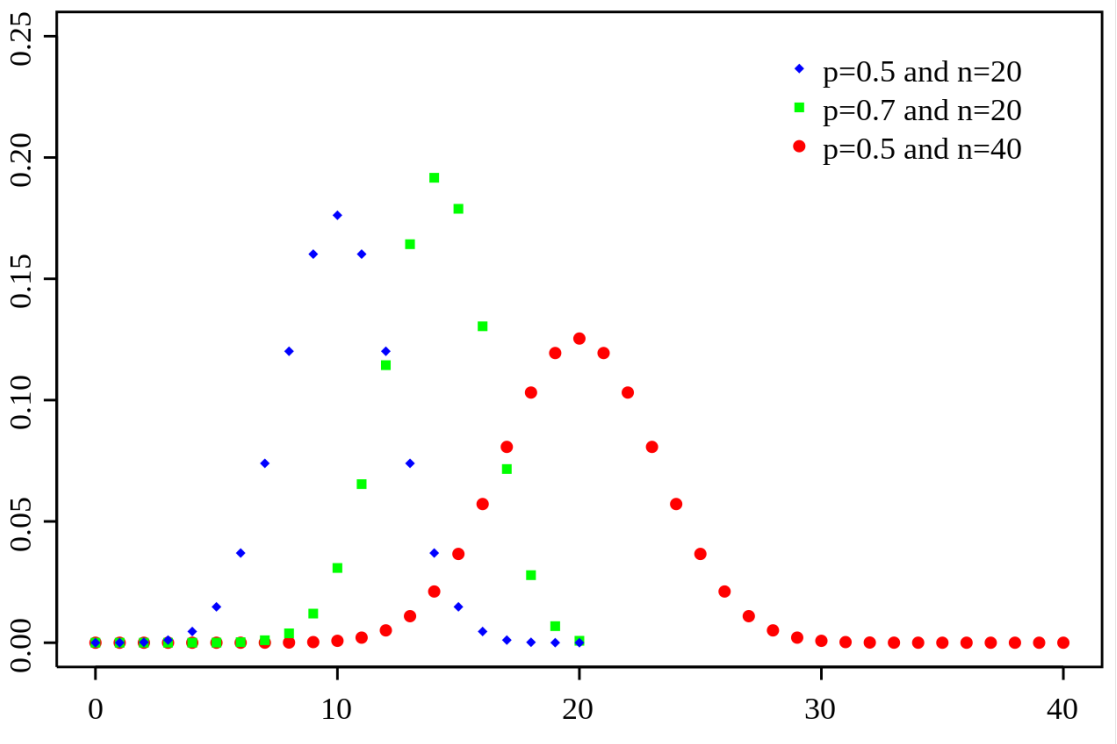
Les GLM sont principalement utilisés lorsque la variable expliquée correspond à des données de comptage. On peut distinguer deux grands types de données de comptage : * Les données de comptage simple, lorsqu'il n'y a pas de maximum théorique connu avant le recueille des données. On parle alors de variable suivant une **loi Poisson** (ex : combien de voitures neuves chaque français va encore acheter avant de prendre conscience de l'impasse climatique dans laquelle nous conduit ce mode de transport ?) Pour ce type de données de type Poisson, la fonction de lien utilisée est le logarithme (ln). * Les données de comptage catégorisées, chaque observation que l'on comptabilise est allouée à une catégorie. Typiquement, il n'y a que deux catégories, l'une que l'on nomme 'échecs', et l'autre que l'on nomme 'succès'. On parle alors de données suivant une **loi binomiale**. Lorsqu'il y a plus de deux catégories, on parle de loi multinomiale. (ex de variable binomiale : Pour chaque partie de 'la fresque du climat' qui est jouée, combien de participant e s prennent durablement conscience de l'extrême gravité de la situation dans laquelle l'humanité s'est mise ? Pour cette variable, il y a un individu statistique par fresque, et deux valeurs à chaque fois : nombre de prises de consciences durables et nombre de non-prises de consciences durables). Pour ce type de données binomiales, on utilise généralement comme fonction de lien la fonction logit et plus rarement, on utilise la fonction probit. Un lien sur ces 2 fonctions. Mathématiquement, la fonction logit est donnée par $\ln\left(\frac{p}{1-p}\right); p \in]0;1[$, et sa fonction réciproque est $\frac{1}{1+e^{-x}}; x \in]-\infty;\infty[$. La fonction probit est compliquée, il s'agit de la fonction inverse de la fonction $\phi(x)$, laquelle est définie ici.

Illustration de ces 2 lois :

Exemple de lois de de Poisson :



Exemples de lois Binomiales :



On peut constater 2 choses sur ces figures : * Bien que la variance ne soit pas explicitement mentionnée, les modifications des paramètres affectent la variance. * Pour certaines gammes de paramètres, les distributions semblent proches d'une loi normale.

1.3.2 Les hypothèses sur la variance et le problème de surdispersion

En fait, pour la loi de Poisson, le paramètre λ correspond à la fois à la moyenne et à la variance. Et pour la loi Binomiale, la moyenne est égale à np, et la variance à $np(1-p)$.

En termes de modélisation statistique, ceci signifie que si que l'on utilise à GLM de type Poisson, *par exemple*, pour expliquer une variable de type Poisson par une variable explicative catégorielle à trois niveaux (I, II, et III), le modèle suivant sera ajusté aux données : $y_i = \exp\left(\{a_I,\ a_{II},\ a_{III}\}_i\right) + e_i$, la fonction exponentielle (exp) étant la réciproque de la fonction de lien ln qui est utilisé pour les GLM Poisson. Les coefficients $\{a_I,\ a_{II},\ a_{III}\}$ correspondent aux prédictions linéarisées du modèle : $\ln(\widehat{y})$. Autrement-dit, le modèle a ajusté trois distributions de Poisson aux données, avec les paramètres suivant : $\lambda_{I,\ II \text{ ou } III} = \exp(a_{I,\ II \text{ ou } III})$. Pour estimer la significativité de la différence entre ces trois valeurs de λ, le modèle suppose que la variance des e_i au sein de chaque groupe est égale à la moyenne du groupe : $\lambda_{I,\ II \text{ ou } III}$. Or si des variables affectant les valeurs de la variable expliquée y n'ont pas été prises en compte, et qu'elles sont indépendantes de la variable explicative, alors, au sein de chacune des catégories I, II, et III, ces variables cachées vont augmenter la variabilité des y_i et donc des e_i. Ainsi, il y aura plus de variance résiduelle que supposé par le modèle (c.-à-d. de la surdispersion), ce qui conduit à surestimer la significativité des effets (sous-estimer les p.values.) En résumer, lorsque l'on utilise des GML Poisson (**tout comme pour Binomiaux**) pour que les p.value soient correctes, il est **indispensable** que toutes les variables (et interactions) ayant des effets importants sur la variable expliquée y soient incluses dans le modèle (c.-à-d., il ne doit pas y avoir de variable cachée). En comparaison, les modèles linéaires classiques supposent que, s'il y a des variables cachées, celle-ci sont indépendantes des variables explicatives étudiées.

Lorsque, à l'inverse, il y a moins de variance dans les résidus que ce que le modèle le suppose, on parle de **sousdispersion**, de qui conduit à sous-estimer la significativité des effets. La sousdispersion est généralement dû au sur-apprentissage (overfitting) : il n'y a pas suffisamment de données par rapport au nombre de paramètres que le modèle cherche à ajuster aux données.

Ci-après, je discute de différentes approches permettant d'éviter le problème de surdispersion (et parfois le problème de sousdispersion).

En pratique, ces alternatives sont plus complexes que les GLM Poissons et Binomiaux. Classiquement, on commence donc par ajuster un modèle, soit Poisson, soit Binomial selon le type de données, puis on mesure le niveau de dispersion des résidus et on le compare à la dispersion supposée par le modèle. C'est seulement dans les cas (fréquents) où la différence entre les deux est trop importante que l'on s'intéressera aux alternatives.

Pour estimer le niveau de surdispersion (ou sousdispersion), on utilise la formule suivante :

$$\phi = \frac{\chi^2}{N-k}$$

Où N correspond au nombre d'observations et k au nombre de paramètres (coefficients). Le numérateur correspond à la mesure de la qualité de l'ajustement du modèle (goodness-of-fit statistic) donnée par la somme des carrés des résidus de Pearson.

Dans le cadre de données de Poisson, il s'agit de

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \widehat{y_i}}{\sqrt{\widehat{y_i}}} \right)^2$$

Dans le cadre de données Binomiale, il s'agit de

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \widehat{y_i}}{\sqrt{\widehat{y_i}(1 - \widehat{y_i})}} \right)^2$$

(Dans les deux cas, les dénominateurs correspondent à l'erreur standard prédite.)

/!\ : Pour pouvoir estimer la surdispersion dans le cadre modèles binomiaux, il faut les données soient formatées en format 'court', et non en format 'long' : Logistic Regression: Bernoulli vs. Binomial Response Variables

S'il n'y pas de sur ou sousdisersion, on doit avoir $\phi = 1$, en cas de surdispersion, $\phi > 1$ et sousdispersion $\phi < 1$. Des procédures ont été développées pour tester l'hypothèse $H_0 : \phi = 1$. Cependant, lorsqu'il y a peu de données, même une forte surdispersion pourra ne pas être significative et à l'inverse, lorsqu'il y a beaucoup de données, même une surdispersion négligeable pourra être fortement significative. Ceci est une des nombreuses illustrations que dans les situations où l'on souhaiterait idéalement que H_0 soit vrai ou à peu près vrai, la p.value est inutile, seule la taille de l'effet compte. Dans le cas de la surdispersion, la taille de l'effet, c'est ϕ.

De plus, on peut se questionner sur le sens d'ajuster à des données un modèle qui contraint aussi fortement la variance des résidus que le font les GLM Poisson et Binomiaux : Est-ce cela à du sens de supposer que l'on connait toutes les variables explicatives et qu'on les prend en compte, quand on sait que dans la très large majorité des cas, c'est faux ? : LIEN

Certains statisticiens suggèrent qu'il faudrait directement employer des procédures qui ne craignent pas de sur- ou de sousdispersion. Celle-ci sont détaillées ci-après.

1.3.3 Comment gérer le problème de surdispersion (ou de sousdispersion) ?

Approximation normale Comme nous l'avions noté pour les figures montrant des exemples de lois Poisson et de lois Binomiale, pour certaines gammes de paramètres, les distributions semblent proches d'une loi normale.

En effet, 1. on peut approximer une loi Binomiale par une loi de Poisson lorsque [n est grand (≥ 10) et p est faible (≤ 0.1)] ou lorsque [$n \geq 100$ et $np \leq 10$ (lien). 2. une loi de Poisson peut elle-même être approximée par une loi Normale dès que $\lambda \geq 5$ (lien).

Cependant, pour pouvoir utiliser ces approximations de données de comptage par des lois Normales dans le cadre de modèles linéaires, il est aussi nécessaire que les variances soient homogènes. Donc si l'on cherche analyser des données de comptage (de type Poisson ou Binomiale), pour lesquels la variance change généralement avec la moyenne, il faudra prêter une attention accrue au respect de l'hypothèse d'homoscédasticité.

Comme discuté précédemment, il est aussi possible de transformer les données pour respecter lien

La quasivraisemblance Le problème de la surdispersion vient du fait que l'estimation de la vraisemblance d'un GLM Poisson ou Binomial suppose que la variance des résidus est identique à celle supposée par le modèle ($\widehat{y}$ pour la Poisson et $\widehat{y}(1-\widehat{y})$ pour la Binomiale). Une solution autre a été développée pour analyser ces GLMs Poisson et Binomial, elle repose sur une modification de la fonction de vraisemblance, qui du coup n'est plus une vraisemblance, mais une quasi-vraisemblance (quasilikelihood). Cette modification permet que la variance des résidus soit corrigé un facteur multiplicatif qui est lui-même estimé à partir des données. Ainsi, les GLMs quasi-Poisson et quasi-Binomial supposent que la variance des résidus est proportionnelle à celle supposée par les GLM classiques (mentionnées au début du paragraphe).

Cette approche est robuste à la surdispersion. Elle fait une hypothèse forte sur la variance des résidus, que la variance résiduelle est proportionnelle à $\widehat{y}$ pour un GLM quasi-Poisson et à $\widehat{y}(1-\widehat{y})$ pour un GLM quasi-Binomial. Mais cette hypothèse est du même ordre que l'hypothèse d'homoscédasticité des modèles linéaires, si son respect est toujours souhaitable, son non-respect aura toujours des conséquences moindres que la surdispersion.

Comme le note Harrison, X. A. (2015), > the 'quasi' approach is that it does not model the overdispersion in the data, but merely adjusts the resulting parameter estimates with a single correction factor. The assumption that all standard errors are biased to the same degree is an obvious problem, which may not be appropriate.

Distributions alternatives : Poisson → binomiale-négative Binomale → Bêta-binomiale

Binomiale-négative La distribution binomiale négative (parfois appelée 'Poisson-Gamma Mixture') suppose que, le paramètre λ de la loi de Poisson est lui même une variable aléatoire, distribuée selon une distribution Gamma, dont la moyenne et la variance deviennent des paramètres que le modèle doit estimer. Cette loi peut donc être utilisée à la place de la distribution poisson afin de s'affranchir de l'hypothèse faite sur la variance. Cependant, sa fonction de vraisemblance est complexe, et elle peut donc parfois poser des problèmes de convergence du modèle. Enfin, par construction, sa variance est toujours supérieure ou égale à la variance de la loi Binomiale. Elle n'est donc pas adaptée pour des problèmes de sous-dispersion.

Pour plus d'information : lien

Bêta-binomiale La distribution bêta-binomiale suppose que, dans une loi binomiale, la probabilité de réussite change à chaque série d'essais selon une distribution bêta. Autrement dit, pour tirer aléatoirement un nombre de succès et d'échecs dans une loi bêta-binomiale de paramètre n, α et β, on tire aléatoirement le paramètre p dans une loi bêta (α, β) puis on tire aléatoirement le nombre de succès et d'échecs dans une loi binomiale (n, p). Cette loi peut être utilisée à la place de la distribution binomiale afin de s'affranchir de l'hypothèse faite sur la variance. Cependant, sa fonction de vraisemblance est complexe, et elle peut donc parfois poser des problèmes de convergence du modèle. Enfin, par construction, sa variance est toujours supérieure ou égale à la variance de la loi Binomiale. Elle n'est donc pas adaptée pour des problèmes de sous-dispersion.

Pour plus d'information : lien

Lorsque le nombre d'essais n est constant pour chaque observation, il est aussi possible d'ajuster un GLM bêta à la variable proportion de succès. Pour plus d'information, voir les deux vignettes du package `R` `betareg`.

Ajouter un effet aléatoire avec un niveau par observation (Observation-level random effects - OLRE) Globalement, des simulations (1 et 2) ont montré que cette approche, qui a le mérite d'être simple, fonctionne généralement bien quand * la surdispersion n'est pas due à une inflation de 0 * la surdispersion n'est pas trop énorme ($\phi > 15$ à 20) * dans le cas de données binomiales, la surdispersion n'est pas générée par un processus beta-binomial

Bootstrap Les coefficients des GLM classiques peuvent être bootstrappés !