Modèles linéaires généralisés à effets fixes et aléatoires de la variabilité inter et intra-individuelle

Yvonnick Noël[1]

Dans l'arsenal des modèles statistiques bien connus des psychologues, les modèles dits linéaires sont largement dominants. Que les variables prédictives au sein de ces modèles soient numériques, catégorielles (recodées en indicatrices) ou les deux en même temps, on peut exprimer dans une formulation unifiée les méthodes bien connues que sont la régression linéaire, l'analyse de la variance et l'analyse de la covariance. Ce cadre simple, déjà assez intégrateur, est classiquement nommé Modèle Linéaire Général. Il fait l'hypothèse d'une distribution gaussienne sur la variable dépendante, conditionnellement aux prédicteurs, et d'un lien linéaire ou de proportionnalité entre variables explicatives et à expliquer.

Nous rappelons dans la première partie de ce chapitre les fondements de cette classe de modèles, pour présenter ensuite deux extensions plus récentes : les Modèles Linéaires Généralisés (GLM), qui libèrent des contraintes de normalité et de linéarité (Nelder & Wedderburn, 1972), et les Modèles Linéaires Généralisés Mixtes (GLMM), ou à coefficients variables, qui permettent d'amener dans la modélisation des contraintes subtiles, purement distributionnelles, sur les paramètres. Nous illustrons sur plusieurs jeux de données réelles les propriétés et l'application concrètes de ces deux classes de modèles.

Le modèle linéaire général

La régression linéaire

Le modèle linéaire est la formulation statistique la plus simple qui soit d'une relation de proportionnalité entre p variables explicatives X_j ($j = 1, \ldots, p$) et la valeur attendue sur une variable à expliquer Y. Dans ce qui suit, on notera les variables en majuscules (X, Y), pour indiquer clairement leur statut de variables aléatoires, et les modalités observées en minuscules ($x_1, x_2 \ldots$, $y_1, y_2 \ldots$). Le principe sous-jacent est que si l'un des prédicteurs augmente en intensité, la variable à expliquer doit, en espérance, augmenter (ou diminuer) en proportion simple. Pour une observation i donnée, cette hypothèse se traduit statistiquement par une dépendance linéaire de la moyenne attendue $\mu_i = E(Y|\mathbf{x}_i)$ de la variable dépendante, pour un ensemble de valeurs fixées $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ des X_j :

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \qquad \text{(Equation 1)}$$

Dans ce modèle, les coefficients β_j sont des réels, connus ou inconnus, qui ajustent les échelles des prédicteurs, potentiellement différentes, sur celle de Y. Le modèle est incomplet en l'état, car il ne fait que formuler une dépendance déterministe *de la moyenne*

[1] Université Européenne de Bretagne, Rennes 2, Centre de Recherche en Psychologie, Cognition et Communication, place du recteur Henri le Moal, F-35043 Rennes Cedex. Les scripts R pour reproduire les analyses et les graphiques de ce chapitre sont disponibles auprès de l'auteur.

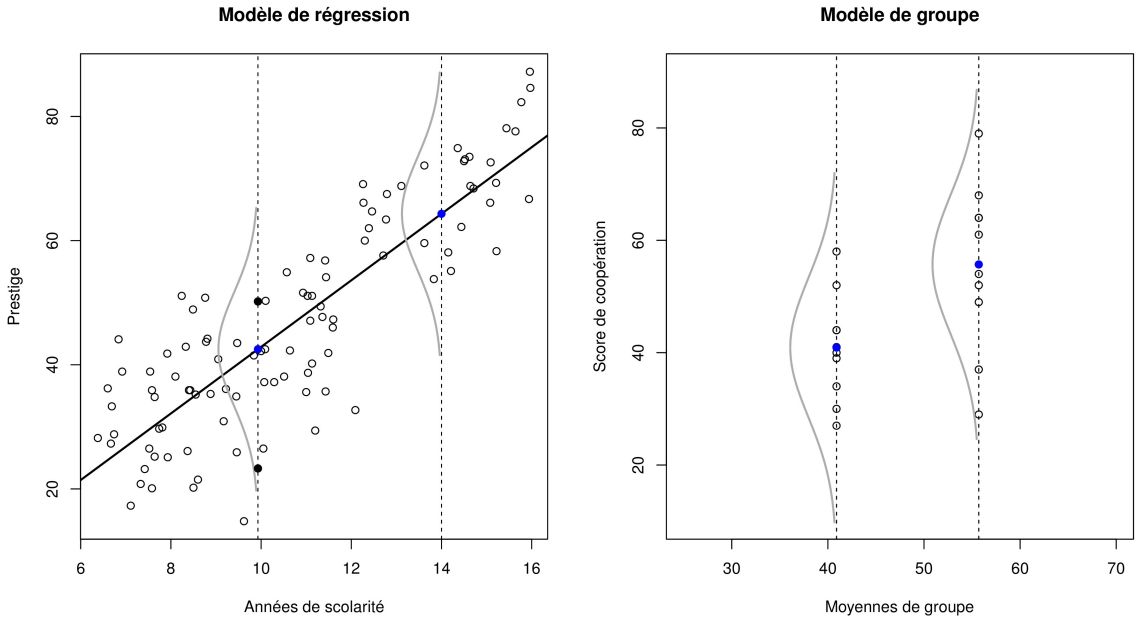
d'une distribution sur les prédicteurs. Nous appellerons *sous-modèle de structure* cette première partie du modèle.

On prend en compte l'incertitude statistique sur les données et la possibilité, pour un même ensemble $\mathbf{x}_i$ de prédicteurs observés, d'engendrer des données y_i différentes au gré de l'échantillonnage, en posant en outre une hypothèse de distribution sur la variable dépendante (VD). Dans le modèle linéaire général, l'hypothèse de distribution sur Y est gaussienne, conditionnellement aux valeurs observées des variables indépendantes (VI) :

$$Y|\mathbf{x}_i \sim N(\mu_i, \sigma^2) \qquad \text{(Equation 2)}$$

On note que l'hypothèse de distribution ne porte pas sur la distribution marginale de Y (un test de normalité sur la VD n'aurait pas de sens dans ce contexte), mais sur la distribution conditionnelle de $Y|\mathbf{x}$, c'est-à-dire de Y pour un ensemble de valeurs particulières $\mathbf{x}$ prises par les X_j (par exemple pour une observation donnée). Selon ce modèle, la moyenne de la distribution gaussienne conditionnelle change avec les valeurs $\mathbf{x}$, mais la variance reste identique (ce qui est raisonnable si on interprète cette variance comme effet d'une erreur de mesure indépendante des prédicteurs). Ce point est illustré graphiquement sur la Figure 1 (panneau de gauche), où l'on montre la régression linéaire simple du prestige perçu de 102 professions sur le niveau d'études requis (en années) pour l'exercer (Fox, 2008).

Figure 1. Distribution gaussienne conditionnelle de la VD pour des valeurs fixées de VI.



Ce graphique illustre trois aspects fondamentaux d'un modèle linéaire : i) l'hypothèse de la linéarité de la relation entre la valeur prise par le prédicteur et la moyenne conditionnelle (ou locale) de la variable dépendante, ii) l'hypothèse de normalité de la variable dépendante pour une valeur choisie de prédicteur (d'où la représentation verticale de la loi normale sur la Figure 1, pour les valeurs exemple $x = 10$ et $x = 14$), et iii) l'hypothèse d'homogénéité de la variance de ces lois normales conditionnelles (voir l'étalement comparable des deux lois normales exemple sur le graphique), qui est attendue si la variabilité est imputable à de

l'erreur. Ces trois aspects sont importants à saisir pour comprendre les extensions non gaussiennes non linéaires du modèle de régression.

L'hypothèse de distribution conditionnelle dans la régression est parfois présentée comme une hypothèse de loi normale sur la « distribution des résidus ». En effet, si la variable dépendante Y suit une loi normale conditionnelle de moyenne μ_i, alors la variable résiduelle $\epsilon_i = Y - \hat{Y}$ suit également une loi normale, mais de moyenne nulle :

$$\epsilon_i \sim N(0, \sigma^2) \qquad \text{(Equation 3)}$$

Cette propriété peut être utilisée pour procéder à un test de normalité unique sur les résidus observés $e_i = y_i - \hat{y}_i$. Il est par contre risqué conceptuellement d'adopter définitivement ce langage car, comme nous le verrons, il ne se généralise pas aux autres distributions. Seule la formulation conditionnelle reste valable dans tous les cas.

L'analyse de la variance

Le formalisme précédent est assez général pour pouvoir exprimer le modèle, traditionnel en psychologie, d'analyse de la variance (ANOVA). Méthodologiquement parlant, la situation d'ANOVA renvoie à la comparaison de distributions sur des populations différentes. Cette situation conduit à modéliser la relation (non paramétrique) entre une VD numérique continue et une VI catégorisée, par exemple un facteur de groupe G. La notion de proportionnalité n'a bien entendu plus de sens avec une VI qualitative, mais certains recodages numériques bien choisis des modalités de VI rendent possible une reformulation de type régression, pour laquelle les coefficients auront du sens.

Dans le cas simple à deux groupes, contrôle et expérimental, on peut par exemple construire la variable artificielle I (ou variable indicatrice) qui prend la valeur 1 pour un sujet du groupe expérimental et 0 pour un sujet du groupe contrôle. Le modèle structural de l'ANOVA fishérienne s'écrit :

$$\mu_j = \beta_0 + \beta_1 I \qquad \text{(Equation 4)}$$

avec pour modèle de distribution $Y|G_j \sim N(\mu_j, \sigma^2)$, pour un groupe j donné ($j = 1, 2$). La signification des coefficients dans ce modèle de régression sur des prédicteurs artificiellement construits apparaît si l'on détaille les moyennes attendues dans ce modèle pour l'un et l'autre groupe :

$$\begin{cases} \mu_1 &= \beta_0 \\ \mu_2 &= \beta_0 + \beta_1 \end{cases} \qquad \text{(Equation 5)}$$

On voit que le coefficient β_0 de cette régression spéciale n'est autre que la moyenne attendue du groupe contrôle et, en réarrangeant, que le coefficient β_1 représente la différence $\mu_2 - \mu_1$ des moyennes attendues des deux groupes. Dans ce modèle, un test de Student sur la différence de β_1 à la norme 0 est donc un test direct de la différence des moyennes. On montrerait facilement que l'expression analytique de ce test, comme test sur un coefficient de régression, est exactement celle d'un *T* de Student de comparaison de deux moyennes, avec hypothèse d'homogénéité des variances. Le principe de recodage en indicatrices permet facilement de traiter la situation à plus de deux groupes, en créant autant d'indicatrices que nécessaire ($K - 1$ pour K groupes).

A nouveau, on voit que c'est la distribution *à l'intérieur d'un groupe donné* qui est supposée gaussienne dans ce modèle, de moyenne spécifique (et non la distribution marginale de la VD, tous groupes confondus). Ce point est illustré Figure 1 (à droite) à

partir des données d'une expérience de psychologie sociale sur le comportement de coopération dans deux conditions expérimentale, anonyme et publique (Fox & Guyer, 1974).

L'analyse de la covariance

L'usage des variables indicatrices, qui donne sens à une régression sur variables catégorisées, ouvre la voie à une régression où prédicteurs numériques et catégoriels peuvent apparaître conjointement dans le modèle de structure. Si l'on étudie par exemple la perception du risque chez des jeunes (Y), en cherchant à la mettre en relation avec l'estime de soi (X), les deux mesures étant supposées numériques, il peut être intéressant de regarder si cette possible relation apparaît différente chez les garçons et les filles. En codant le genre dans une indicatrice ($I = 1$ chez les hommes), le modèle général d'un score dans cette situation peut s'écrire :

$$\mu_i = \beta_0 + \beta_1 X + \beta_2 I + \beta_3 XI \qquad \text{(Equation 6)}$$

Il intègre naturellement l'effet supposé linéaire de l'estime de soi (modulé par β_1), un éventuel effet de niveau de groupe (modulé par β_2, par rapport au niveau de base β_0) et une interaction entre les deux, sous la forme d'un effet supplémentaire qui n'apparaîtrait que chez les hommes (modulé par β_3). Dans cette situation, plusieurs hypothèses psychologiques sont possibles, qui mènent à des modèles différents (nommés M_0 à M_4 sur la Figure 2), selon les contraintes correspondantes posées sur les coefficients. Par exemple :

i) il n'existe pas de relation entre estime de soi et perception du risque, quel que soit le groupe ($\beta_1 = \beta_2 = \beta_3 = 0$),

ii) il n'y a pas de relation entre estime de soi et perception du risque, dans aucun groupe, mais les niveaux moyens de perception du risque diffèrent chez les hommes et les femmes ($\beta_1 = \beta_3 = 0$),

iii) il y a une relation linéaire entre estime de soi et perception du risque, identique dans les deux groupes, et les niveaux généraux de perception du risque diffèrent dans les deux groupes ($\beta_3 = 0$),

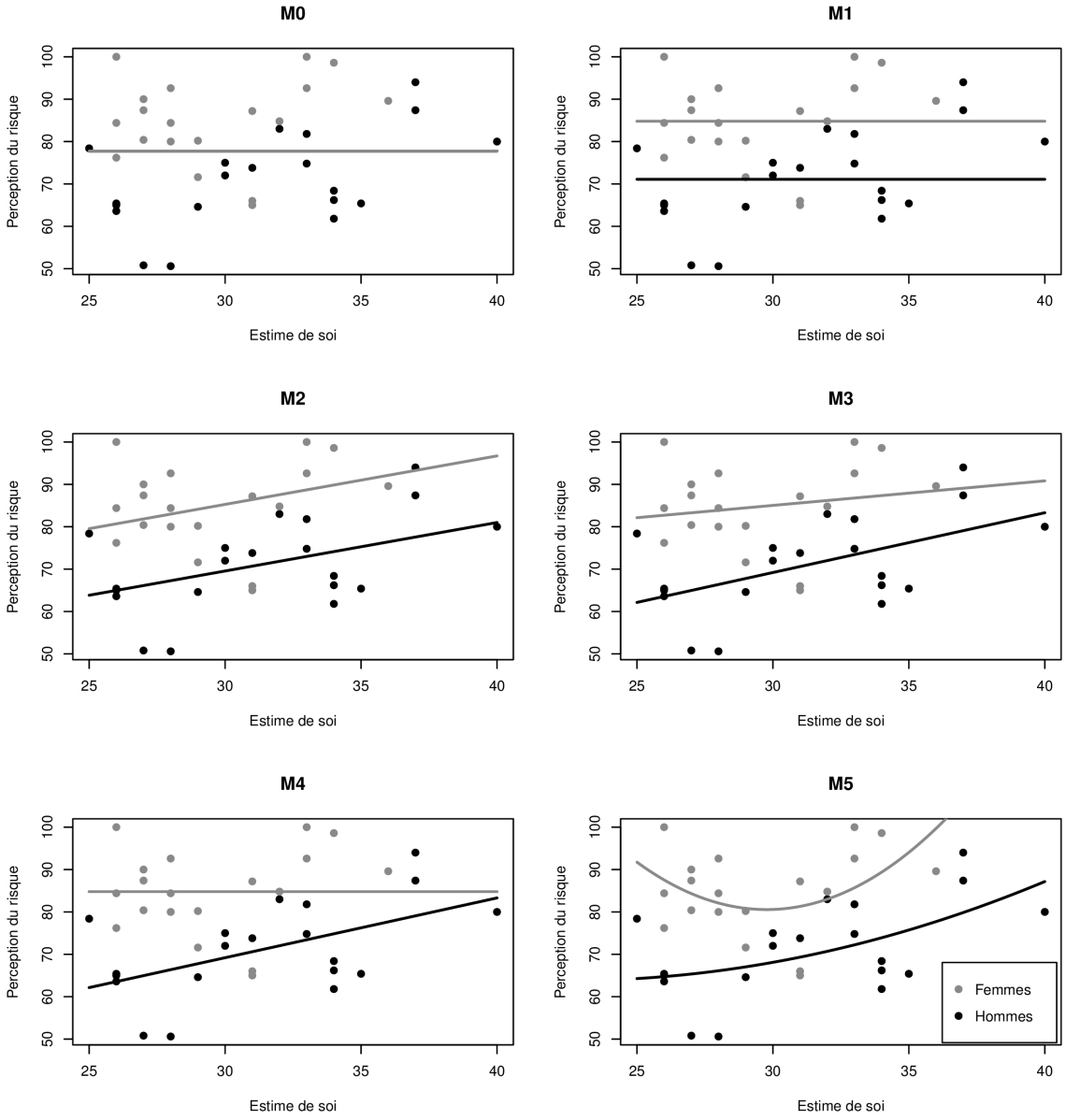
iv) les deux groupes diffèrent à la fois en intensité de la relation et en niveau général de perception du risque (aucune contrainte sur les paramètres)

v) la relation entre estime de soi et perception des risques n'existe que chez les hommes et pas chez les femmes, et leurs niveaux moyens de perception du risque diffèrent ($\beta_1 = 0$).

On note que cette approche qui intègre simultanément variables numériques et catégorisées dans le même modèle n'équivaut pas à faire des régressions séparées sur chacun des groupes. Faire des régressions séparées conduirait à estimer une variance d'erreur différente dans chaque analyse, et la comparaison statistique des pentes de régression ou des intercepts d'un groupe à l'autre ne serait pas possible. Il est donc primordial, dans la modélisation d'un jeu de données, de construire un unique modèle, à l'intérieur duquel les hypothèses psychologiques sont testées statistiquement comme des contraintes (de valeur ou d'égalité) sur les paramètres. Cette approche, qui construit naturellement des séquences de modèles emboîtés les uns dans les autres, permet de faire toutes les comparaisons possibles et protège en outre[2] contre l'inflation de l'erreur de type I qui découlerait de tests multiples séparés sur des effets locaux.

[2]Une discussion sur la collinéarité des prédicteurs dépasse le cadre de ce chapitre.

Figure 2. Modèles gaussiens linéaires et non-linéaires à prédicteurs numériques et catégorisés.



La régression polynomiale

Dans l'exemple précédent, on pourrait être tenté, pour des raisons théoriques ou en étant guidé par les données, de supposer une relation plus complexe entre Estime de soi et Perception du risque. Si la conscience du risque émerge plus facilement à la fois chez les sujets à faible et à haute estime d'eux-mêmes, pour des raisons potentiellement différentes (par exemple crainte vs. préservation de soi), on attend une relation en 'U' entre les deux variables. Cette relation peut par exemple être paramétrée comme une fonction parabolique. Si l'on note Y_{ij} et X_{ij} la perception du risque et l'estime de soi du sujet i dans le groupe j, cette relation s'écrit :

$$\hat{Y}_{ij} = a_j(X_{ij} - b_j)^2 \qquad \text{(Equation 7)}$$

Les paramètres a_j et b_j sont groupe-spécifiques et permettent de modéliser une fonction en 'U' différentes dans les deux groupes, tant en position (de centre b_j) qu'en incurvation (de pente a_j). Une telle relation est facilement ré-exprimée sous une forme linéaire en procédant au développement puis au changement de variables :

$$\hat{Y}_{ij} = a_j(X_{ij}^2 - 2b_jX_{ij} + b_j^2) = \beta_2^{(j)}X_{ij}^2 + \beta_1^{(j)}X_{ij} + \beta_0^{(j)} \qquad \text{(Equation 8)}$$

En régressant simultanément sur le prédicteur et son carré[3], on obtient donc indirectement une régression parabolique pour chaque groupe j (avec $\beta_2^{(j)} = a_j$, $\beta_1^{(j)} = -2a_jb_j$ et $\beta_0 = a_jb_j^2$). L'ajustement correspondant est illustré sur le sixième panneau de la Figure 2 (modèle M_5).

L'approche par comparaison de modèles

Les pratiques inférentielles en psychologie se sont longtemps appuyées sur un ensemble de « tests statistiques », communément associés à des « types de problèmes » (comparaison de deux moyennes par un *T* de Student, de deux proportions par une statistique *Z*, de deux variances par un *F* de Fisher, etc.). Cette approche, largement diffusée dans l'enseignement, permet d'apporter des réponses rapides dans des situations expérimentales relativement simples qui se réduisent à des comparaisons de conditions expérimentales.

Comme on le voit dans l'exemple précédent, elle est évidemment insuffisante quand l'objet de l'analyse est de modéliser un phénomène observé, en même temps que de répondre à des questions théoriques sur l'existence d'effets. On ne peut par exemple conclure à une différence entre les hommes et les femmes quant au lien Estime de soi / Prise de risque sans dans le même temps modéliser de façon paramétrique cette liaison, et poser en outre une hypothèse de distribution sur la VD conditionnellement à la VI. C'est l'ensemble de ces deux niveaux d'hypothèse, structurale (sur la forme fonctionnelle de lien, incluant les effets attendus) et distributionnelle (sur la forme fonctionnelle de distribution de probabilité ou de densité) qui constitue ce qu'on appelle un *modèle de régression*.

Naturellement, la mise en concurrence de plusieurs modèles, comme autant de scénarios scientifiques alternatifs, suppose une évaluation comparative pour sélectionner le « meilleur modèle ». Dans un modèle gaussien, la déviance (somme des résidus au carré) suit (à un facteur d'échelle près) une loi de χ^2, de même que la différence entre les déviances de deux modèles emboîtés. Il est donc possible de construire une statistique de type *F* de Fisher en construisant le rapport de ces deux variables χ^2, préalablement divisées par leurs degrés de liberté (voir Noël, 2013, p. 269). Le tableau 1 résume les différentes statistiques *F* de comparaison de chaque modèle à son successeur, dans la séquence des modèles rangés par complexité croissante.

La lecture de ces statistiques *F* est comparative : elles mesurent l'importance de la réduction de la déviance apportée par un terme supplémentaire dans le modèle, au regard de l'erreur de mesure du modèle. On voit par exemple que, par rapport au modèle constant M_0, l'introduction d'un effet de genre dans M_1 se traduit par une réduction significative de la déviance ($R_{01} = 6086,8 - 4260,9 = 1825,97$ pour $38 - 37 = 1$ d.d.l. soit $F_{1,37} = 18,9662$, $p < 0,00013$).

Cette table de comparaisons, dite table d'analyse de la déviance, fait apparaître cependant un problème dans l'approche incrémentielle par *F* de Fisher. Comme on le voit sur la ligne

[3]Eventuellement centrés pour limiter les problèmes de collinéarité.

du modèle M_4, ou aucun *F* n'apparaît, la construction de la statistique est impossible si deux modèles ont le même nombre de paramètres. Dans ces cas en effet, les déviances des deux modèles successifs ont le même nombre de degrés de liberté, la différence de ces d.d.l. est nulle et la division par cette différence dans la construction de la statistique *F* n'est donc pas possible.

Tableau 1. Comparaison de modèles gaussiens

Modèles	d.d.l. résiduels	Déviance résiduelle	d.d.l. différence	Diff. en déviance	*F* de Fisher	*p*	*BIC*
M_0	38	6086,8					314,9669
M_1	37	4260,9	1	1825,97	1896,62	0,1399	304,7210
M_2	36	3596,7	1	664,22	6,8993	0,00003	301,7752
M_4	36	3575,1	0	21,58			301,5405
M_3	35	3520,4	1	54,70	0,5681	0,4563	304,6027
M_5	33	3177,1	2	343,31	1,7830	0,1840	307,9281

Ce problème ne se pose pas quand on adopte une approche bayésienne de comparaison de modèles, par le facteur de Bayes (Kass & Raftery, 1995). Le facteur de Bayes est une statistique qui compare deux modèles en faisant simplement le rapport de leurs vraisemblances (c'est-à-dire de la probabilité des données observées, d'après chacun des modèles). Cette approche est conceptuellement différente de celle, usuelle, par valeur *p*, qui calcule la probabilité de données i) aussi extrêmes (par exemple la réduction de la déviance), ii) sachant un certain modèle de référence arbitrairement supposé vrai (l'hypothèse dite nulle). En calculant pour chaque modèle sa vraisemblance, on obtient une statistique plus simple à comprendre (on choisira simplement le modèle le plus vraisemblable), sans avoir à supposer que l'un des modèles est vrai *a priori*. A cet égard, il est donc tout à fait possible de conclure qu'une hypothèse (dite) nulle est la plus vraisemblable dans une certaine situation. Cette statistique est en outre *consistante* : la probabilité de choisir le bon modèle, s'il est dans l'ensemble testé, tend vers 1 quand la taille de l'échantillon tend vers l'infini (voir Raftery, 1995, pour une discussion détaillée).

Le tableau 1 fournit une approximation de (-2 fois le log de) la vraisemblance des modèles par la statistique simplifiée *BIC* (Schwarz, 1978). La décision par *BIC* se révèle très bonne dans la comparaison des modèles gaussiens (Kuiper & Hoijtink, 2010). Elle se lit sur une échelle (logarithmique) inversée : c'est le *BIC* le plus faible qui révèle le modèle le plus vraisemblable. Dans cette analyse, on gardera donc le modèle M_4, qui affirme qu'un lien positif Estime de soi / Perception du risque existe bien, d'allure simplement linéaire, mais uniquement chez les hommes. On note que cela revient à *affirmer* que ce lien *n'existe pas* chez les femmes, une interprétation rendue possible par le cadre de décision bayésien.

Les modèles linéaires généralisés

La régression binomiale

Dans une réplication de l'expérience classique de Solomon-Wynne (Solomon, Kamin & Wynne, 1953), trente chiens apprennent à éviter un choc électrique. Les chiens sont dans

une cage à double compartiment, dont l'un est à plancher électrifié. A chaque essai, la lumière s'éteint, la barrière inter-compartiments s'ouvre, et un choc intervient 10 secondes plus tard. S'il a déjà vécu la situation, le chien a donc 10 secondes pour sauter dans l'autre compartiment, non électrifié. Du point de vue de la théorie de l'apprentissage, nous sommes ici dans le cadre du conditionnement opérant, avec renforçateur négatif. Chaque chien a été soumis à 25 essais. Les nombres de chiens qui sautent, par essai, sont rapportés dans le tableau 2. On cherche à montrer qu'il y a bien apprentissage.

Tableau 2. Une expérience de conditionnement opérant

Essais	0	1	2	3	4	5	6	7	8	9	10	11	12
Sauts	0	3	4	5	12	12	16	18	20	23	21	19	26
Essais	13	14	15	16	17	18	19	20	21	22	23	24	
Sauts	21	26	27	27	28	29	30	30	29	30	30	29	

Modèles de groupe

La question revient à argumenter que la proportion (vraie, inconnue) de chiens qui sautent augmente significativement avec les essais. La figure 3 (en haut à gauche) illustre une première tentative de modélisation de cette évolution par un modèle M_1, de la classe des modèles linéaires gaussiens, discutée à la section précédente.

Sans même évoquer la qualité de l'ajustement, il y a au moins quatre raisons de ne pas accepter le principe même de cette modélisation : i) les données traitées sont bornées sur $[0;30]$ et une loi de distribution sans bornes ne convient donc pas (on voit comment la normale sur les données à l'essai 20 se trouve tronquée par la borne supérieure), ii) les données sont des valeurs discrètes et une loi de distribution continue est donc inappropriée, iii) pour certaines valeurs d'essai (25 par exemple), une fonction d'évolution linéaire amène des prévisions qui sortent de l'intervalle des comptages possibles, ce qui n'a pas de sens, et iv) des effets de bords sont attendus près des bornes naturelles 0 et 30 de la VD, et l'hypothèse d'homogénéité de la variance d'erreur apparaît inadaptée.

Sur des données de type comptage, il est plus raisonnable d'utiliser l'un des modèles de distribution spécifiquement conçu pour les données discrètes doublement bornées, par exemple la loi binomiale. A chaque essai j, on peut considérer le nombre de chiens qui sautent parmi N comme la réalisation d'une variable binomiale X_j de probabilité inconnue π_j , si les comportements des chiens sont bien indépendants. La probabilité d'un certain nombre k de chiens parmi N qui sautent à l'essai j est alors donnée par :

$$P(X_j = k|\pi_j) = C_N^k \pi_j^k (1-\pi_j)^{N-k} \qquad \text{(Equation 1)}$$

La question de savoir s'il y a apprentissage devient alors celle de comparer un modèle où cette probabilité est constante contre un modèle où elle croît avec les essais.

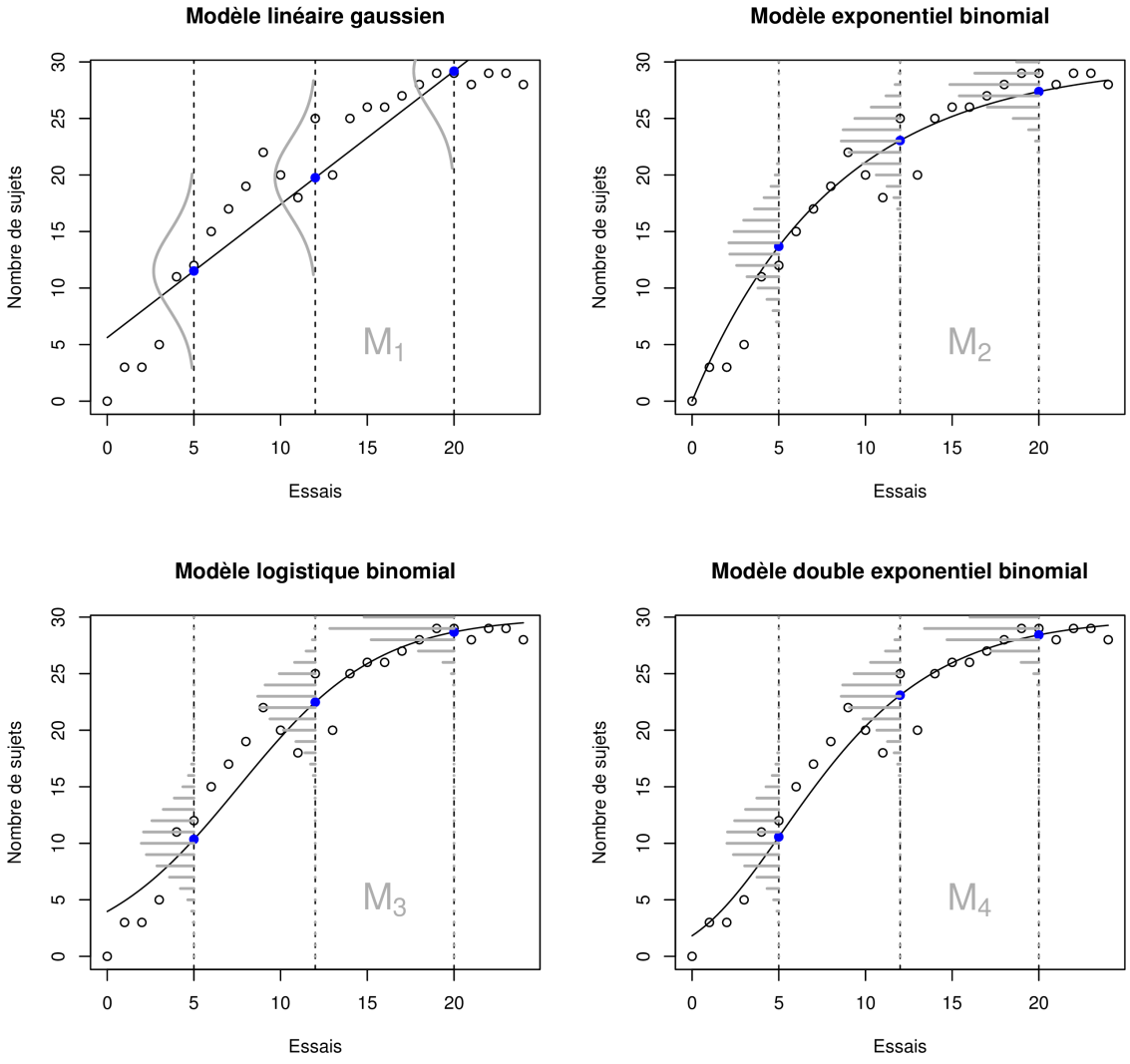
Naturellement, cette croissance doit elle aussi faire l'objet d'une hypothèse fonctionnelle qui prend en compte la nature bornée des données. Il est peu raisonnable de penser que la probabilité puisse jamais atteindre la valeur 1 car aucun comportement animal n'est jamais définitivement fixé. Une fonction qui tend vers l'asymptote 1 est donc souhaitable. Par contre, on peut discuter la pertinence d'un modèle qui affirmerait que la probabilité de sauter à l'essai 0 (en l'absence de toute expérience dans la situation) est nulle. Si on acceptait cette hypothèse, il faudrait une fonction d'apprentissage qui passe par le point

(0,0), puis croît vers l'asymptote 1 avec les essais. Un candidat possible pour cela est une fonction exponentielle complémentaire de la forme :

$$\pi_j = 1 - \exp\left[\beta_1 j\right] \qquad \text{(Equation 2)}$$

avec $\beta_1 < 0$ (figure 3, modèle M_2).

Figure 3. Quatre modèles de l'évolution d'un effectif.



Si l'on pense peu réaliste d'avoir à supposer que la probabilité d'un comportement puisse être nulle, même en l'absence de toute raison de sauter, on peut supposer un modèle qui tend aussi vers une asymptote 0, à gauche. Dans ce cas, un modèle candidat possible est la fonction logistique :

$$\pi_j = \frac{\exp\left[\beta_0 + \beta_1 j\right]}{1 + \exp\left[\beta_0 + \beta_1 j\right]} \qquad \text{(Equation 3)}$$

avec $\beta_1 > 0$, qui possède cette propriété d'avoir une double asymptote en $y = 0$ et $y = 1$, et de tendre vers ces limites à la même vitesse (figure 3, modèle M_3). Si l'on pense que cette symétrie est peu souhaitable, et que la probabilité de la réponse comportementale

croît plus vite vers 1 qu'elle ne décroît vers 0, on peut vouloir tester un modèle de double exponentielle de la forme :

$$\pi_j = \exp\left[-\exp[-(\beta_0 + \beta_1 j)]\right] \qquad \text{(Equation 4)}$$

avec $\beta_1 > 0$ (figure 3, modèle M_4).

On note que ces différents modèles comportent tous un terme linéaire en j, à l'intérieur d'une fonction monotone. On peut faire apparaître plus clairement ce terme linéaire en écrivant ces modèles sous la forme :

$$\begin{aligned} -\log\left[-\log \pi_j\right] &= \beta_0 + \beta_1 j \\ \log\left[\frac{\pi_j}{1-\pi_j}\right] &= \beta_0 + \beta_1 j \\ \log\left[1-\pi_j\right] &= \beta_0 + \beta_1 j \end{aligned} \qquad \text{(Equation 5)}$$

On parle de modèle linéaire généralisé (GLM) quand on peut réécrire le modèle sous cette forme linéarisée :

$$g(\mu_j) = \beta_0 + \beta_1 j \qquad \text{(Equation 6)}$$

où g est une fonction de lien (strictement parlant, l'inverse de la fonction de régression) bien choisie. La fonction de lien ne doit pas être vue comme une fonction de transformation, car elle ne porte pas sur les données, mais sur l'espérance conditionnelle (théorique) de la VD. Le cas où g est la fonction identique $g(\mu) = \mu$ nous ramène au modèle linéaire. On note que ces modèles sont donc en général *non linéaires*, mais facilement linéarisables par une fonction de lien monotone.

Dans le cas présent, nos modèles ne sont pas strictement parlant définis pour la moyenne conditionnelle du nombre de sauts, mais pour la probabilité de sauter. Mais cela correspond sur l'échelle des comptages (et non plus de la probabilité de sauter) à un modèle sur le nombre espéré (effectif théorique) de chiens qui sautent, de la forme :

$$\mu_j = N\pi_j \qquad \text{(Equation 7)}$$

Les trois modèles binomiaux correspondants sont représentés sur les sous-panneaux 2, 3 et 4 de la Figure 3. Au final, les quatre *BIC* de ces modèles sont respectivement de 132.29, 101.61, 96.83 et 95.54. Le modèle de double exponentielle est donc statistiquement le meilleur des quatre, ce que l'examen graphique pouvait laisser supposer.

On note qu'une différence importante que ces modèles présentent par rapport à un modèle gaussien est que, par propriété de la binomiale, la variance conditionnelle des données (à un essai j donné) est entièrement déterminée par la moyenne :

$$\sigma_j^2 = N\pi_j(1-\pi_j) = \mu_j\left(1 - \frac{\mu_j}{N}\right) \qquad \text{(Equation 8)}$$

Dans un modèle binomial, la variance est une fonction quadratique de la moyenne et ne représente donc pas un paramètre séparé du modèle. Elle est faible pour les valeurs de probabilité du comportement proches de 0 ou de 1, et élevée pour les valeurs intermédiaires, avec un maximum pour $\pi_j = 0.5$. Cela rend bien compte du fait qu'au voisinage des valeurs limites 0 et 1 de π_j, la distribution de la VD s'écrase sur ces bornes et est donc nécessairement moins étalée (comparer par exemple les distributions binomiales conditionnelles pour les essais $j = 5$ et $j = 20$ dans le modèle retenu). C'est la raison pour laquelle sur des données binomiales, vouloir calculer un écart-type, vouloir

centrer-réduire ou vouloir faire une ANOVA (qui suppose l'homogénéité) n'a aucun sens. Il y a une hétérogénéité de la variance qui est naturelle dans ces modèles, et la distribution binomiale, associée à une fonction de réponse bornée bien choisie, prennent cela naturellement en compte.

On note aussi qu'on ne peut ici parler d'une distribution binomiale des *résidus* : les écarts à la courbe modèle sont positifs ou négatifs, et la loi binomiale n'est définie que pour des valeurs positives entières. L'équivalence de famille de distribution de la VD et des résidus n'est retrouvée que dans le cas gaussien, comme indiqué plus haut.

Au final un modèle binomial constant (fixant $\beta_1 = 0$ dans le modèle) est de *BIC* 354.19 : il y a donc bien évolution de la probabilité de sauter dans cette expérience. Que la double exponentielle, qui tend vers 1 beaucoup plus rapidement que la logistique, ait été ici retenue fait également sens au regard du paradigme classique de Solomon-Wynne : ils notaient dès leurs premières expériences que l'acquisition de la réponse était rapide et l'extinction quasiment non-mesurable, même sur de longues périodes de temps (Mosteller, 2010, p. 38).

Modèles individuels

L'approche par loi générale d'apprentissage peut paraître limitative. Dans une approche plus différentielle, on peut penser que les chiens ont une fonction d'apprentissage individuelle. Certains d'entre eux peuvent être plus réactifs par nature et acquérir la réponse plus tôt dans la séquence des essais. Si c'est le cas, l'analyse sur des comptages agrégés est au mieux grossière, et au pire pourrait bien masquer la structure vraie du phénomène d'apprentissage, surtout si elle est couplée à une typologie sur les chiens. De nouvelles analyses peuvent être réalisées sur un jeu de données explicitant toutes les séries individuelles de sauts et de non-sauts pour chaque chien à chaque essai. Le fichier correspondant doit contenir trois colonnes : i) les séries empilées de comportements pour chaque chien à chaque essai, ii) les numéros d'essais correspondants et iii) les identifiants sujets. Si les sauts et non-sauts sont codés par les valeurs numériques 1 et 0, les données correspondantes (Bernoulli) peuvent être traitées par les modèles binomiaux (la Bernoulli étant une binomiale à *N*=1). Le modèle porte alors sur une probabilité *individuelle* π_{ij} que le chien i saute à l'essai j.

Il existe autant de manières d'introduire une hétérogénéité de sujets dans un modèle structural qu'il a de paramètres. Dans les modèles précédents, les deux paramètres de position (β_0) et de pente (β_1) peuvent être rendus individuels. Dans l'approche la plus simple, on peut décider de rendre le paramètre de position β_0 individuel, ce qui donne, pour les modèles logistique et double exponentiel :

$$\pi_{ij} = \frac{\exp\left[\beta_{0i} + \beta_1 j\right]}{1 + \exp\left[\beta_{0i} + \beta_1 j\right]} \qquad \text{(Equation 9)}$$

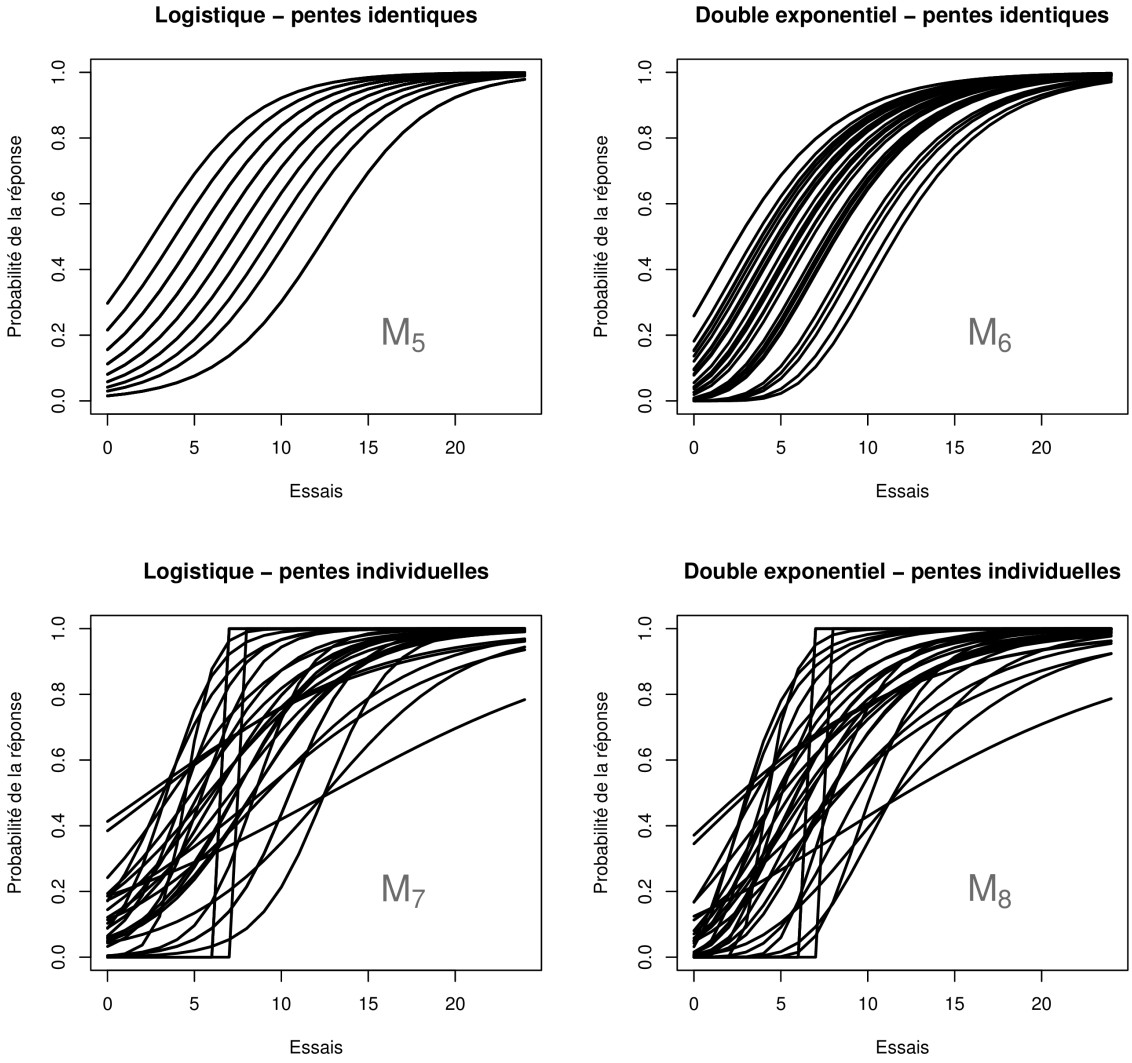
et

$$\pi_{ij} = \exp\left[-\exp[-(\beta_{0i} + \beta_1 j)]\right] \qquad \text{(Equation 10)}$$

Ces deux propositions sont représentées Figure 4 (panneaux du haut). Comme on le voit, l'hétérogénéité inter-individuelle est modélisée par un décalage horizontal des courbes d'apprentissage, certains sujets démarrant leur acquisition plus tôt dans la séquence des essais. L'avantage des propositions simples ci-dessus est que, si elles offrent une bonne description des comportements, les différences individuelles seront immédiatement

interprétables en termes de seuils d'acquisition. Par exemple, dans le modèle logistique, l'essai où la probabilité de sauter devient 0.5 peut servir de repère pour caractériser la précocité de l'apprentissage chez un sujet. Elle sera atteinte à l'essai pour lequel $\beta_{0i} + \beta_1 j = 0$ soit au moment théorique $j^* = -\beta_{0i}/\beta_1$. Ce modèle n'est pas autre chose qu'un modèle de Rasch, dans la version contrainte du *Linear Logistic Test Model* (Fisher & Forman, 1982). On peut constater qu'à l'instar de tous les modèles de la famille de Rasch, il modélise autant de courbes d'apprentissages qu'il y a de *scores individuels* (nombres de sauts par chien) différents, et non de sujets : il y a 30 chiens, ayant sauté de 12 à 21 fois (sans le 13), soit 9 courbes modèles distinguées (voir Figure 4, premier panneau).

Figure 4. Modèles individuels de l'apprentissage d'une réponse



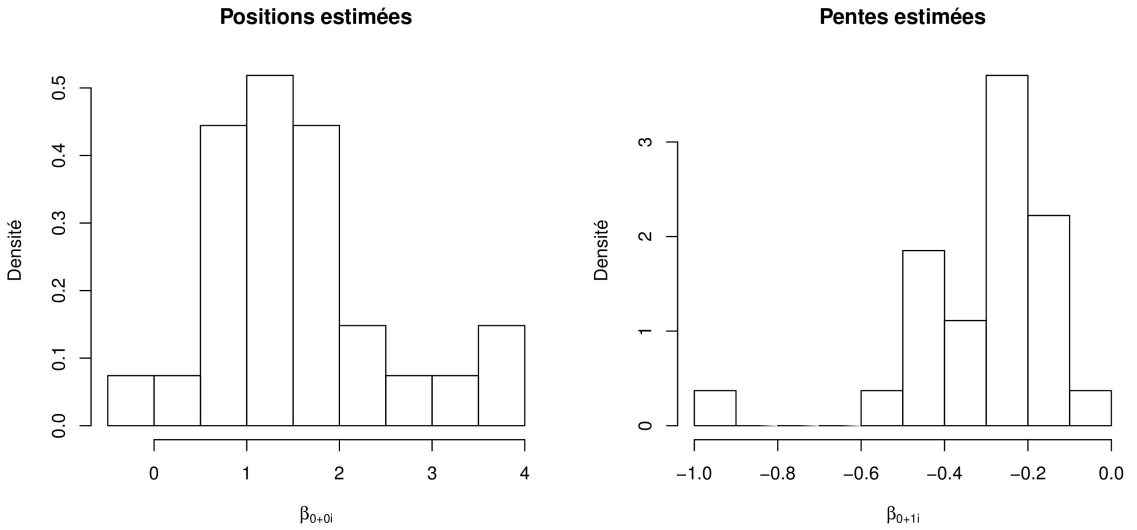
C'est une propriété purement mathématique liée à la forme logistique de la fonction choisie, qui a son avantage quand le modèle de Rasch décrit bien les données : on sait que les paramètres estimés $\hat{\beta}_{0i}$ sont liés de façon simple aux comptages des comportements de chaque sujet et c'est une statistique très simple à utiliser en pratique, sans avoir besoin de logiciel. En contrepartie, c'est aussi une rigidité du modèle car sa capacité de discrimination en termes de position horizontale des courbes s'en trouve affectée. On peut l'apercevoir visuellement en comparant les deux premiers graphiques de la Figure 4 (en

haut) : la double exponentielle permet de discriminer plus finement la précocité de l'acquisition de la réponse.

Dans la même logique, on pourrait être tenté de rendre individuels à la fois les positions et les paramètres de pente (β_1) des courbes. Ici comme ailleurs, il n'est pas certain qu'on ait intérêt à donner aux modèles statistiques du comportement trop de souplesse, car ce qu'on gagne en qualité descriptive ou vraisemblance de modèle, on le perd simplement en interprétabilité. L'hétérogénéité est dans ces modèles intégrée à la fois en pente et en intercept, ce qui ne permet plus d'en donner une interprétation simple. Dans le modèle logistique par exemple, la valeur repère $j^* = -\beta_{0i}/\beta_{1i}$ ne représente plus un seuil comparable de sujet à sujet sur la même dimension, car la variabilité sur β_{1i} correspond à autant d'échelles de temps différentes pour les sujets.

Ces deux modèles libres, pour les fonctions de réponse logistique et double exponentielle, sont représentés sur la Figure 4 (en bas). Comme on peut le voir, elles font apparaître un autre problème potentiel de la libération des deux paramètres : pour les séquences de comportements constituées d'une série ininterrompue de 0 puis de 1, cette souplesse du modèle amène un sur-ajustement sur les données observées. La valeur estimée de β_{1i} s'enfuit vers l'infini, la fonction de réponse (estimée) est en escalier et les valeurs de probabilités prévues deviennent numériquement 1, à la précision de la machine près (ce qui n'a pas de sens mathématique). Une solution technique possible à ce problème est d'introduire une contrainte distributionnelle sur les paramètres eux-mêmes : c'est la voie des *Generalized Linear Mixed Models* (GLMM), ou modèles à effets aléatoires, ou encore « modèles mixtes » (qui mélangent des effets fixes et des effets aléatoires).

Figure 5. Distributions des paramètres individuels estimés



Modèles à effets aléatoires de la variabilité inter-individuelle

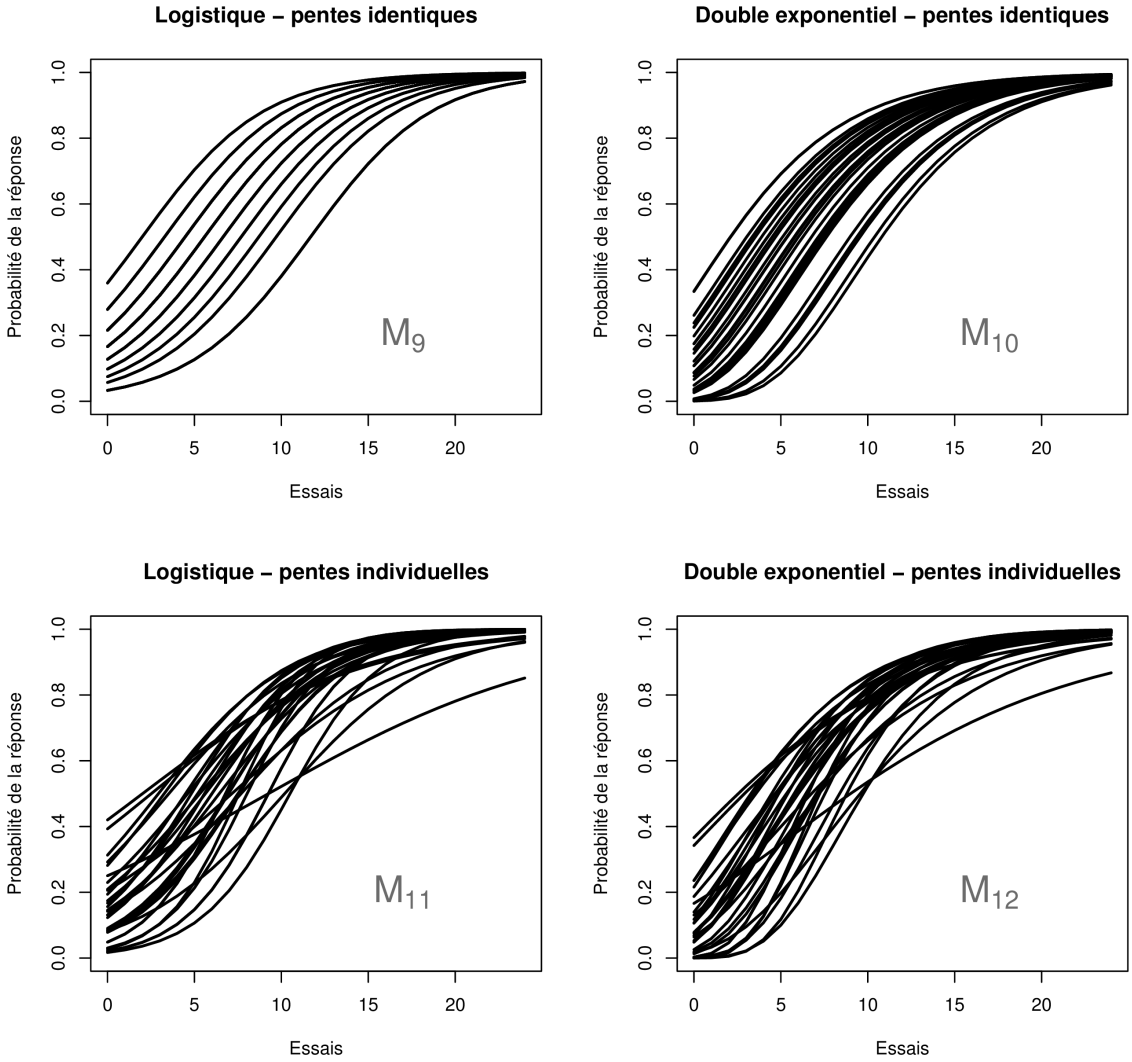
Les paramètres $\hat{\beta}_{0i}$ et $\hat{\beta}_{1i}$ estimés dans le modèle double exponentiel individuel ci-dessus varient majoritairement entre -0,5 et 4, et entre -1 et 0, respectivement, si l'on écarte les trois sujets qui ont des séquences ininterrompues de 0 puis de 1, pour lesquels les paramètres estimés prennent des valeurs aberrantes (voir Figure 5). L'examen de ces

distributions suggère de construire des modèles où non seulement les données, mais aussi les paramètres pourraient être conçus comme issus d'une distribution. Par exemple, si l'on pense que la distribution latente des positions de sujets respecte une structure massée autour d'une valeur centrale, on pourrait imposer dans le modèle que :

$$\beta_{0i} \sim N(\mu_0, \sigma_0^2) \qquad \text{(Equation 11)}$$

Naturellement, le calcul de la vraisemblance (binomiale) du modèle devra inclure cette hypothèse en intégrant une partie distributionnelle (gaussienne) sur le paramètre β_{0i} et ce mélange de distributions (sur les données, sur les paramètres) ne donnera pas toujours d'expression analytique close de la vraisemblance complète[4]. Ces calculs doivent souvent être faits numériquement à l'aide de logiciels spécialisés, mais ceux-ci sont librement disponibles (Bates et coll., 2013, Noël, 2013).

Figure 6. Modèles d'apprentissage individuel à effets aléatoires.

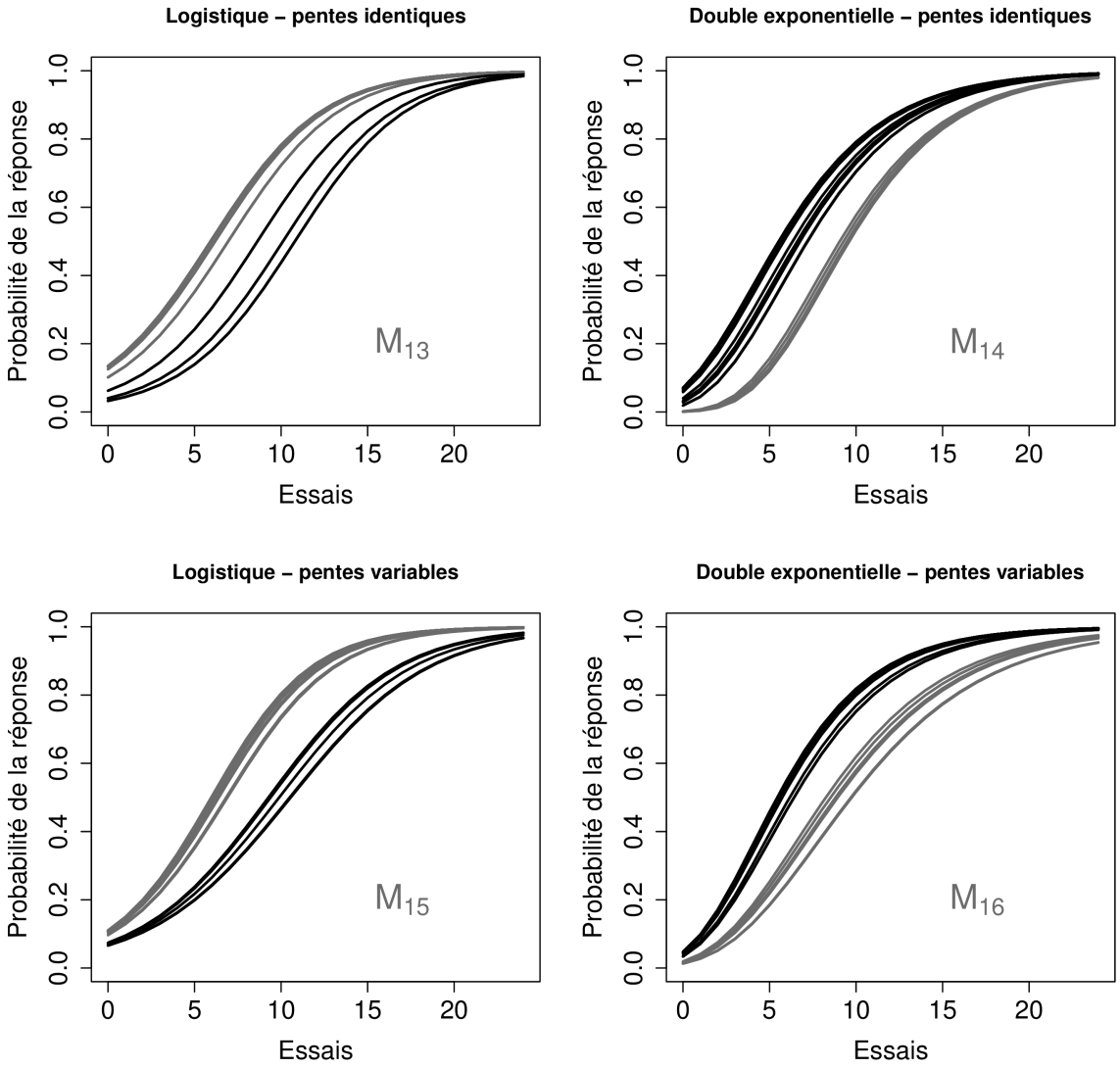


Le même type de contrainte distributionnelle peut être posée sur les pentes individuelles β_{1i} si l'on pense que ces vitesses d'acquisition de la réponse se distribuent autour d'une

[4]C'est notamment le cas ici car la seule distribution qui se « mélange » bien avec une loi binomiale des données est la loi Beta. La librairie `lme4`, utilisée ici, fait par défaut l'hypothèse d'une loi normale sur les paramètres.

valeur centrale. Le résultat de ces modélisations est représenté sur la Figure 6. Comme on le voit, cette contrainte « douce » se traduit par une régularisation des estimations qui, sans effacer l'hétérogénéité, lui donne une structure. Le problème de l'estimation numérique égale à 1 ou 0 pour les probabilités de sauter se trouve résolu. Les *BIC* de ces quatre modèles sont 635.76, 618.97, 633.16 et 624.50, à comparer avec les *BIC* des modèles logistique et double exponentiel de groupe, précédemment testés : 588.45 et 580.64[5]. L'inclusion d'effets individuels sous cette forme ne nous donne pas un meilleur modèle, du point de vue du *BIC*.

Figure 7. Modèles de régressions binomiales en deux classes latentes.



Modèles en classes latentes

L'examen des représentations graphiques des modèles précédents pourraient laisser penser que les sujets sont structurés en classes, par exemple deux classes, certains d'entre eux semblant avoir un apprentissage plus tardif (ou simplement plus lent). Nous ne disposons

[5]Ces nouvelles valeurs sont obtenues sur les modèles précédents, re-testés sur le nouveau tableau des données binaires individuelles. Les valeurs de vraisemblance et les nombres d'observations ne sont plus les mêmes, mais sont exactement proportionnelles, de sorte que la hiérarchie des modèles est préservée.

pas d'information a priori sur une telle classification sous-jacente mais, à partir d'une hypothèse sur le nombre de classes sous-jacentes, nous pourrions introduire des paramètres spécifiques de probabilités d'appartenance, à estimer dans le modèle en même temps que les paramètres de la régression. Plusieurs propositions de ce type existent dans la littérature, reposant souvent sur l'algorithme EM pour mélanges de distributions (Aitkin, 1996 ; Gruen & Leisch, 2007). On suppose que k fonctions de régression latentes déterminent la structure des données, chaque sujet relevant de l'une d'entre elle, sans qu'on puisse dire laquelle a priori. Les probabilités que le sujet relève de l'une ou de l'autre sont estimées comme des paramètres du modèle.

Cette possibilité est illustrée Figure 7 pour nos données, en supposant qu'il existe deux classes sous-jacentes de sujets. Les modèles logistique et double exponentiel ont été testés[6], ainsi que deux variantes du modèle, selon que les constantes seules où les pentes sont laissées libres de varier par classe latente. Les *BIC* de ces quatre nouveaux modèles sont 587.18, 582.03, 590.64 et 584.41, ce qui (dans le cas présent) n'est pas meilleur que le modèle double exponentiel global (580.64). Le potentiel de cette approche est important pour la psychologie différentielle, car elle permet de mener simultanément deux tâches courantes du différentialiste qui sont souvent réalisées successivement : le travail de modélisation d'un lien ou d'un comportement et l'extraction de classes de sujets ou de stratégies différentielles. Des nombres variables de classes latentes peuvent être testés dans le modèle et sélectionnés sur la base du *BIC*.

La régression poissonienne

Les données en psychologie sont très souvent des événements qualitatifs, et la modélisation statistique impose qu'une forme de numérisation préalable de telles données soient produites. Le moyen le plus simple est de dénombrer les apparitions de chaque modalité qualitative, et la modélisation portera alors sur ces comptages. On peut distinguer deux types de comptages : les comptages bornés à droite par une limite supérieure (nombres de personnes produisant un certain comportement sur un échantillon fixé), cas examiné à la section précédente, et les comptages sans borne supérieure connue (nombres de réponses produites par une personne dans un intervalle de temps fixé), examinés dans cette partie.

Modèles log-linéaires

Dans une expérience sur l'utilisation d'un robot mobile pour assister des personnes handicapées dans leurs tâches quotidiennes, on observe comment des sujets se familiarisent avec le robot qu'ils doivent amener, à l'aide d'une télécommande, vers une lieu cible dans un appartement. On observe en particulier comment évolue le nombre d'arrêts du robot par un obstacle, au cours des 12 essais. Pour l'un des sujets, on obtient les comptages : 7, 9, 15, 5, 10, 7, 6, 8, 5, 4, 4, et 3. Peut-on dire que ces nombres d'erreurs révèlent un processus d'acquisition de maîtrise au cours des essais ?

Les données sont des comptages, mais qui n'ont cette fois pas de borne supérieure connue. Un modèle pour comptages bornés comme la loi binomiale serait donc inapproprié. Le modèle de la loi de Poisson correspond exactement à ce type de données, où sont comptés sur un certain intervalle de temps fixé (le temps d'un essai) des événements arrivant à débit constant (considérés comme tels sur cet intervalle assez court).

[6] A l'aide de la librairie `npmlreg` sous R.

La probabilité que la variable de comptage X_j à l'essai j prenne la valeur fixée k est donnée par :

$$P(X_j = k) = \frac{\mu_j^k}{k!} e^{-\mu_j} \qquad \text{(Equation 12)}$$

Cette loi de probabilité est définie pour n'importe quel entier naturel, sans limite supérieure, et convient bien à la modélisation d'un comptage sans borne à droite. C'est la loi conditionnelle des comptages dans un essai donné qui est supposée Poisson de moyenne μ_j, mais s'il y a apprentissage, la moyenne (inconnue) des erreurs doit diminuer avec le temps. A côté du modèle de distribution, il nous faut donc définir un modèle structural sur la moyenne conditionnelle (ou locale) du nombre d'erreurs. Ce modèle de régression doit prendre en compte les deux particularités des données d'être sans borne supérieure mais d'avoir une borne gauche naturelle à 0. La fonction exponentielle est un candidat possible :

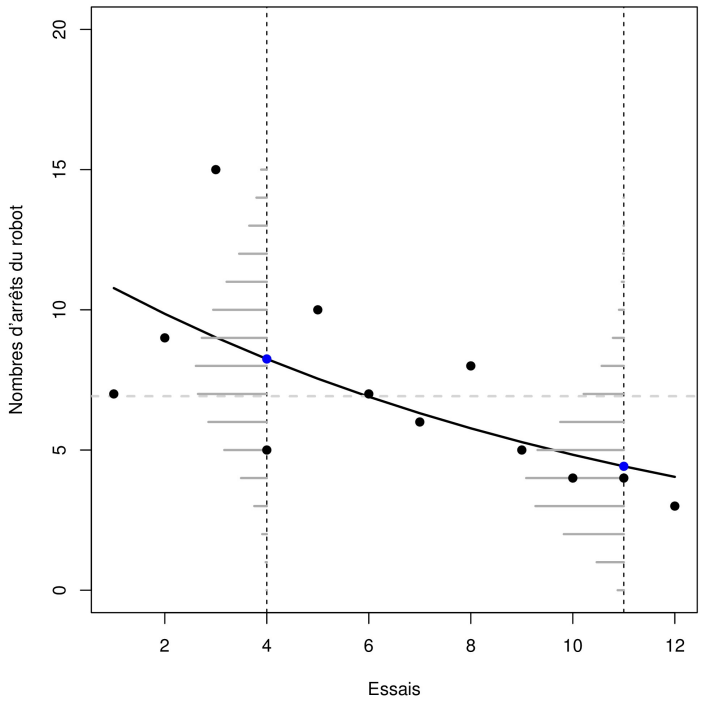
$$\mu_j = \exp[\beta_0 + \beta_1 j] \qquad \text{(Equation 13)}$$

soit encore :

$$\log \mu_j = \beta_0 + \beta_1 j \qquad \text{(Equation 14)}$$

On parle dans ces cas de GLM poissonien à lien log, ou encore de *modèle log-linéaire*. Cette fonction de moyenne conditionnelle correspond au modèle M_1 de la diminution moyenne des nombres d'erreurs au cours des essais, représenté par la courbe en noir sur la Figure 8. Au sein de ce modèle, la contrainte $\beta_1 = 0$ construit un modèle M_0 de moyenne constante au cours des essais (ligne horizontale grise). Les deux modèles sont de *BIC* 57.82 et 62.98, de sorte que l'acquisition de maîtrise est argumentable.

Figure 8. Evolution temporelle d'un nombre d'erreurs.



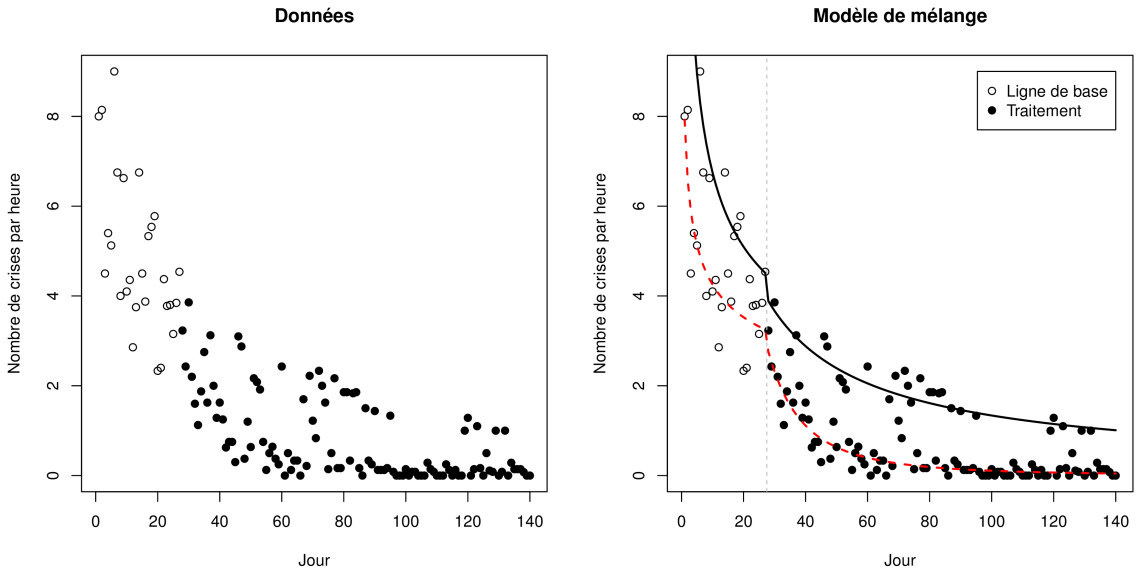
Une des propriétés notables de la loi de Poisson est que sa variance est exactement égale à sa moyenne ($\sigma_j^2 = \mu_j$), et ne représente donc pas un paramètre supplémentaire. Quand un sujet est susceptible de faire beaucoup d'erreurs en moyenne, par exemple dans les premiers essais, il est aussi susceptible d'avoir des performances plus variables, et inversement. Cette propriété est directement visible sur la Figure 8 où les distributions Poisson conditionnelles ont été représentées pour les essais $j = 4$ et $j = 11$, en superposition avec la courbe de régression exponentielle. A ces essais, les moyennes conditionnelles sont estimées (points sur la courbe) à 8.25 et 4.42. Ces deux valeurs correspondent aussi aux variances estimées des deux lois de Poisson conditionnelles correspondantes (représentées en grisé), la première étant en effet plus dispersée que la seconde.

Comme dans le cas binomial, on comprend que cette liaison fonctionnelle moyenne-variance prend en compte l'effet de borne inférieure 0 : la variance conditionnelle est nécessairement plus faible au voisinage des valeurs nulles de comptages. Cela permet de comprendre que la description courante en psychologie en moyenne et variance, comme descripteurs indépendants, est inappropriée dans de tels cas. Ce serait inapproprié aussi sous M_0, car bien que de moyenne conditionnelle constante, toute l'information sur la variance est déjà incluse dans la moyenne. Dans la classe des distributions usuelles, dites de la famille exponentielle, en réalité seule la loi normale a cette propriété d'avoir des paramètres explicites de moyenne et variance indépendants.

Modèles de mélanges de régression

Wang et al. (1996) ont étudié un enfant épileptique soumis à des crises journalières très fréquentes. Le nombre de crises par heure a été enregistré systématiquement par les parents pendants 140 jours. Les 27 premiers jours ont constitué la ligne de base, et un traitement par injection intra-veineuse d'immuno-globulines a été démarré à partir du 28ème jour. L'objectif est de tester l'efficacité du traitement dans la diminution du nombre de crise horaire.

Figure 9. Evolution du nombre horaire de crises d'épilepsie.



Ce jeu de données fait apparaître un phénomène étrange et intéressant : les comptages de crises semblent se distribuer, à un jour donné, selon une distribution bimodale (Figure 9, panneau de gauche). Sur l'ensemble de la série, cela semble suggérer l'existence (intra-sujet) d'une double série événementielle, comme si le régime des crises était à deux niveaux : les jours « hauts » et les jours « bas ». Cet exemple peut nous aider à appréhender les situations où des sujets sont susceptibles de basculer de façon imprévisible et non manifeste d'un état latent à un autre. Il pourrait s'agit d'un nombre d'erreurs qui varie selon qu'un sujet utilise l'une ou l'autre de deux stratégies de résolution d'une tâche répétée par exemple.

Le double régime latent des crises peut être modélisé par un mélange de deux régressions exponentielles poissonniennes simultanées. Cela ne représente pas deux régressions séparées, mais un couple de régressions simultanées modélisant la distribution des enregistrements, sur un seul et unique sujet, et à un jour donné, par une distribution à deux modes, dont il dépend à travers deux coefficients pondérateurs de somme 1 :

$$\mu_j = w f_1(j) + (1 - w) f_2(j) \qquad \text{(Equation 15)}$$

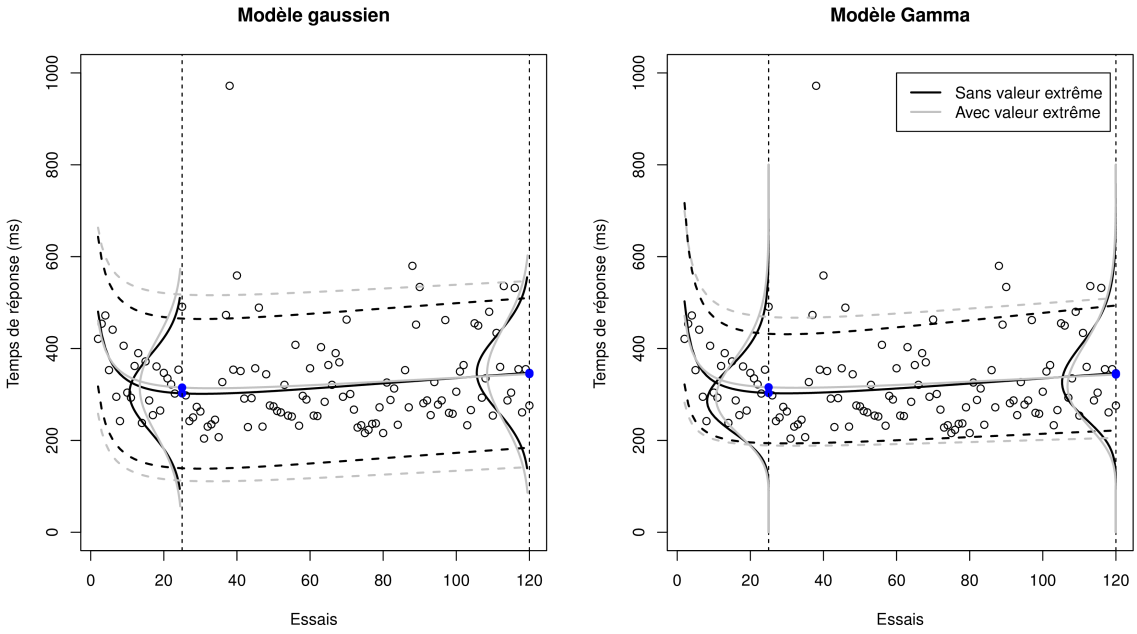
Ces coefficients peuvent être interprétés comme les probabilités que la réponse considérée relève du premier ou du second régime, à un jour donné. Pour la plupart des points sur la figure, le coefficient w est proche de 0 ou de 1 (le point est proche de l'une ou de l'autre des courbes) et pour certains points, situés entre les deux courbes (au voisinage du jour 70 par exemple), les coefficients de pondération peuvent être plus proche de 0.5-0.5 (ce qui peut être interprété comme un processus de transition vers une domination du régime bas). Au sein de ce modèle, on peut vouloir tester l'impact de l'introduction du nouveau traitement. On dira que celui-ci n'a rien changé si l'ensemble de la série, avant et après son introduction, est modélisable par la même exponentielle (à deux régimes). Un tel modèle a un *BIC* de 873.04, contre 796.83 pour le modèle de l'impact du traitement. Comme on le voit sur la figure, l'introduction du traitement a provoqué une diminution des nombres horaires de crises, plus importante que ce que laissait anticiper la (double) courbe de base.

La régression gamma

Les paradigmes de tâches répétées sont parfois utilisés pour mettre en évidence certaines caractéristiques de la variabilité intra-individuelle. La question de la structure de cette variabilité, en fonction des VI, ne peut cependant pas être pensée indépendamment du choix de la variable dépendante, car comme on l'a vu plus haut, le type de variable dépendante et le modèle de distribution choisi implique toujours une certaine forme de dépendance entre moyenne et variance conditionnelle de réponse. Si l'on souhaite isoler, pour analyse, une « variance » intra-individuelle signifiante psychologiquement, il convient donc de correctement modéliser l'hétérogénéité *structurale* de la variance uniquement imputable à la nature de la VD, en rendant explicite le lien moyenne-variance.

L'équipe d'Anik de Ribeaupierre à Genève poursuit depuis plusieurs années une étude longitudinale de grande ampleur (la *Geneva Variability Study*) sur l'évolution de compétences cognitives élémentaires au cours de la vie, avec des tâches cognitives de difficultés variées (Fagot, Chicherio & de Ribaupierre, 2013). Dans l'un des paradigmes, le sujet doit appuyer le plus rapidement possible sur un bouton dès qu'il identifie une cible visuelle sur l'écran. La figure 10 illustre l'évolution des temps de réponse au fil des essais pour un sujet âgé exemple. La nature de la variable dépendante est cette fois-ci numérique, continue et bornée à gauche par 0.

Figure 10. Evolution d'un temps de réponse à une tâche répétée.



Avec ce type de variable, on souhaite un modèle de distribution capable d'accomoder le fait que la dispersion des temps quand ils sont courts est nécessairement moindre que lorsque ceux-ci sont longs. On souhaite également définir une distribution conditionnelle du temps de réaction qui est dissymétrique à un essai fixé. On prévoit que cette dissymétrie s'estompe lorsque les temps sont longs, et que l'effet de bord à 0 s'atténue. Un candidat possible comme modèle de distribution dans ces cas est la loi Gamma du temps de réaction à l'essai j :

$$f(T_j = t|s, \lambda_j) = \frac{\lambda_j}{\Gamma(s)} (\lambda_j t)^{s-1} e^{-\lambda_j t} \qquad \text{(Equation 16)}$$

Cette loi peut s'interpréter comme celle du temps d'apparition d'une réponse qui résulte d'un processus cognitif à s étapes ou opérations (latentes), quand chacune est de durée élémentaire identique en moyenne, et de débit temporel λ_j (nombre d'étapes par unité de temps). Elle fournit donc un modèle très général du temps de réponse pour un comportement structuré en étapes (ou opérations élémentaires) dont les durées de traitement s'ajoutent. Comme dans les modèles précédents, cette hypothèse distributionnelle est posée conditionnellement au prédicteur, c'est-à-dire ici à un essai donné. Dans la loi Gamma, la variance évolue comme le carré de la moyenne. Elle tend en forme vers la loi normale lorsque les temps sont longs.

Au-delà de l'hypothèse distributionnelle, plusieurs hypothèses structurales peuvent être faites sur l'évolution du débit λ_j sous-jacent au processus de réponse (ou indirectement sur la moyenne attendue du temps de réponse) à un essai donné. Il est très courant dans ces paradigmes de tâches répétées de voir les sujets acquérir une certaine compétence dans la tâche au fil des essais, et les temps de réaction diminuent en moyenne. Sur des séries très longues, une forme de fatigue attentionnelle peut s'installer, qui compense et renverse parfois le premier processus évolutif, et on voit alors les temps ré-augmenter en fin de série. Cette hypothèse sur un double processus latent, inhérent à la réponse du sujet, est traduite ici sous la forme d'une régression Gamma polynomiale, c'est-à-dire à l'aide d'une

fonction quadratique des essais. En pratique, nous régressons sur le logarithme des essais (et son carré), plutôt que sur le numéro d'essai, en utilisant par ailleurs une fonction de régression exponentielle : cela a pour effet, dans les unités d'orgine, de produire une fonction double exponentielle, capable d'accomoder le fait que le processus d'augmentation du temps de réaction (fatigabilité) est plus lent que celui d'acquisition de compétence.

Le choix d'une loi dissymétrique permet de sous-pondérer de manière naturelle les valeurs très grandes de temps, en leur affectant une très faible densité, ce qui rend pratiquement inutile le procédé assez répandu qui consiste à éliminer certaines valeurs « aberrantes » (on n'est jamais bien sûr ce faisant de ne pas écarter des données utiles). Ce phénomène est illustré sur la figure 10, ou une régression gaussienne est comparée à une régression Gamma, dans les cas où l'on inclut ou non le temps extrême 972 ms. à l'essai 37. La courbe de moyenne s'en trouve modérément affectée, tant dans le cas gaussien que Gamma, mais la dispersion estimée des données selon le modèle gaussien s'en trouve notablement majorée (les intervalles de confiance à 95 % ont été représentés dans les deux cas pour matérialiser ce point, avec des courbes en pointillé). Cette majoration de l'erreur estimée n'est pas triviale car elle sera utilisée dans les tests sur coefficients ou de réduction de la déviance, et cela pourrait masquer des effets existants.

La régression inverse-gaussienne

Une vision alternative du temps de réponse, spécifiquement dans les tâches perceptives d'identification, est fournie par plusieurs auteurs (Hohle, 1965 ; Schwarz, 2001). On modélise le temps de réponse du sujet comme un processus d'accumulation (par incrément gaussiens) d'informations perceptives, jusqu'à atteindre un seuil minimal qui déclenche la décision. On sait qu'un tel processus, dit de Wiener (apparenté au mouvement brownien avec dérive positive en physique), mène à des temps d'atteinte du seuil qui suivent une loi inverse gaussienne (ou de Wald) :

$$f(t_j|\mu_j, s) = \sqrt{\frac{s}{2\pi t_j^3}} \exp\left[-\frac{s}{2t_j}\left(\frac{t_j-\mu_j}{\mu_j}\right)^2\right] \qquad \text{(Equation 17)}$$

Cette distribution a pour propriété que la variance est proportionnelle au cube de la moyenne. Couplée avec la fonction de régression double exponentielle, c'est le meilleur de tous les modèles de régression que nous avons testés sur cette série de temps, autant du point de vue du *BIC* que de l'examen des graphiques quantile-quantile des résidus. Mais une déviation apparaît quand même dans les temps courts, qui sont moins fréquents en réalité que ce que prédit cette distribution.

Pour corriger ce problème, Schwarz (2001) a proposé de considérer que le temps total de réponse est en fait dans ces tâches constitué d'une phase d'observation jusqu'à décision et d'une phase de réponse motrice. Ce temps supplémentaire de réponse finale, après décision, est modélisé soit par une constante (on parle de loi Wald décalée) ou par une variable aléatoire de loi exponentielle (qui est simplement la loi Gamma avec $s = 1$). Dans ce dernier cas, la somme des temps sur les deux phases mène à un temps total qui suit une loi dite ex-Wald. Cette distribution montre souvent un excellent ajustement aux distributions marginales de temps de réponse observées en psychologie (Matzke & Wagenmakers, 2009). Elle n'est actuellement pas implémentée sous forme de module de régression dans les logiciels usuels.

Conclusion

En dépit du fait que les modèles linéaires généralisés sont désormais des outils classiques en statistiques, ils ne sont encore que trop rarement intégrés aux enseignements de statistique en psychologie. Leurs extensions récentes, avec inclusion d'effets aléatoires ou de classes latentes, rendent ces outils d'une importance considérable pour le psychologue, et le différentialiste en particulier. Leur maîtrise suppose d'avoir clairement à l'esprit les hypothèses inhérentes à tout modèle de régression, en particulier les concepts de distribution, de moyenne et de variance conditionnelles. D'autres extensions, non discutées dans ce chapitre, permettent de modéliser la fonction variance de la même façon que nous avons interrogé la fonction moyenne dans les exemples qui précèdent. Elles permettent d'expliquer directement la variabilité des réponses en fonction de variables instrumentales, avec des choix distributionnels et de lien spécifiques pour ce type de paramètres (on parle alors de GLM doubles, à la fois en moyenne et en dispersion, Smyth, 1989). On comprend que dans l'absolu, n'importe quel paramètre distributionnel peut faire l'objet d'une hypothèse structurale, au moins tant qu'on peut y donner un sens psychologique.

Le choix d'un bon modèle de distribution est une étape souvent ignorée dans les analyses appliquées en psychologie, le choix par défaut de la loi normale étant trop peu souvent questionné. Dans l'idéal, la recherche d'un modèle de distribution, comme celle d'une fonction de lien, devrait être guidé par des considérations théoriques. Sur la distribution, c'est bien une hypothèse sur le mécanisme même de génération des données qui est posée, et cette réflexion ne peut pas être indépendante de la théorie psychologique sous-jacente. Les auteurs travaillant sur les temps de réponse l'ont bien compris, et leurs conceptions du travail cognitif latent a mené à des propositions argumentées en termes de distribution. La démarche ne relève plus alors d'une « statistique appliquée à la psychologie », mais bien d'une *psychologie statistique*, que nous voyons peu à peu émerger dans la littérature.

Références

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6, 251-262.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-4, http://cran.r-project.org/package=lme4.

Fagot, D., Chicherio, C., & de Ribaupierre, A. (2013). Différences individuelles dans la capacité en mémoire de travail et variabilité intra-individuelle dans les temps de réponse Effets de l'âge et de la complexité de la tâche. In M. Carlier & P.-Y. Gilles (Eds.) Vive(nt) les différences. Psychologie différentielle fondamentale et applications (pp. 103-108). Aix en Provence, France : Presses Universitaires de Provence.

Gruen, B. & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis, 51, 5247-5252.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, 69, 382-386.

Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Kuiper, R.M. & Hoijtink, H. (2010). Comparisons of Means Using Exploratory and Confirmatory Approaches. *Psychological Methods*, 15, 69-86.

Matzke, D. & Wagenmakers, E.J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis . *Psychonomic Bulletin & Review*, 16, 798-817 .

Mosteller, F. (2010). Learning theory. In Finberg, S.E., Hoaglin, D.C. & Tanur, J.M. (Eds.), The Pleasure of Statistics : The Autobiography of Frederick Mosteller, New York : Springer.

Nelder, J.A. & Wedderburn, R.A. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A, 135, 370–384.

Noël, Y. (2013). *Psychologie statistique avec R*. Coll. Pratique R, Paris : Springer.

Noël, Y. (2013). R2STATS: A GTK GUI for fitting and comparing GLM and GLMM in R, R package version 0.68-34, http://cran.r-project.org/package=R2STATS.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. Journal of the Royal Statistical Society, Series B, 51, 47–60 .

Raftery, A.E. (1995). Bayesian model selection in social research (with Discussion). *Sociological Methodology*, 25, 111-196.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers* , 33, 457-469 .

Solomon, R.L., Kamin, L.J. & Wynne, L.C. (1953). Traumatic avoidance learning: the outcomes of several extinction procedures with dogs. *Journal of Abnormal and Social Psychology*, 48, 291-302.

Wang, P., Puterman, M., Cockburn, I. & Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*, 52, 381-400.