Dans le cas gaussien, en revanche, les paramètres du modèle peuvent être déduits directement par le biais de formules fermées.

La présence de formules fermées est synonyme de gain en complexité, et donc de gain en temps de calculs pour l'estimation des modèles. Aussi, Brouste, Dutang et Rohmer (2020) se sont intéressés à l'identification de cas où des formules fermées peuvent être extraites pour l'estimation des paramètres, quelle que soit la loi exponentielle retenue et quelle que soit la fonction de lien utilisée. Ces cas apparaissent lorsque les variables explicatives x_i considérées sont catégorielles ou qualitatives (par exemple : le type de véhicule ou l'usage du véhicule en assurance auto). Ceci correspond à une situation relativement courante dans un contexte d'assurance non-vie où la tarification des contrats est établie sur une grille, et où les variables continues comme l'âge ou le nombre de kilomètres parcourus par un véhicule sont préalablement discrétisées.

Sous réserve de respecter une condition d'identifiabilité des paramètres du modèle, ces travaux permettent alors d'exhiber des estimateurs explicites pour des modèles GLM présentant une ou deux variables catégorielles. Prenons par exemple le cas d'une seule variable explicative, le prédicteur linéaire du modèle peut alors se réécrire pour $x_i = (x_i^{(1)}, x_i^{(2)})$ *via*

$$g(E(Y_i)) = \theta_{(1)} + \sum_{j=1}^{d_2} x_i^{(2),j}\theta_{(2),j},$$

avec $\theta_{(1)} = 1$ l'intercept du modèle et $x_i^{(2),j}$, la j-ième modalité parmi les d_2 modalités prises par la variable $x_i^{(2)}$. En supposant que $< R, \theta > = 0$ et que $\sum_{k=1}^{d_2} r_k - r_0 \neq 0$, avec $R=(r_1, r_2, \ldots r_{d_2})$ un vecteur à valeurs réelles de taille d_2, l'estimateur du maximum de vraisemblance est cohérent, asymptotiquement normal et s'obtient alors de manière unique par

$$\hat{\theta}_{n,(1)} = \frac{\sum_{k=1}^{d_2} r_k g(\overline{Y_n}^{(k)})}{\sum_{k=1}^{d_2} r_k - r_0},$$

$$\hat{\theta}_{n,(2),j} = g(\overline{Y_n}^{(j)}) - \frac{\sum_{k=1}^{d_2} r_k g(\overline{Y_n}^{k})}{\sum_{k=1}^{d_2} r_k - r_0}, \; j=1,\ldots, d_2,$$

avec la moyenne empirique pour la variable réponse $\overline{Y_n}^{(j)} = \frac{1}{m_j}\sum_{i\in I} Y_i \times x_i^{(2),j}$ et la fréquence de la j-ième modalité $m_j = \sum_{i\in I} x_i^{(2),j}$. L'extension de ce type d'estimateurs en présence de plus de deux variables catégorielles est possible tant que l'on considère des termes d'interactions entre ces variables et une condition à somme nulle décrite dans Brouste, Dutang et Rohmer (2020).

L'estimation de modèles GLM intervient dans plusieurs algorithmes de *machine learning*. Sous la forme de modèles de régression logistiques ou poissoniens, ils peuvent par exemple intervenir dans des méthodes de

Quentin GUIBERT
Actuaire certifié IA, professeur associé à l'université Paris Dauphine et actuaire consultant chez Prim'Act

Christophe DUTANG
Actuaire certifié IA, enseignant -chercheur à Paris-Dauphine et responsable du M2 actuariat