Modélisation GLM

# Aspect théorique des modèles linéaires généralisés

Les « modèles linéaires généralisés » ont été introduits par Nelder et Wedderburn en 1972 puis exposés en 1989 par MC Cullagh et Nelder. Les GLM sont des extensions du modèle linéaire, l’objectif est de généraliser le modèle gaussien à un ensemble de lois plus large. Ces modèles présentent également l’avantage de pouvoir introduire des variables qualitatives dans le modèle. Ce qui, par exemple, correspondrait plus au taux d’absentéisme que l’on souhaite modéliser dans cette étude.

## Intérêt des modèles linéaires généralisés

Les modèles linéaires généralisés sont des modèles régulièrement utilisés en assurance que ce soit en assurance santé (remboursements soins, frais d’hospitalisation), en assurance auto (dommage matériel, vol, …), en assurance MRH (incendie, vol, dégâts des eaux, …).

Les GLM permettent de :

* Modéliser des réponses diverses
* Intégrer toute type d’information exogène susceptible d’influer sur la variable dépendante (réponse Y)
* Quantifier l’impact des facteurs de risque X (sens/intensité)
* Résidus hétéroscédastiques (la loi varie par profil)

Cependant, leur mise en place nécessite d’introduire deux hypothèses fondamentales :

* Les données que l’on cherche à expliquer sont indépendants entre elles
* Les variables explicatives X sont indépendantes deux à deux

Ces hypothèses sont très importantes notamment car l’un des intérêts des modèles linéaires généralisés en assurance est la tarification et donc le calcul d’une prime pure, calculée généralement de la façon suivante avec une approche fréquence-sinistre :

## Principe des GLM

### Composant d’un GLM :

Les différents composants d’un modèle linéaire généralisé sont :

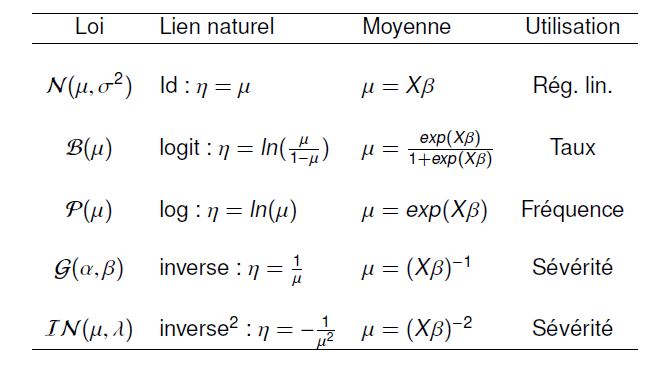
* **La loi de la réponse aléatoire Yi :** par hypothèse la distribution de cette loi appartient à la famille exponentielle

Rappel : La densité de probabilité d’une loi appartenant à la famille exponentielle s’écrit de la façon suivante :xdc v

Avec :

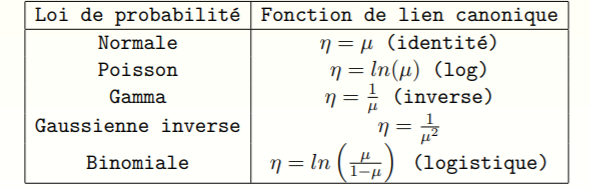
* Avec θ le paramètre naturel de la dispersion,
* φ le paramètre de dispersion, b une fonction définie sur deux fois dérivable et de dérivée première injective,
* c une fonction définie sur
* **Le prédicteur** noté ηi et défini tel que est linéaire et déterministe : il est calculé via les facteurs de risques explicatifs
* **La fonction de lien g :** elle est par définition monotone, dérivable et inversible. Elle est définie et utilisée en pratique de la façon suivante :

En pratique, il faut adapter la fonction de lien au domaine de définition de Y, voici un tableau qui présente les différentes fonction lien utilisées en pratique :



La fonction de lien canonique est la fonction lien qui associe la moyenne µ au paramètre canonique θ. Elle est telle que :

On trouvera ci-dessous la fonction de lien canonique associée à certaines lois usuelles :



A présent, présentons quelques modèles fréquemment utilisés dans le contexte assurantiel :

### Modèle gaussien :

Dans le cas d’un échantillon gaussien, les densités d’une famille de lois s’écrivent :

Et appartiennent à la famille exponentielle en posant :

On remarque avec la première égalité que la famille gaussienne se met sous forme canonique

### Modèle de Poisson :

Le modèle de Poisson est un modèle de comptage pour le définir, on considère n variables indépendantes Yi de loi de Poisson de paramètre µi = E(Yi). Les Yi sont par exemple les effectifs d’une table de contingence. Ces variables admettent pour densités :

Avec :

Et donc la fonction de lien canonique de ce modèle est la fonction logarithme népérien.

### Modèle Tweedie :

Les variables de réponses Yi admettent dans ce modèle la densité suivante :

Avec : =

Avec cette formalisation, et avec un facteur de dispersion positif.

### Validation du modèle :

Pour valider le modèle linéaire généralisé étudié, il faut passer par les étapes suivantes :

* Validation croisée : construction de 2 échantillons indépendants par tirage aléatoire : un échantillon d’apprentissage et un échantillon de validation
* Validation de la significativité globale du modèle
* Validation de la significativité individuelle des coefficients de la régression
* Etude des résidus
* Comparaison des résultats du modèle et des résultats observés