Modélisation GLM

# Aspect théorique des modèles linéaires généralisés

Les modèles linéaires généralisés ont été introduits par Nelder et Wedderburn en 1972, ils constituent la base de référence pour modéliser l’effet des variables de segmentation sur un tarif en assurance. Ces modèles sont adaptés à de nombreuses problématiques courante dans le domaine de la statistique et de l’actuariat.

## 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 :

Avec :

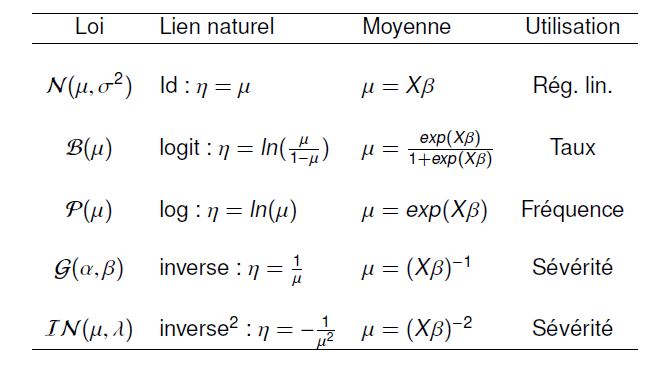
* Avec θ le paramètre naturel de la dispersion,
* φ le paramètre de dispersion, b une fonction définie sur deux fois dérivable et de dérivée première injective,
* c une fonction définie sur

De plus on a en particulier avec les lois de cette famille :

ainsi que

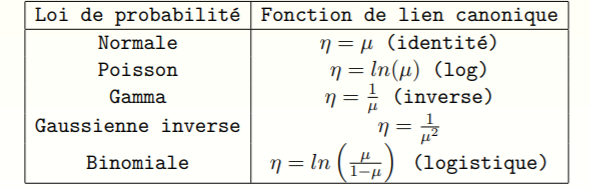
* **Le prédicteur** noté ηi et défini tel que est linéaire et déterministe : il est calculé via les facteurs de risques explicatifs
* **La fonction de lien g :** elle est par définition monotone, dérivable et inversible. Elle est définie et utilisée en pratique de la façon suivante :

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.

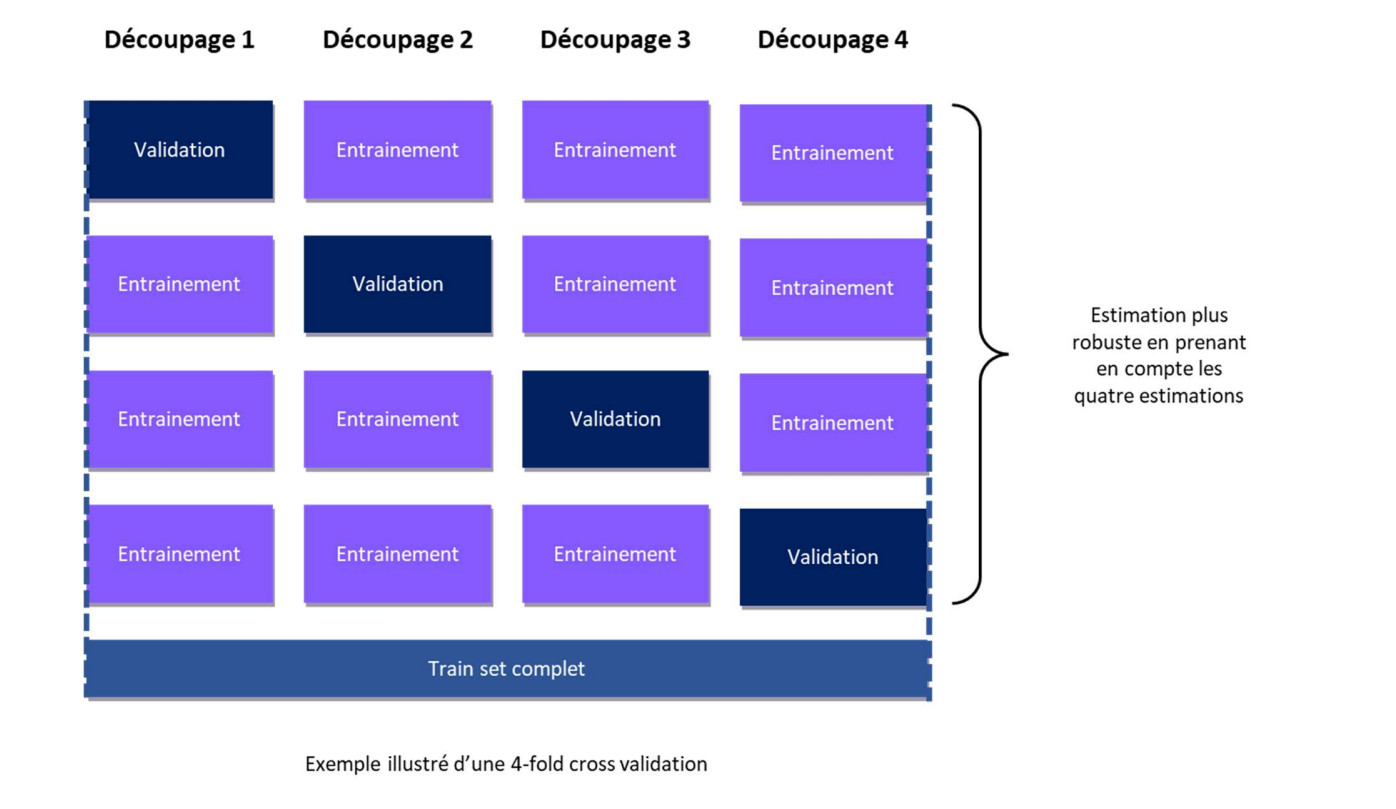
## Validation d’un modèle GLM :

Pour valider le modèle linéaire généralisé étudié, il existe les différentes méthodes suivantes :

### Validation croisée (k-fold)

Le découpage de l’échantillon en apprentissage et en validation induit une perte d’un certain volume de données pour calculer les estimateurs mais permet de calculer les erreurs moyenne de l’estimateur, cette segmentation de l’échantillon se fait via les étapes suivantes :

* On coupe aléatoirement l’échantillon de données en k sous-échantillons de taille égale
* On sélectionne un échantillon de validation et k-1 échantillon d’apprentissage
* On fait tourner k fois les étapes précédentes pour que chaque sous-échantillon ait été utilisé une fois en validation
* On calcule la moyenne sur les k estimation pour obtenir l’estimateur final
* On calcule l’erreur moyenne sur les échantillons de validation



### Validation de la significativité globale du modèle

### Validation de la significativité individuelle des coefficients de la régression

### Etude des résidus

Pour évaluer l’erreur du modèle, nous pouvons calculer les résidus du modèle de différentes manières, les méthodes que l’on utilisera dans cette étude sont les principales utilisées pour calculer des résidus, les résidus de Pearson et les résidus de déviances définis tels que :

* Les résidus de Pearson :

Le résidu de Pearson est défini comme :

* Les résidus de Déviance :

On peut considérer que chaque observation yi contribue à une quantité di à la déviance (), le résidu de déviance est défini comme :

La somme des carrés des résidus dans les deux cas est asymptotiquement une statistique du Khi-2 à n-p-1 degrés de liberté.

### Comparaison des résultats du modèle et des résultats observés