Data Science pour l'assurance non vie

Dafnis Krasniqi

SAMM - Statistique, Analyse et Modelisation Multidisciplinaire
Universite Paris 1 Pantheon-Sorbonne (EA 4543)

5 octobre 2022

Overview

Generalized Linear Models (GLM)

Les modèles linéaires généralisés ont été développés dans les années 1970 par deux statisticiens, John Nelder et Robert Wedderburn. Les GLM couvrent de nombreuses distributions paramétriques classiques.

Tous ces modèles ont trois choses en commun :

- Une distribution de même forme : Y admet pour densité $f(y_i, \theta_i, \phi)$
- Prédicteurs linéaires : Choix de la combinaison linéaire des prédicteurs $\eta(x) = x^T\beta$
- Fonction de lien : $E[Y]$ est lié X selon $g(E[Y]) = \eta(x)$

La famille des distributions exponentielles

Definition

Si y_i appartient au GLM alors la distribution de probabilité prend la forme :

$$f(y_i, \theta_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right), \quad y_i \in \mathbb{R} \tag{1}$$

où θ_i et ϕ sont des paramètres inconnus et a , b, c sont des fonctions déterministes connues et ils sont spécifiées en fonction du type de la famille exponentielle. θ_i est appelé paramètre naturel et ϕ est considéré comme un paramètres de nuisance.

L'espérance et la variance pour les GLM sont définies comme suit

$$\mathbb{E}(Y_i|X_i = x_i) = b^{'}(\theta_i)$$

$$\mathbb{Var}(Y_i|X_i = x_i) = b^{''}(\theta_i)\phi$$

Prédicteurs linéaires et Fonction de lien

Definition

Le prédicteur linéaire met en relation le paramètre η avec les prédicteurs X. Dans un modèle GLM, cela se définie comme suit :

$$\eta(x_i) = \theta_i = x_i^T \beta \tag{2}$$

Definition

Nous allons appeler cette fonction de lien g et elle se définie ainsi :

$$g(\mu_i) = \eta(x_i) \tag{3}$$

Fonction de lien

Distribution	Lien	Lien inverse
Normale	$\mu_i = \eta(x_i)$	$\mu_i = \eta(x_i)$
Binomial	$\text{Logit}(\mu_i) = \eta(x_i)$	$\mu_i = \frac{exp(\eta(x_i))}{1+exp(\eta(x_i))}$
Poisson	$\log(\mu_i) = \eta(x_i)$	$\mu_i = exp(\eta(x_i))$
Gamma	$\frac{1}{\mu_i} = \eta(x_i)$	$\mu_i = \frac{1}{\eta(x_i})$

(4)

Estimation des paramètres β

Pour estimer les β, nous allons utiliser le maximum de la vraisemblance. Avec l'hypothèse que les que $Y_1, ..Y_n$ des variables aléatoires indépendantes, la vraisemblance et la log-vraisemblance peuvent s'écrire :

$$L(\theta) = \prod_{i=1}^{N} f(y_i, \theta_i, \phi) = \prod_{i=1}^{N} exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]$$

$$l(\theta) = \frac{1}{a(\phi)}\left[\sum_{i=1}^{N} y_i\theta_i - \sum_{i=1}^{N} b(\theta_i)\right] + \sum_{i=1}^{N} c(y_i, \phi)$$

Avec $\theta = (\theta_1, ...\theta_n)$
Ainis au finale, le β finaux seront égales a,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}}\ l(\beta)$$

avec $\hat{\beta} = \hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$

Sélection de modèle : la déviance

Une fois que nous avons estimé nos différents paramètres β nous allons pouvoir calculer la déviance. La déviance va comparer log-vraisemblance avec les paramètres estimés et la log-vraisemblance avec les valeurs observées (ce que l'on appelle modèle saturé).

$$\mathcal{D} = 2\phi \left[\log\left(L(Y)\right) - \log\left(L(\mu)\right)\right]$$

avec $\log\left(L(Y)\right)$ le modèle saturé et $\log\left(L(\mu)\right)$ la log-vraisemblance estimée.
Plus la déviance converge vers 0 et plus la qualité de la régression est bonne.

$$GLM\ \ Normale \rightarrow \mathcal{D} = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

$$GLM\ \ Poisson \rightarrow \mathcal{D} = \sum_{i=1}^{n} \left(y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right)$$

$$GLM\ \ Gamma \rightarrow \mathcal{D} = \sum_{i=1}^{n} \left(-\log\left(\frac{y_i}{\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i} \right)$$

Sélection de modèle : AIC et BIC

Malheureusement, la déviance souffre d'un biais car il augmente automatiquement lorsque des variables sont ajoutées au modèle. Pour cette raison, il doit être associé à d'autres critères tels que l'AIC ou le BIC.
L'AIC et le BIC se définissent comme suit :

$$AIC = -2 * \log(L(\beta, y_i)) + 2 * k$$

$$BIC = -2 * \log(L(\beta, y_i)) + \log(n) * k$$

avec $\log(L(\beta, y_i))$ la log-vraisemblance de la régression, k le nombre de paramètres et n la taille de l'échantillon.

Sélection de modèle : Différence entre les prédictions et les observations.

Un GLM a une fonction de liaison canonique (Normale $\Rightarrow$ identité, Poisson $\Rightarrow$ log, Gamma $\Rightarrow$ 1/x, Binomiale $\Rightarrow$ logit) avec un terme d'interception a la propriété dite **d'équilibre**. En négligeant les petites déviations par l'opimiseur utilisé pour l'ajustement, les résultats s'accomplissent sur l'échantillon d'entraînement :

$$\sum_{i \in training} t_i y_i = \sum_{i \in training} t_i \hat{\mu}_i$$

En additionnant les prédictions $\hat{\mu}_i$, on obtient exactement les observations y_i.

Modèle GLM de poisson

Definition

Le GLM de Poisson est le modèle standard pour modéliser les données de comptage (situations où évènements rares). La distribution de probabilité du GLM Poisson se définie ainsi :

$$f(Y_i|\lambda_i) = \frac{exp\left(-t_i\lambda_i\right)\left(t_i\lambda_i\right)^{y_i}}{y_i!} = exp\left(y_i \log(t_i\lambda_i) - t_i\lambda_i - \log(y_i!)\right)$$

avec $y_i \in \mathbb{N}$.

Nous avons $\theta = \log(\lambda)$, $\phi = 1$, $b(\theta) = exp(\theta) = \lambda$ et $C(y, \phi) = -\log(y!)$.

La fonction de liaison est log(), et elle est définie comme suit :

$$\log\left(\frac{\lambda_i}{t_i}\right) = x_i^T\beta \Longleftrightarrow \lambda_i = t_i exp(x_i^T\beta)$$

Modèle GLM de poisson

Definition

Si Y suit une distribution de Poisson la moyenne et la variance sont définies comme suit :

$$\mathbb{E}\left(Y_i\right) = Var\left(Y_i\right) = \lambda_i = t_i exp(x_i^T \beta)$$

avec x_i^T est un vecteur de covariables et $\beta = \beta_0, \beta_1, ..., \beta_p$ est un vecteur de paramètres.

Avec hypothèse que les que $Y_1, ..Y_n$ sont indépendant, la log-vraisemblance s'écrire :

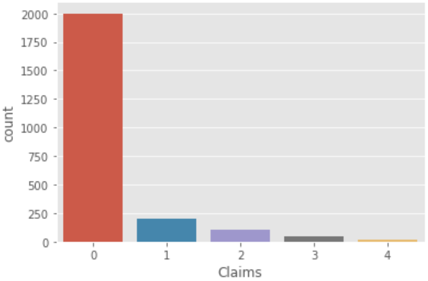
$$l(\beta, y_i) = \log(L(\beta, y_i)) = \sum_{i=1}^{n} -exp(x_i^T \beta)t_i + y_i(x_i^T \beta + \log(t_i)) - \log(y_i)!$$

En fixant les dérivées à zéro, la condition de maximum de vraisemblance de premier ordre est la suivante :

$$\sum_{i}^{n} x_i \left(y_i - exp(x_i^T \beta)t_i\right) = 0$$

Modèle GLM de poisson

Voici un exemple de données pour les données de comptage



Generalized Additive Model (GAM)

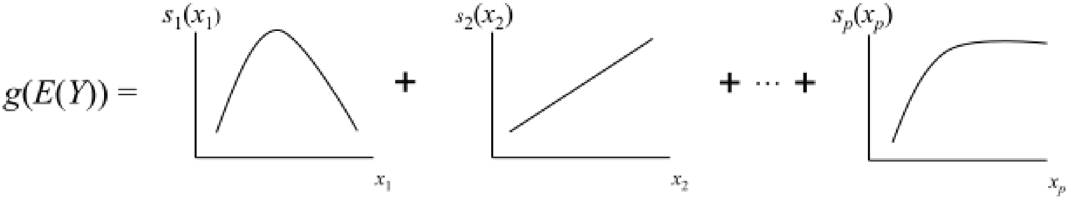


Figure – GAN

Generalized Additive Model (GAM)

Dans cette partie, nous travaillerons uniquement avec des splines cubiques naturelles.

Definition

Les splines cubiques s'écrivent ainsi :
Considérons P nœuds $x_1, x_2, ..., x_P$ tels que $0 < x_1 < x_2 < x_3 < ... < x_p$,

$$s(x) = \beta_0 + \beta_1 x^3 + \beta_2 (x - x_1)_+^3 + \beta_3 (x - x_2)_+^3 + \beta_4 (x - x_3)_+^3 ... + \beta_{P+1} (x - x_P)_+^3$$

On impose que $s''(x_1) = 0$ et $s''(x_P) = 0$

Generalized Additive Model (GAM)

La courbe a deux objectifs :

- se rapprocher le plus possible des points de notre jeu de données. Pour atteindre cet objectif, nous utiliserons $\sum_{j=1}^{n} \{y_i - s(x_j)\}^2$.
- donner la priorité à la variation générale par rapport à la variation locale. Pour atteindre cet objectif, nous utiliserons $\lambda \int_{x_1}^{x_n} s''(x)^2 \, dx$.

Une façon simple de trouver λ est d'utiliser le score GCV (generalized cross validation).

$$V_g = \frac{n \sum_{i=1}^{n} (y_i - s_i)}{[tr(I - A)]^2}$$

avec $A = X(X^T X + \lambda S)^{-1} X^T$

Generalized Additive Model (GAM)

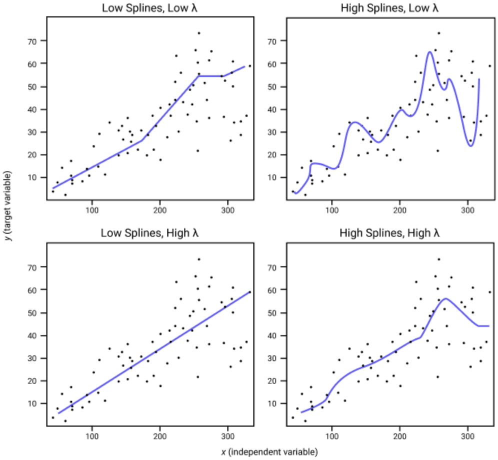


Figure – GAN

Generalized Additive Model (GAM)

Au lieu d'avoir une seule courbe pour la variable de l'âge du conducteur, nous aurons plusieurs courbes en fonction de ce que nous choisissons pour P.

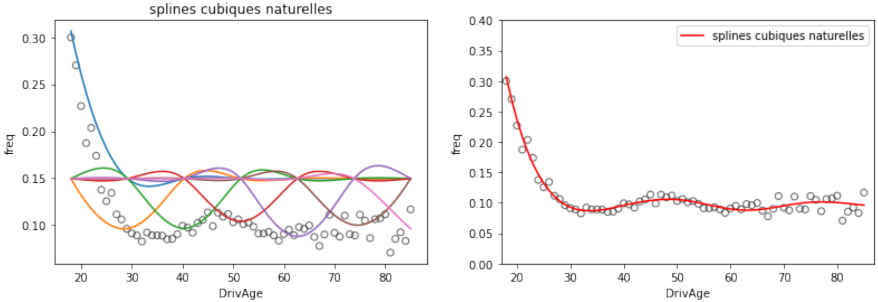


Figure – GAN

Generalized Additive Model (GAM)

Remarque

Ces méthodes ont l'avantage d'être plus puissantes car elles s'adaptent mieux aux données mais ont l'inconvénient du sur-apprentissage (pas de généralisation).

Conclusion

Ces modèles ont l'avantage d'être très simples à implémenter et à interpréter. Grâce aux coefficients $\hat{\beta}$, nous pouvons directement savoir quelle variable a eu le plus fort impact sur les prédictions.

The End