Dauphine | PSL
UNIVERSITÉ PARIS

SUJET A5 - TARIFICATION EN ASSURANCE IARD AVEC LES GLM ET LES GAM

MÉMOIRE DE MASTER 1 MATHÉMATIQUES APPLIQUÉES DE L'UNIVERSITÉ PARIS DAUPHINE

Adnane EL KASMI | Samuel TEBOUL | Antonin AUBRY

Soutenance M1, 11 mai 2021

Plan

INTRODUCTION

Plan

2-1- Présentation des données

⋆ La base police **freMPL3** possède 30 595 observations et 23 variables.

⋆ la base sinistre **freMPL4** possède 36 295 observations et 23 variables.

Intitulé de la variable	Type de variable	Valeurs de la variable	Descriptif
VehMaxSpeed	Qualitative (polytomique)	"1-130 km/h" à "200-220 km/h "	Vitesse maximum du vehicule
SocioCateg	Qualitative (polytomique)	"CSP1" à "CSP99"	Catégorie sociale
VehUsage	Qualitative (polytomique)	"Private", "Professional" ...	Usage du vehicule
Deductype	Qualitative (polytomique)	"Majorized", "Normal Partially" ...	Type de franchise
VehBody	Qualitative (polytomique)	"bus", "cabriolet" ...	Carosserie du vehicule
VehPrice	Qualitative (polytomique)	"A" jusqu'à "Z", "Z1"	Prix du vehicule
VehEngine	Qualitative (polytomique)	"carburation direct", "electric" ...	Type de moteur du vehicule
VehEnergy	Qualitative (polytomique)	"diesel", "electric" ...	Type d'energie du vehicule
VehClass	Qualitative (polytomique)	"0", "A", "B", "H", "M1" et "M2"	Classe du vehicule
Garage	Qualitative (polytomique)	"Collective garage", "None"	Type de garage
VehAge	Qualitative (polytomique)	"0" à "10+"	Âge du vehicule

Intitulé de la variable	Type de variable	Valeurs de la variable	Descriptif
Exposure	Quantitative (continue)	[0,1]	Exposition au risque dans l'année
LicAge	Quantitative (discrète)	Entier	Âge du permis de conduire en mois
BonusMalus	Quantitative (discrète)	Entier de 50 à 250	Bonus si > 100 et Malus si < 100
DrivAge	Quantitative (discrète)	Entier de 18 à 97	Âge du conducteur
ClaimAmount	Quantitative (continue)	Réel	Le montant déclaré
RiskVar	Quantitative (discrète)	Entier de 1 à 20	Variable du risque

Intitulé de la variable	Type de variable	Valeurs de la variable	Descriptif
ClaimInd	Qualitative (dichotomique)	{0,1}	Indicateur de la réclamation
Gender	Qualitative (dichotomique)	{Male, Female}	Le genre
MariStat	Qualitative (dichotomique)	{Alone, Other}	Statut marital
HasKmLimit	Qualitative (dichotomique)	{0,1}	Limite de kilométrage

Intitulé de la variable	Type de variable	Valeurs de la variable	Descriptif
RecordBeg	Temporelles	JJ-MM-AAAA	Date début du contrat
RecordEnd	Temporelles	JJ-MM-AAAA	Date fin du contrat

TABLE: Les variables de nos bases de données freMPL3 et freMPL4.

2-2- Analyse statistique

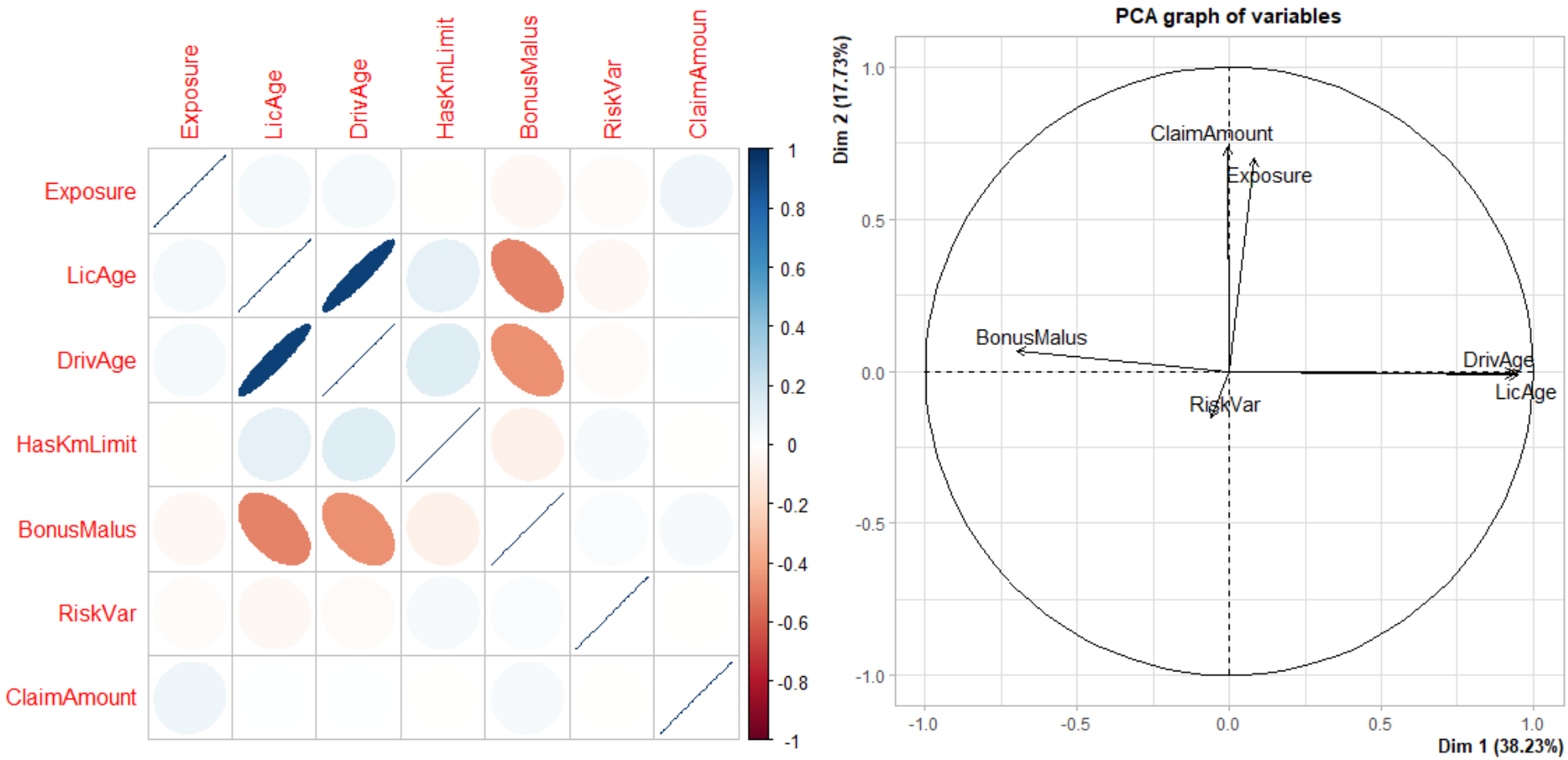

Figure: Corrélogramme et graphe de l'ACP (variables quantitatives).

Plan

3-1 Modélisation de la fréquence des sinistres (ClaimInd)

Régression logistique $y_i =$ " $ClaimInd_i$" (GLM) :

- $y_i \in \{0, 1\}$ donc $y_i|x_i \sim Bernoulli(p_i)$ avec $p_i = P(y_i = 1|x_i) = E[y_i|x_i]$.
- Fonction lien canonique **logit**: $g(E[y_i|x_i]) = logit(p_i) = log(\frac{p_i}{1-p_i}) = x_i^T \beta$.
- Sélection du modèle optimal avec l'algorithme "AIC" :

```
GLM.ClaimInd <- glm(ClaimInd ~ LicAge3 + BonusMalus + VehAge10 + VehAge6 +
                      VehAge8 + SocioCategCSP6 + VehUsageProfessional +
                      VehUsageProfessional_run + DeducTypePartially_refunded +
                      DeducTypeRefunded + VehBodystation_wagon + VehPriceO +
                      VehPriceP + VehPriceQ + VehMaxSpeed140_150_km_h +
                      VehMaxSpeed150_160_km_h + VehMaxSpeed160_170_km_h +
                      VehMaxSpeed170_180_km_h + VehMaxSpeed180_190_km_h +
                      VehMaxSpeed190_200_km_h + VehMaxSpeed200_220_km_h +
                      VehMaxSpeed220_km_h + VehClassA + GaragePrivate_garage,
                      family = binomial(link = 'logit'),
    data = cbind.data.frame(ClaimInd = freMPL34.app$ClaimInd, freMPL34.app.reg))
```

- On trouve que l'erreur de prédiction sur l'échantillon test est estimée à 2.3%.
- D'après la matrice de confustion on trouve un Accuracy de 53.5%.

3-2 Modélisation du coût des sinistres (ClaimAmount)

Régression Gamma $y_i =$ "$ClaimAmount_i$" (GLM) :

- $y_i \in \mathbb{R}^+$ donc $y_i|x_i \sim Gamma(\alpha, \beta)$.
- Fonction lien **log**: $g(E[y_i|x_i]) = log(E[y_i|x_i]) = x_i^T \beta$.
- Sélection du modèle optimal avec l'algorithme "AIC" :

```
GLM.claimAmount <- glm(ClaimAmount ~ LicAge + BonusMalus + HasKmLimit +
                       RiskVar + VehAge3 + VehAge4 + VehAge6 + VehAge8 +
                       DeducTypePartially_refunded + VehPriceE +
                       VehPriceJ + VehPriceM + VehPriceN + VehPriceQ +
                       VehEnergyregular + VehMaxSpeed140_150_km_h +
                       VehMaxSpeed160_170_km_h + VehMaxSpeed170_180_km_h +
                       VehMaxSpeed180_190_km_h + VehMaxSpeed190_200_km_h +
                       VehMaxSpeed200_220_km_h + VehClassA + VehClassM1 +
                       GarageNone + VehUsagePrivate_trip_to_office
                       + VehBodycabriolet,
                       family = Gamma(link = "log"),
    data = cbind.data.frame(ClaimAmount = freMPL34.app.sinistre$ClaimAmount,
                       freMPL34.app.sinistre.reg),
                       weights = VehBodymicrovan + VehBodycoupe +
                       SocioCategCSP2 + SocioCategCSP4 +
                       SocioCategCSP5 + SocioCategCSP7 +
                       SocioCategCSP9)
```

Plan

4-1 Modélisation de la fréquence des sinistres (ClaimInd)

Régression logistique $y_i =$ " $ClaimInd_i$" (GAM) :

- $y_i \in \{0, 1\}$ donc $y_i|x_i \sim Bernoulli(p_i)$ avec $p_i = P(y_i = 1|x_i) = E[y_i|x_i]$.
- Fonction lien canonique **logit**: $g(E[y_i|x_i]) = logit(p_i) = log(\frac{p_i}{1-p_i}) = x_i^T \beta$.
- Sélection du modèle optimal avec l'algorithme "AIC", "REML" et "GAM.CHECK":

```
GAM.claimInd <- gam(formula = ClaimInd ~ s(LicAge, k=34) + BonusMalus + VehAge10 +
                      VehAge6 + VehAge8 + SocioCategCSP6 + VehUsageProfessional +
                      VehUsageProfessional_run + DeducTypePartially_refunded +
                      DeducTypeRefunded + VehBodystation_wagon + VehPrice0 +
                      VehPriceP + VehPriceQ + VehMaxSpeed140_150_km_h +
                      VehMaxSpeed150_160_km_h + VehMaxSpeed160_170_km_h +
                      VehMaxSpeed170_180_km_h + VehMaxSpeed190_200_km_h +
                      VehMaxSpeed200_220_km_h + VehMaxSpeed220_km_h + VehClassA +
                      GaragePrivate_garage,
                      family = binomial("logit"),
    data = cbind.data.frame(ClaimInd = freMPL34.app$ClaimInd, freMPL34.app.reg),
                      method="REML")
```

- L'erreur de prédiction sur l'échantillon test est estimée à 2.3%;
- D'après la matrice de confustion on trouve un Accuracy de 55.8%;

Visualiser l'effet de LicAge sur l'échelle du logit:

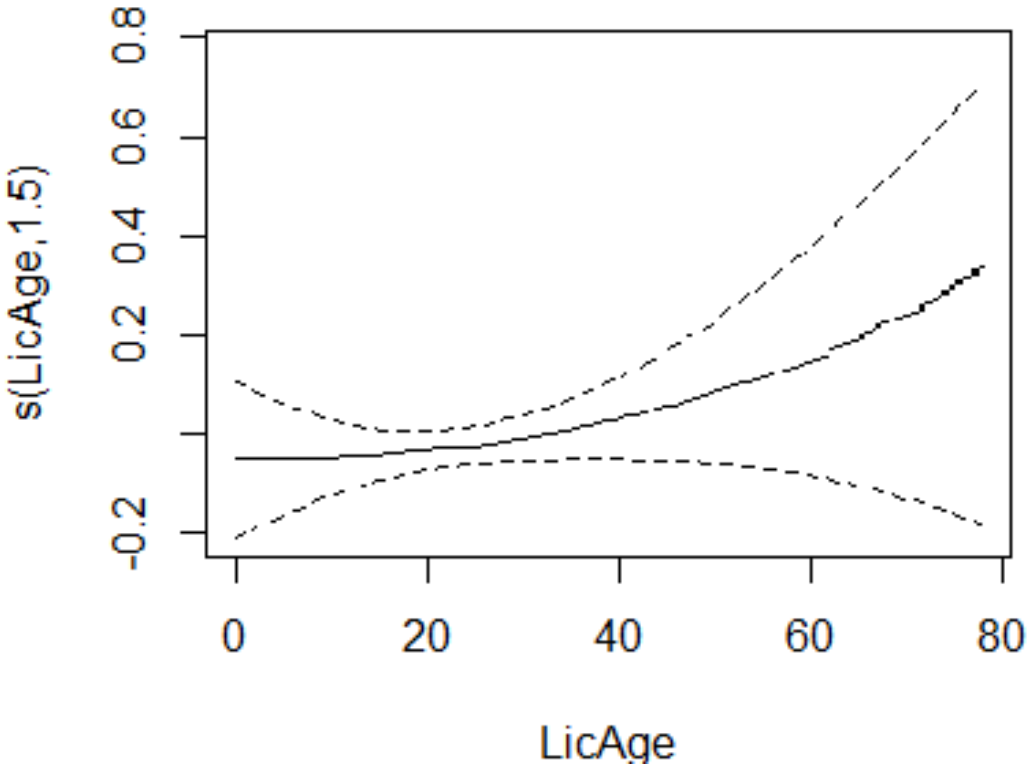

FIGURE: Sortie R du plot(GAM.ClaimInd).

4-2 Modélisation du coût des sinistres (ClaimAmount)

Régression Gamma $y_i =$ " $ClaimAmount_i$" (GAM) :

- $y_i \in \mathbb{R}^+$ donc $y_i|x_i \sim Gamma(\alpha, \beta)$.
- Fonction lien **log**: $g(E[y_i|x_i]) = log(E[y_i|x_i]) = x_i^T \beta$.
- Sélection du modèle optimal avec l'algorithme "AIC", "REML" et "GAM.CHECK":

```
GAM.claimAmount <- gam(ClaimAmount ~ s(LicAge, k=62)+ s(BonusMalus) + HasKmLimit +
                         s(RiskVar) + VehAge3 + VehAge4 + VehAge6 + VehAge8 +
                         DeducTypePartially_refunded +  VehPriceE + VehPriceJ +
                         VehPriceM + VehPriceN + VehPriceQ + VehEnergyregular +
                         VehMaxSpeed140_150_km_h + VehMaxSpeed160_170_km_h +
                         VehMaxSpeed170_180_km_h + VehMaxSpeed180_190_km_h +
                         VehMaxSpeed190_200_km_h + VehMaxSpeed200_220_km_h +
                         VehClassA + VehClassM1 + GarageNone +
                         VehUsagePrivate_trip_to_office + VehBodycabriolet,
                         family = Gamma(link = "log"),
       data = cbind.data.frame(ClaimAmount = freMPL34.app.sinistre$ClaimAmount,
                         freMPL34.app.sinistre.reg),
                         weights = VehBodymicrovan + VehBodycoupe +
                         SocioCategCSP2 + SocioCategCSP4 +SocioCategCSP5 +
                         SocioCategCSP7 + SocioCategCSP9, method = "REML")
```

Verification du modèle additif généralisé $y_i =$ " $ClaimAmount_i$" :

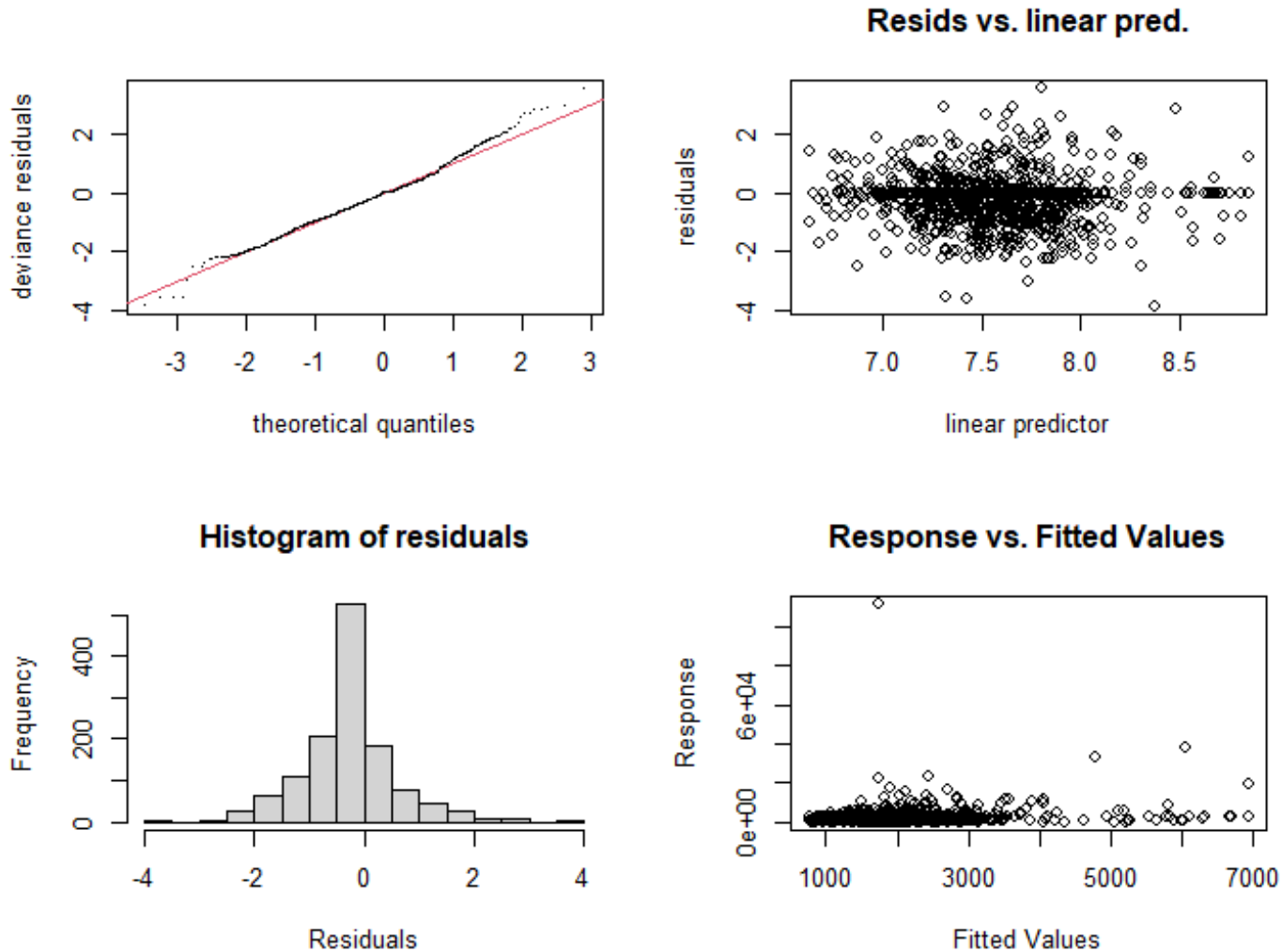

FIGURE: Sortie R gam.check(GAM.claimAmount).

Plan

5-1 Calcul de la prime pure

En assurance, la prime pure correspond au montant moyen des sinistres auquel devra faire face l'assureur.

- $Sinistre.Individu_i = I_i * B_i$.
- Fréquence et Sévérité supposées indépedantes.
- Le calcul de la prime pure est donc donné par :

$$Prime_{pure.i} = P(ClaimInd_i = 1).Predict.ClaimAmount_i.$$

La moyenne de primes pures est de **49.50** pour les GLMs et **49.88** pour les GAMs.

$\Rightarrow$ Mais problème de rentabilité.

5-2 Simulation de la prime avec chargement

Une prime avec chargement est de la forme :

$$Prime_{AvecChargement} = (1 + \eta) Prime_{pure}.$$

Objectif: être rentable dans 99% des cas.

La simulation:

- Tirer aléatoirement des ClaimAmount puis les sommer.
- Le faire 1000 fois.
- Sommer les primes avec différents taux de chargement.
- Comparer somme des primes et sommes de sinistres.

On obtient un chargement $\eta = 0.151$ pour les GLMs et $\eta = 0.142$ pour les GAMs.

5-3 Comparaison des primes avec chargement selon le modèle utilisé

Différents moyens de comparaisons:

- Répartition des primes selon le modèle:

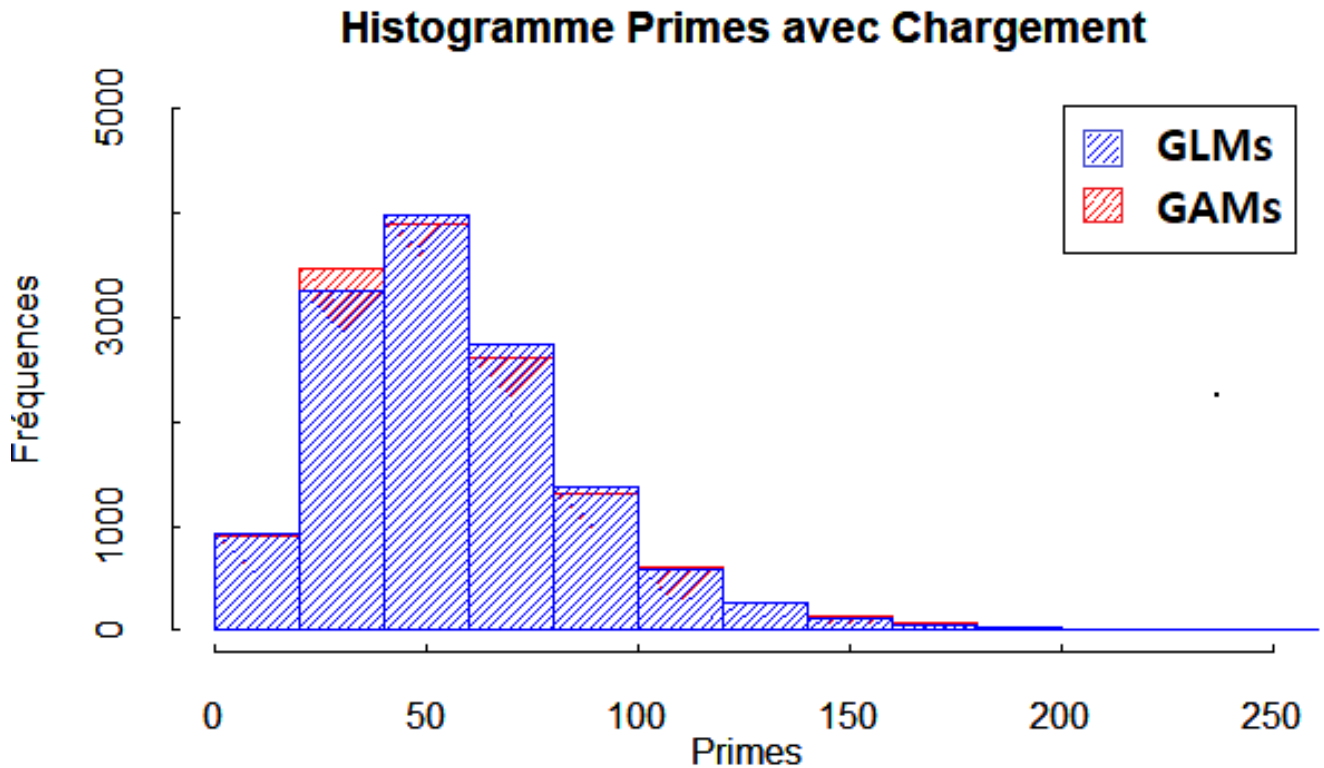

FIGURE: Histogrammes des primes avec chargement, selon la méthode GLM ou GAM.

- Ecart entre les primes: $Ecart_i = |Prime_{GAM} - Prime_{GLM}|$.

Plan

Limites et problèmes rencontrés

Au cours de notre travail, nous avons rencontré différentes difficultées :

- Réalisation de l'ACP et de l'AFC $\surd$.
- Non convergence du GLM pour ClaimAmount (problème avec la fonction de lien) $\surd$.
- Calibrage des modèles additifs généralisés (GAMs) $\surd$.
- Calcul de la prime $\surd$.
- Simulation du chargement technique $\surd$.

Plan

CONCLUSION