Statistique asymptotique
TD/TP 2 : Régression de Poisson : inflation de zéro

Avant de commencer,

- récupérer sur les données "claims_small" et "claims". Vous trouverez une description ainsi qu'une proposition d'analyse ici

Les données contiennent

- `IDpol` : policy number (unique identifier)
- `ClaimNb` : number of claims on the given policy
- `Exposure` : total exposure in yearly units
- `Area` : area code (categorical, ordinal)
- `VehPower` : power of the car (categorical, ordinal)
- `VehAge` : age of the car in years
- `DrivAge` : age of the (most common) driver in years
- `BonusMalus` : bonus-malus level between 50 and 230 (with reference level 100) ; 9. VehBrand : car brand (categorical, nominal)
- `VehBrand` : brand of the car (categorical)
- `VehGas` : diesel or regular fuel car (binary)
- `Density` : density of inhabitants per km2 in the city of the living place of the driver
- `Region` : regions in France (prior to 2016), these are illustrated in Figure 1 (categorical).

On veut expliquer et prédire la variable `ClaimNb`.

Exercice 1

1. Charger les données et vérifier le nombre d'observations et de variables ainsi que leur type. Faire les changements de type nécessaires.
2. Créer la variable `Frequency` définie comme le quotient des variables `ClaimNb` et `Exposure`
3. Créer un jeu de données plus petit (contenant 100 000 observations).
4. Sur ce jeu de données, calculer par variable le nombre de valeurs manquantes

Exercice 2

Faire les graphiques suivants à partir de `claims_small`

1. distribution de la variable `ClaimNb`, vous pouvez utiliser l'option `scale_y_log10()` de `ggplot`
2. distribution de la variable `Exposure`
3. distribution de la variable `Frequency`.
4. Calculer la proportion d'observations avec `ClaimsNb=0` et la moyenne de `ClaimsNb=0`

5. Etudier graphiquement le lien entre `Frequency` et les variables explicatives.

Exercice 3

1. Créez un jeu de données d'apprentissage `claims_small_train` contenant 80 000 observations et un jeu de données de test `claims_small_train` contenant 20 000 observations.
2. Suite aux observations précédentes, proposez un premier modèle `model_full` entrainé sur `claims_small_train` en prenant en compte toutes les variables explicatives potentielles et la variable `Exposure` comme `offset`.
3. Faites un choix de modèle par AIC backward, on appellera le modèle obtenu `model_full_aic`
4. Entraînez également le modèle `model_null` avec l'intercept seul et la variable `Exposure` comme `offset`
5. Calculer les prédictions obtenues avec ces modèles sur les individus du train et du test, puis calculer la MAPE et le log-MSE (il faudra coder une fonction qui renvoie ces mesures sur le jeu de données de train et de test).

$$\log - \text{MSE} = \frac{1}{n} \sum_i (\log(\hat{Y}_i^P + 1) - \log(Y_i + 1))^2$$

Exercice 4

Entrainez suivant les étapes de l'exercice précédent (modèle full, modèle full avec sélection aic et modèle null) un modèle de Hurdle, voir https ://en.wikipedia.org/wiki/Hurdle_model. Attention à bien coder les prédictions, il faudra faire une fonction à part.

Exercice 5

Essayer d'améliorer les modèles précédent via

- du features engineering
- des pénalisations
- ...