École d'actuariat
Université Laval

Travail pratique 1

Guillaume Michel
Nathanaël Pelchat
Mikael Robertson
Olivier Turcotte

Automne 2018

1 Sommaire exécutif

Table des matières

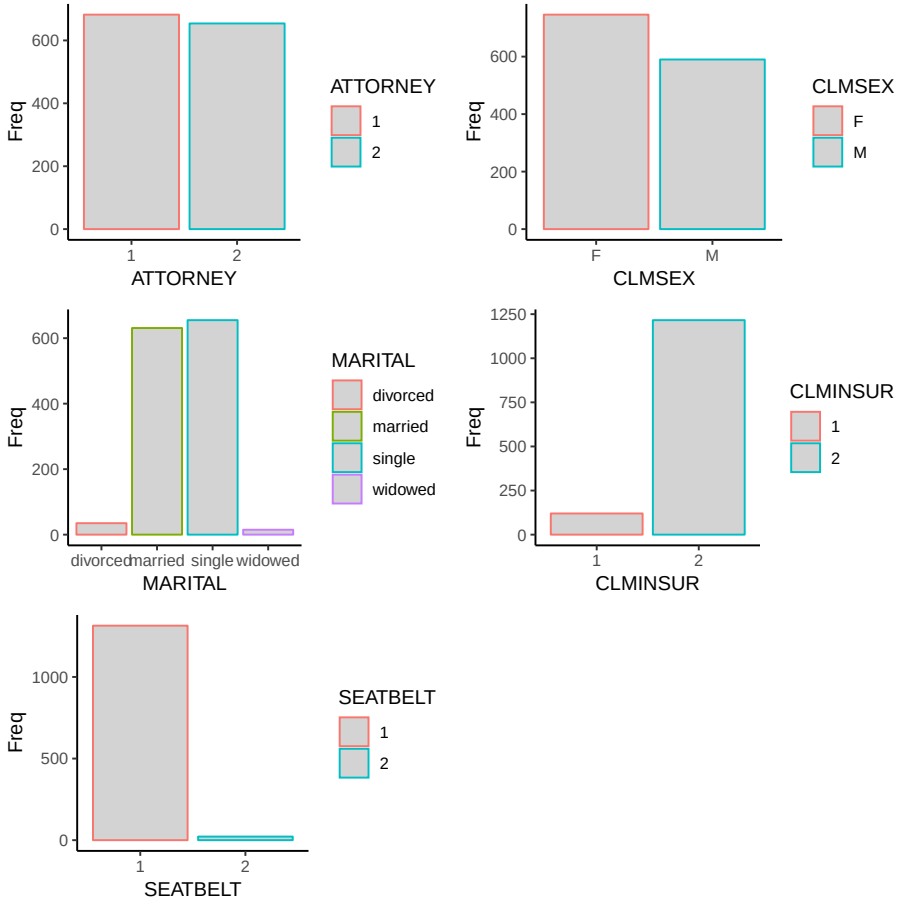
2 Analyse des données

Voici les variables disponibles afin d'effectuer un modèle prédictif de la perte économique :

Tableau 1 – Description des variables

Variables	**Type**	**Description**
CASENUM	Valeur entière	Numéro d'identification de la réclamation
ATTORNEY	Variable indicatrice	Indique si le réclamant est représenté par un avocat
CLMSEX	Variable indicatrice	Indique le sexe du réclamant
MARITAL	Variable polytomique	Indique le statut marital du réclamant
CLMINSUR	Variable polytomique	Indique si le réclamant est assuré
SEATBELT	Variable polytomique	Indique si le réclamant portait une ceinture de sécurité
CLMAGE	Valeur entière	Âge du réclamant
LOSS	Valeur continue	Perte économique totale du réclamant en milliers de dollars

Ces variables sont en majorité qualitative. Une analyse de fréquences de celles-ci permet d'avoir un meilleur ressenti quant à leurs interaction avec la variable exogène Loss :




```
##      CLMAGE          LOSS
##  Min.   : 0.0   Min.   :   0.005
##  1st Qu.:21.0   1st Qu.:   0.640
##  Median :33.0   Median :   2.331
##  Mean   :32.6   Mean   :   5.965
##  3rd Qu.:41.0   3rd Qu.:   3.998
##  Max.   :95.0   Max.   :1067.697
```

3 Modèle proposé

3.1 Équation

Le modèle choisit est donné par l'équation suivante

$$\begin{aligned}\ln Y =&\beta_0 + \beta_1 x_{i,CLMAGE} + \beta_2 x_{i,ATTORNEY} + \beta_{3,1} x_{i,MARITAL,2} + \\ &\beta_{3,2} x_{i,MARITAL,3} + \beta_{3,3} x_{i,MARITAL,4} + \beta_4 x_{i,SEATBELT} + \\ &\beta_5 x_{i,CLMAGE} * x_{i,ATTORNEY}\end{aligned}$$

3.2 Traitement des variables qualitatives

Les variables qualitatives du modèle, soit *ATTORNEY, SEATBELT et MARITAL* ont chacune été converti en *factor* car c'est le type de données usuelle de R afin d'effectuer des régression linéaire comportant des variables qualitatives.

3.3 Interactions

3.4 Interprétation

3.5 Statistiques

Voici les intervalles de confiances à 95% pour chacun des paramètres du modèle :

Tableau 2 – Intervalles de confiances des paramètres du modèle

	2.5%	97.5%
β_0	0.28583388	1.262939359
β_1	0.01334127	0.026414523
β_2	-1.32630903	-0.708680308
$\beta_{3,1}$	-0.51501270	0.340256198
$\beta_{3,2}$	-0.72345837	0.148361570
$\beta_{3,3}$	-1.69700396	-0.155725880
β_4	0.46855029	1.528929789
β_5	-0.02025211	-0.003201663

De plus, voici le R_a^2 : 0.2754. Ainsi, une grande variabilité du modèle n'est pas expliqué par le modèle. Ceci est en parti dû au grand nombre de variables qualitives dans le modèle qui n'ont pas assez de valeurs possibles afin de réfléter l'étendue des valeurs possible de la perte économique.

Voici la table anova du modèle :

```
## Analysis of Variance Table
##
## Response: log(LOSS)
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## CLMAGE             1   78.57   78.57  49.9248 2.572e-12 ***
## ATTORNEY           1  674.73  674.73 428.7572 < 2.2e-16 ***
## MARITAL            3   23.54    7.85   4.9859 0.0019223 **
## SEATBELT           1   21.20   21.20  13.4721 0.0002517 ***
## CLMAGE:ATTORNEY    1   11.46   11.46   7.2819 0.0070538 **
## Residuals       1328 2089.85    1.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Analyse des résidus

5 Prévisions

6 Recommendations

Annexes

A Erreurs de données

La base de données originelles utilisé dans la création du modèle a dû subir quelques modifications afin d'être utilisable. Voici les quelques erreurs répertoriés ainsi que les techniques utilisés pour les rectifier :

1. Fautes d'ortographe
   (a) MARITAL
   Cette colonne contient à l'origine plusieurs faute de frappes des états maritaux. Afin d'unifier le tout, il a fallut substituer les états dans ces quatres variables distinctes : *divorced, widowed, married, single.*
   (b) CLMSEX
   Cette colonne supposé contenir les états **F** ou **M** contient à l'origine quelques états *male*. Afin d'unifier le tout, ces états ont été substituer en **M**.
2. Données aberrantes
   (a) LOSS
   Cette colonne contient une valeur très extrême de *1067.697*. En analysant le boxplot ci-dessous, on voit bien que la valeur est très énorme

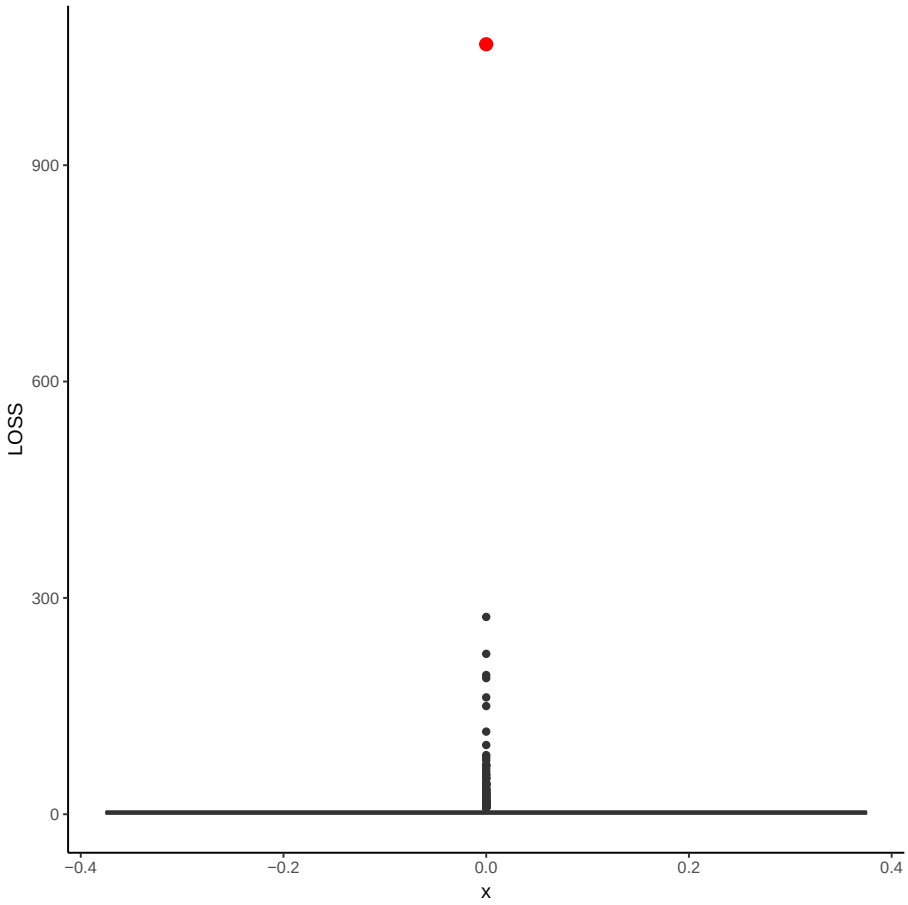
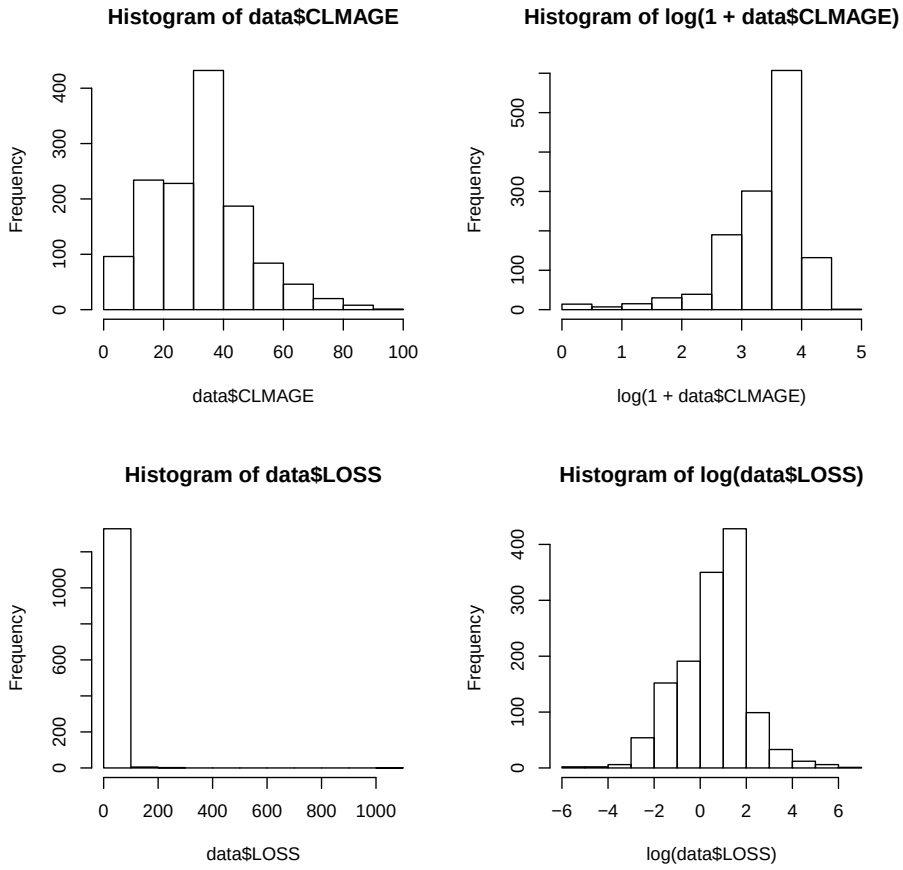


FIGURE 1 – Boxplot de LOSS

à comparer au reste. Ainsi, la ligne contenant cette valeur a été retiré afin de préserver au mieux les données jugés adéquates.

(b) CLMAGE
Cette colonne contient une valeur de *610*, valeur impossible selon la description de la variable, soit l'âge du réclamant. Afin de ne pas fausser les résultat, la ligne contenant cette valeur a été retiré.

B Transformation



C Sélection des variables