

ÉCOLE D'ACTUARIAT  
UNIVERSITÉ LAVAL

## Travail pratique 1

Guillaume MICHEL  
Nathanaël PELCHAT  
Mikael ROBERTSON  
Olivier TURCOTTE

AUTOMNE 2018

# **1 Sommaire exécutif**

## Table des matières

<b>1</b>	<b>Sommaire exécutif</b>	<b>2</b>
<b>2</b>	<b>Analyse des données</b>	<b>4</b>
<b>3</b>	<b>Modèle proposé</b>	<b>5</b>
3.1	Équation . . . . .	5
3.2	Traitement des variables qualitatives . . . . .	6
3.3	Interactions . . . . .	6
3.4	Interprétation . . . . .	7
3.5	Statistiques . . . . .	7
<b>4</b>	<b>Analyse des résidus</b>	<b>8</b>
4.1	Linéarité . . . . .	8
4.2	Homogénéité . . . . .	9
4.3	Indépendance . . . . .	9
4.4	Normalité . . . . .	10
4.5	Test pour manque d'ajustement . . . . .	10
<b>5</b>	<b>Prévisions</b>	<b>11</b>
<b>6</b>	<b>Recommandations</b>	<b>11</b>
	<b>Annexes</b>	<b>11</b>
<b>A</b>	<b>Erreurs de données</b>	<b>11</b>
<b>B</b>	<b>Transformation</b>	<b>13</b>
<b>C</b>	<b>Sélection des variables</b>	<b>14</b>

## 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

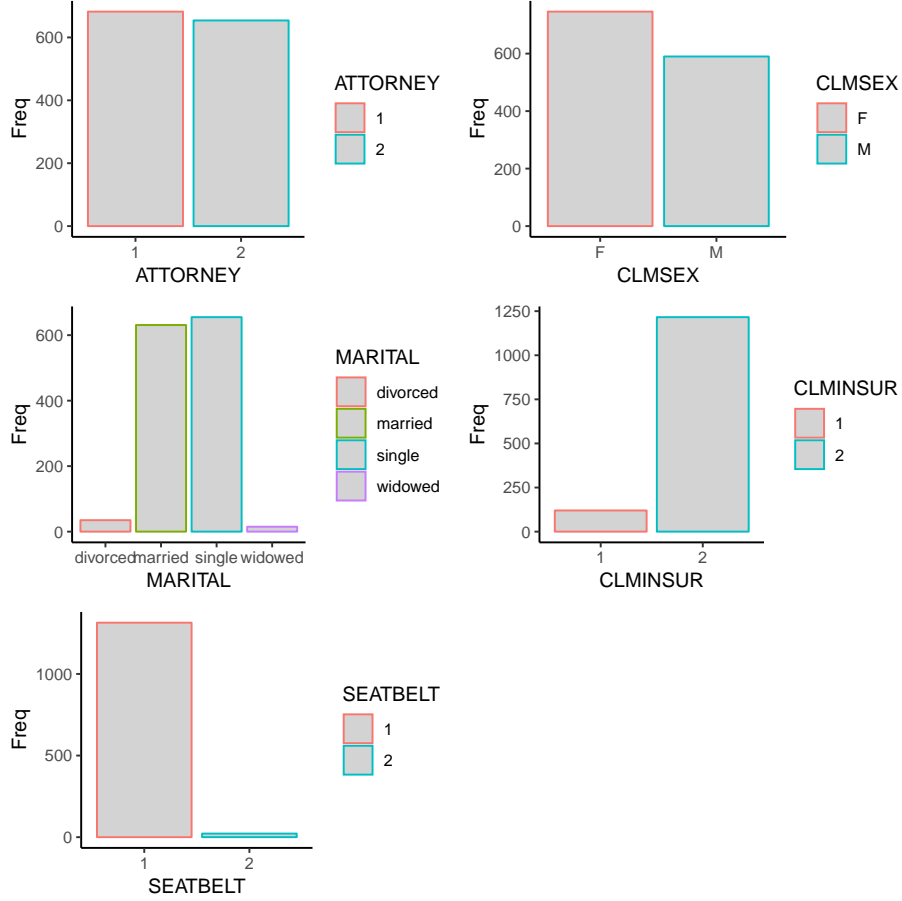
En effectuant l'analyse préliminaire univariée des différentes variables de la base de données, il est déjà possible d'avoir une bonne idée de leur pertinence et de leur pouvoir prédictif potentiel. La variable CASENUM n'est pas utilisée dans le modèle final puisqu'elle n'a pas d'incidence directe sur la variable du montant des pertes. La variable LOSS est la variable endogène continue, et la seule autre variable continue employée dans le modèle est la variable exogène CLMAGE.

Tableau 2 – Analyse des variables continues

	CLMAGE	LOSS
$\min(X)$	0.0	0.005
$F_X^{-1}(0.25)$	21	0.640
$F_X^{-1}(0.5)$	33	2.331
$E[X]$	32.6	5.965
$F_X^{-1}(0.75)$	41	3.998
$\max(X)$	95	1067.697

On s'intéresse donc à la capacité de la variable CLMAGE à prédire la valeur de la variable LOSS, tout en gardant en tête que cette variable endogène prend quelques valeurs extrêmes, dont la sévérité n'est pas nécessairement expliquée par les valeurs entières de la variable CLMAGE. L'analyse préliminaire univariée des variables continues projette donc la possibilité que le modèle ne puisse pas expliquer parfaitement la valeur du montant relié à une réclamation donnée.

La majorité de ces variables sont qualitatives. Une analyse de fréquences de celles-ci permet d'avoir un meilleur ressenti quant à leurs interactions avec la variable exogène LOSS :



### 3 Modèle proposé

#### 3.1 Équation

Le modèle choisi est donné par l'équation suivante

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

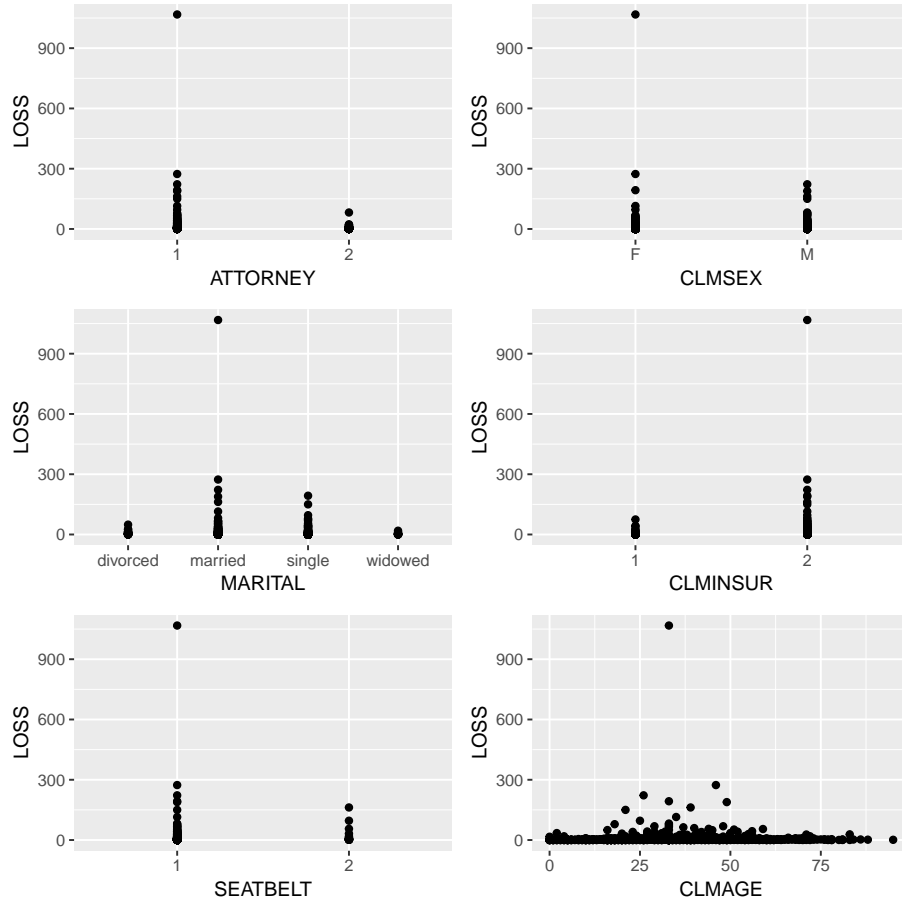


FIGURE 1 – Variable LOSS en fonction de chaque variable explicative

### 3.2 Traitement des variables qualitatives

Les variables qualitatives du modèle, soit *ATTORNEY*, *SEATBELT* et *MARITAL*, ont chacune été converties en *factor* car c'est le type de données qui est compatible avec R lorsqu'on veut utiliser ce logiciel pour effectuer des régressions linéaires comportant des variables qualitatives.

### 3.3 Interactions

Suite à la sélection des variables, il nous a été possible de déterminer qu'il n'y avait qu'une seule interaction non-redondante (qui ne cause pas de multicolinéarité) et significative au modèle. Il s'agit de l'interaction entre les variables *CLMAGE* et *ATTORNEY* qui représentent respectivement l'âge du réclamant ou de la réclamante et la présence d'un avocat pour la réclamation. Cette inter-

action est logique car, en effet, l'âge d'un réclamant peut influencer la décision de prendre un avocat.

### 3.4 Interprétation

### 3.5 Statistiques

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

Tableau 3 – Intervalles de confiance des paramètres du modèle

	2.5%	97.5%
$\beta_0$	0.28583388	1.262939359
$\beta_1$	0.01334127	0.026414523
$\beta_2$	-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
$\beta_4$	0.46855029	1.528929789
$\beta_5$	-0.02025211	-0.003201663

De plus, voici le  $R_a^2$  : 0.2754. Ainsi, une grande variabilité de la variable endogène n'est pas expliquée par le modèle. Ceci est en parti dû au grand nombre de variables qualitatives dans le modèle qui n'ont pas assez de valeurs possibles afin de refléter l'étendue des valeurs possible de la perte économique. (structure de la phrase à revoir)

Voici la table anova du modèle :

Tableau 4 – Table anova du modèle

Source	Dl	SS	MS	F
SSR	7	809.4939	115.642	73.48
SSE	1328	2089.85386	1.573685	
SST	1335	2899.348	2.171796	

Selon la statistique F du tableau 4, on peut effectuer un test de validité globale de la régression linéaire. Ce faisant, nous obtenons une p-value inférieur à  $2.2 \times 10^{-16}$  et donc nous concluons que la régression est tout à fait valide.

## 4 Analyse des résidus

### 4.1 Linéarité

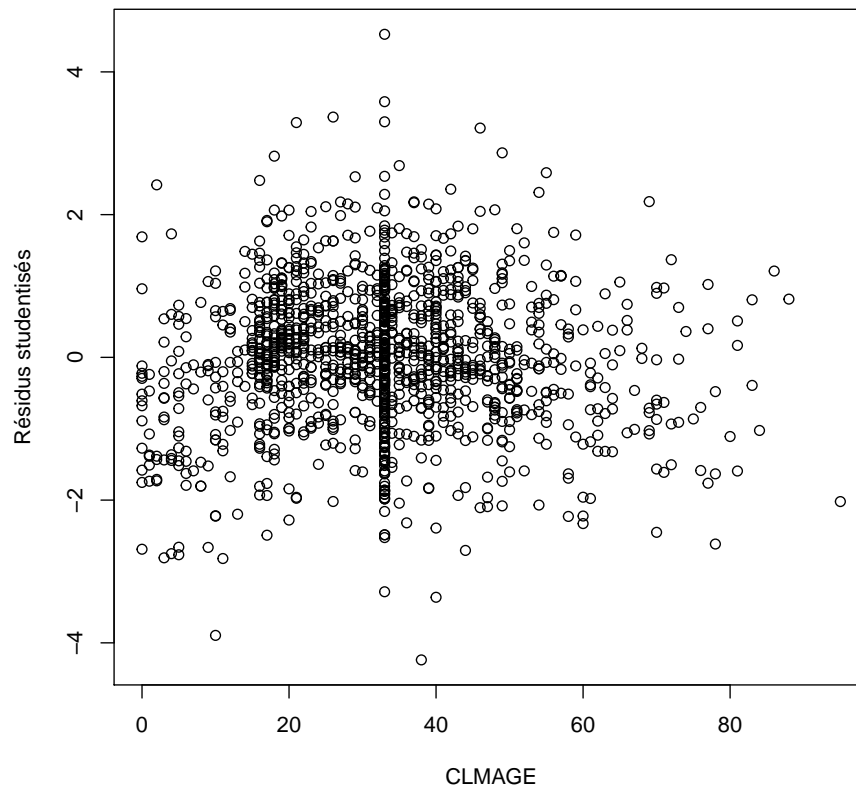


FIGURE 2 – Résidus studentisés en fonction de CLMAGE



## 4.2 Homogénéité

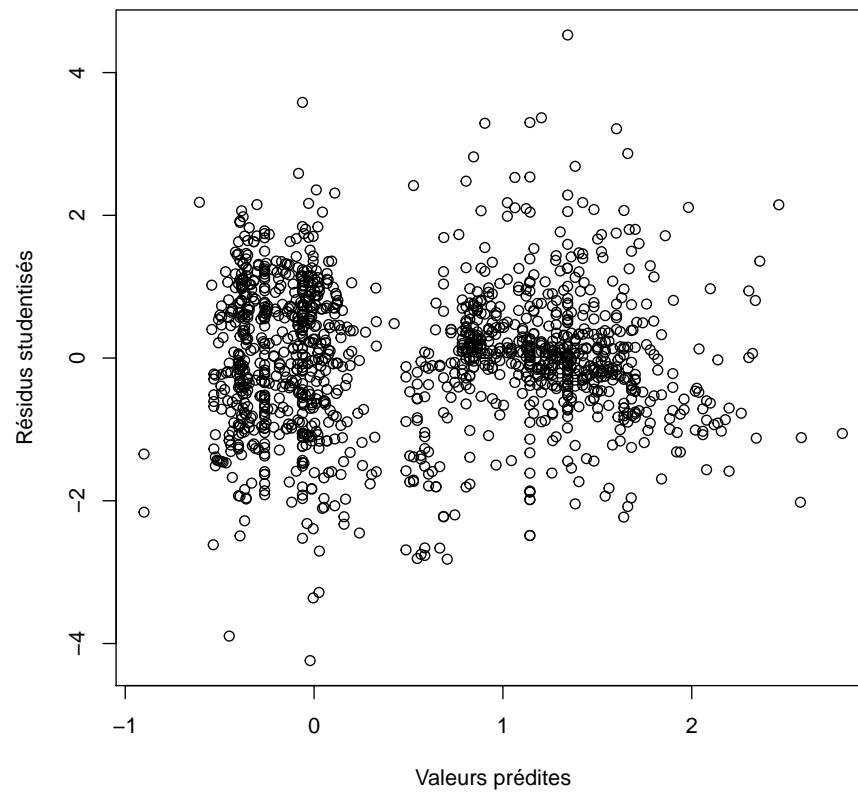


FIGURE 3 – Résidus studentisés en fonction de Y

## 4.3 Indépendance

```
##  
## Durbin-Watson test  
##  
## data: modele  
## DW = 1.9557, p-value = 0.2091  
## alternative hypothesis: true autocorrelation is greater than 0
```

## 4.4 Normalité

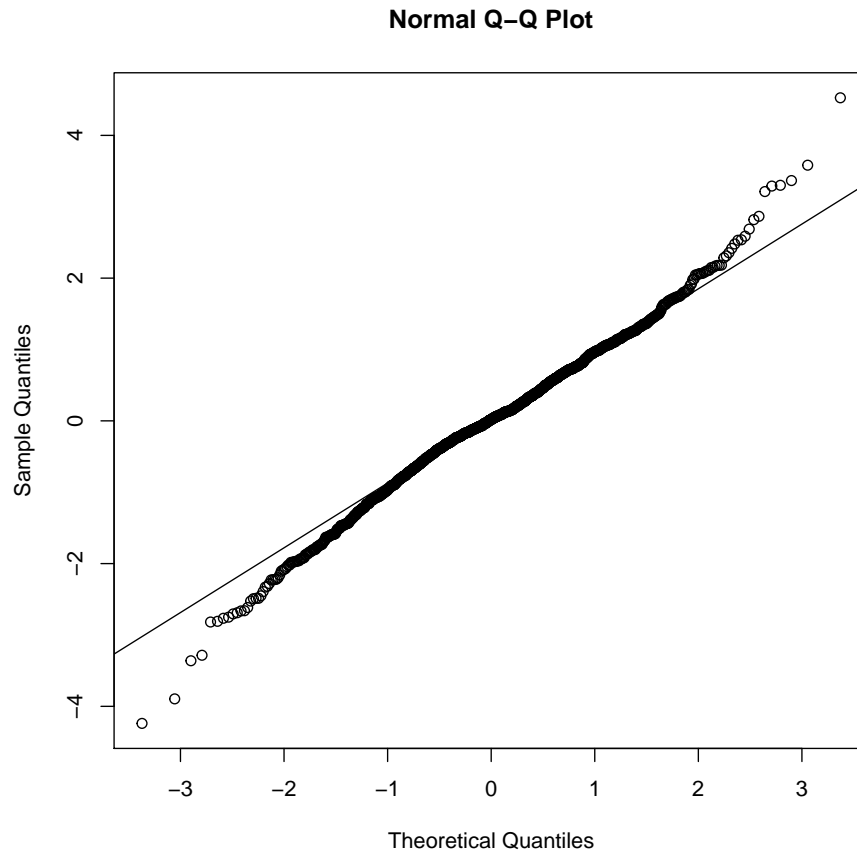


FIGURE 4 – Quantile théorique versus quantile pratique

## 4.5 Test pour manque d'ajustement

Après avoir effectué un test pour manque d'ajustement, on obtient une p-value de  $8.623 \times 10^{-5}$ , valeur inférieure au seuil de 5%. Ainsi, on doit rejeter  $H_0$ , ce qui implique que le modèle actuel ne s'ajuste pas bien aux données (sens de la phrase à clarifier).

## 5 Prévisions

Pour répondre à la question du directeur, un individu répondant aux caractéristiques suivantes : CLMAGE=45, SEATBELT=1, ATTORNEY=1, MARITAL="single" et CLMINSUR = 1, aura selon le modèle actuel une perte économique de :  $Y \in [0.3378229, 46.89545]$  à un niveau de 95%.

Avec les mêmes caractéristiques, on obtient :  $E[Y] \in [3.369781, 4.701301]$ . Ainsi, on constate la grande variabilité de la perte économique en comparant les différentes bornes des deux prévisions.

Tableau 5 – Intervalles de confiance des moyennes de la perte économique totale prévues

CLMAGE	MARITAL	CLMSEX	SEATBELT	CLMINSUR	ATTORNEY	2.5%	97.5%
70	single	M	1	1	1	4.83	8.86
45	married	M	1	1	1	4.26	5.54
45	divorced	M	1	1	1	3.47	8.10
45	widowed	M	1	1	1	1.09	4.05
45	single	F	1	1	1	3.37	4.70
45	single	M	2	1	1	6.26	18.65
45	single	M	1	2	1	3.37	4.70
45	single	M	1	1	2	0.72	1.00
22	single	F	2	1	2	1.11	3.28

## 6 Recommandations

# Annexes

## A Erreurs de données

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

### 1. Fautes d'orthographe

#### (a) MARITAL

Cette colonne contient à l'origine plusieurs fautes de frappe dans la manière de noter les états maritaux. Afin d'uniformiser le tout, il a fallu substituer les états écrits différemment dans ces quatre variables distinctes : *divorced*, *widowed*, *married*, *single*.

(b) CLMSEX

Cette colonne supposée contenir les états **F** ou **M** contient à l'origine quelques états *male*. Afin d'unifier le tout, ces états répertoriés de façon hétérogène ont été substitués en **M**.

2. Données aberrantes

(a) LOSS

Cette colonne contient une valeur très extrême de *1067.697*. En analysant la figure 5, on voit bien que la valeur est très énorme comparativement à l'ensemble des autres valeurs prises par cette variable. Néanmoins, comme il s'agit de perte économique aux États-Unis suite à une blessure corporelle, cette valeur est possible donc nous ne l'avons pas retirée de l'étude.

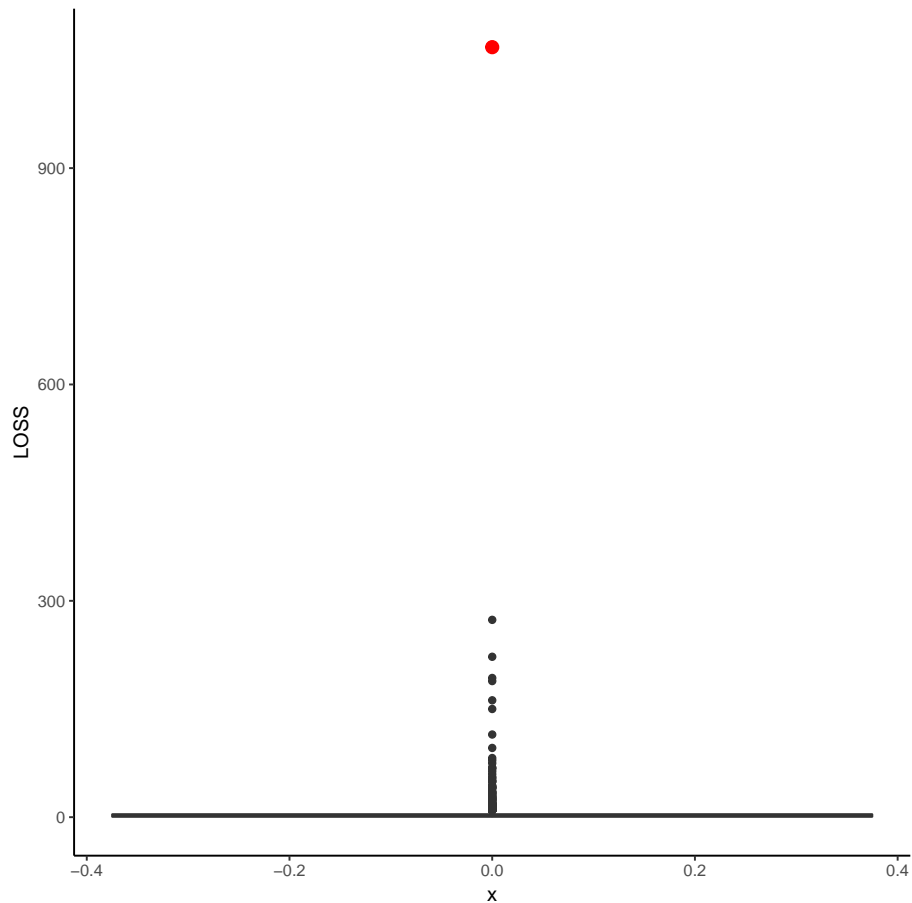
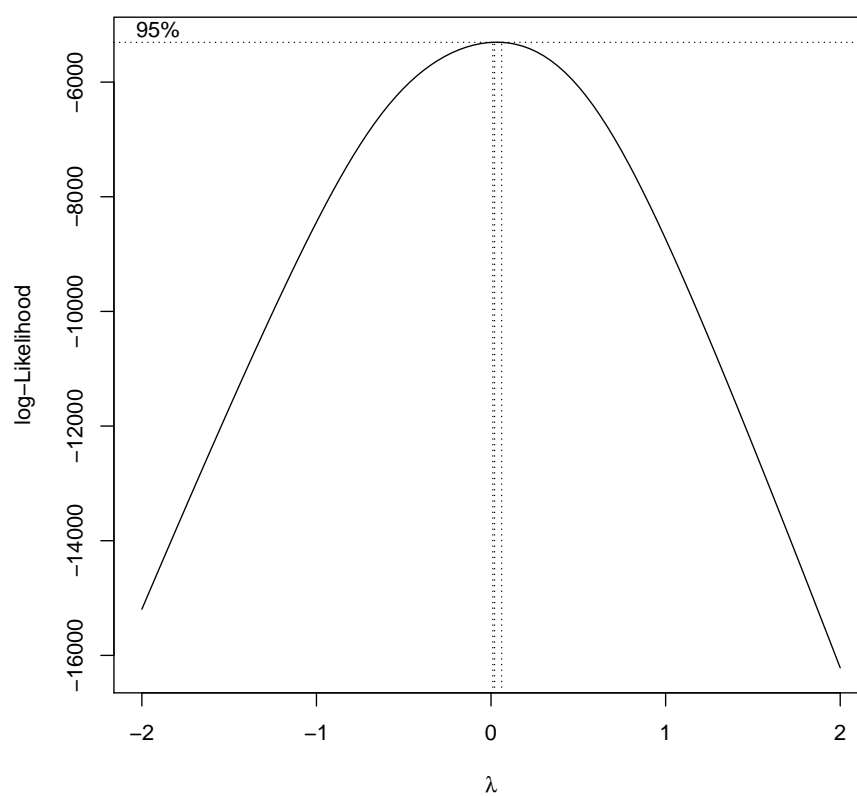


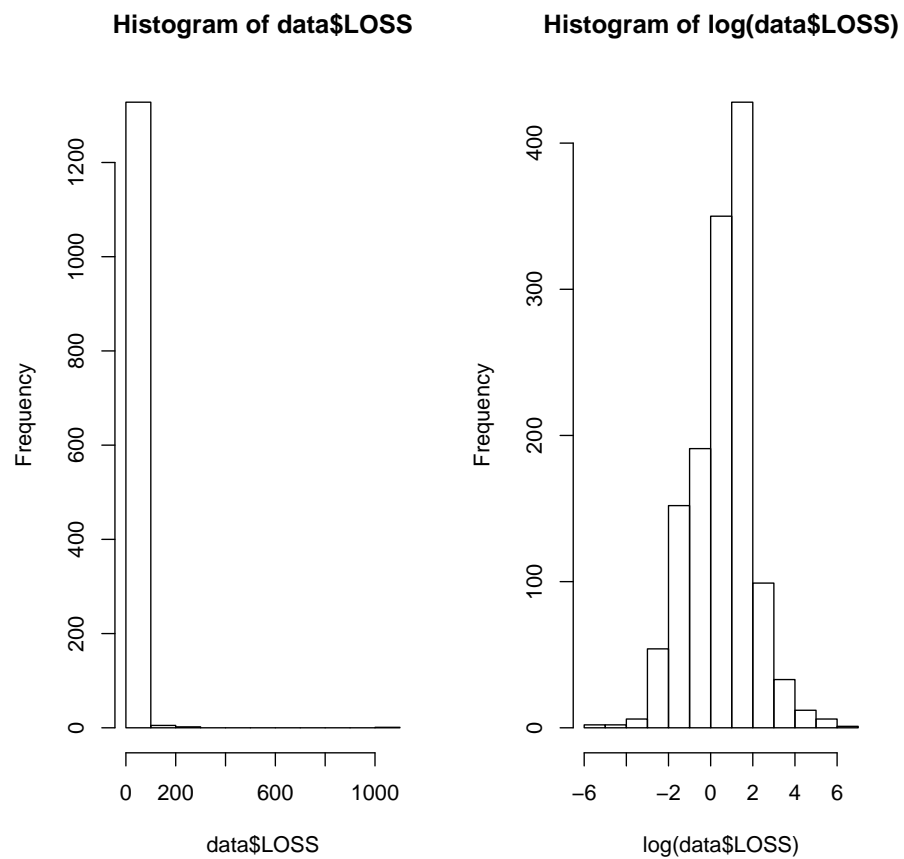
FIGURE 5 – Boxplot de LOSS

(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 résultats, cette valeur a été modifiée pour 61.

## B Transformation





## C Sélection des variables