

THÉORIE ET APPLICATIONS DES MÉTHODES DE RÉGRESSION

TRAVAIL 1

Présenté Par

RAHMA JEBALI  
HAJAR TAQIF  
OUMAIMA OUFFY

Novembre 2020

## Table des matières

<b>1</b>	<b>Exercice 1</b>	<b>3</b>
1.1	Chargement des données . . . . .	3
1.2	Traitement des données . . . . .	4
1.3	Sélection de variables . . . . .	5
1.3.1	Résolution de la multicolinéarité . . . . .	5
1.3.2	Sélection du meilleur AIC . . . . .	8
1.4	Prédiction demandée . . . . .	9
<b>2</b>	<b>Exercice 2</b>	<b>9</b>
2.1	Chargement des données . . . . .	10
2.2	Définir les variables . . . . .	10
2.3	Multicolinéarité . . . . .	11
2.3.1	Détection du problème . . . . .	11
2.3.2	Correction de la multicolinéarité . . . . .	12
2.4	Sélection du modèle : Modèle GLM . . . . .	13
2.5	Sélection du meilleur modèle . . . . .	13
<b>3</b>	<b>Exercice 3</b>	<b>15</b>
3.1	Notations . . . . .	15
3.2	Distribution de la variable réponse et fonction de lien . . . . .	15
3.3	Modèle adopté . . . . .	15
<b>4</b>	<b>Annexe</b>	<b>22</b>

## Table des figures

1	Échantillon des données du taux de mortalité, de la pollution environnementale et des caractéristiques socio-démographiques .	3
2	Ajustement du modèle . . . . .	4
3	Tableau ANOVA . . . . .	4
4	Les Facteur VIFs du modèle complet . . . . .	5
5	Les indices de conditionnement et les proportions de variabilités	6
6	Les indices de conditionnement et les proportions de variabilités (suite) . . . . .	7
7	Facteur VIF du modèle sans la variable A13 . . . . .	8
8	Modèle avec le meilleur AIC . . . . .	8
9	Ajustement du modèle sélectionné . . . . .	9
10	L'estimer ponctuel et l'intervalle de confiance à 95% pour le taux de mortalité . . . . .	9
11	Echantillon des données . . . . .	10
12	Déclaration des variables . . . . .	11
13	Les Facteur VIFs du modèle complet . . . . .	12
14	Facteur VIF du modèle sans les variables ca . . . . .	12
15	Ajustement du modèle . . . . .	13
16	Méthode d'inclusion . . . . .	14
17	Tous les sous-modèles . . . . .	14
18	Calcul des VIFS . . . . .	16
19	Ajustement du modèle avec poisson ordinaire . . . . .	17
20	Ajustement du modèle avec poisson ordinaire en ajoutant une variable offset . . . . .	18
21	Ajustement du modèle avec quasi-poisson . . . . .	19
22	Ajustement du modèle avec une loi binomiale négative . . . . .	20
23	Test d'hypothèse . . . . .	21
24	Méthode de sélection 1 . . . . .	21
25	Méthode de sélection 2 . . . . .	21
26	Méthode de sélection 3 . . . . .	22

# 1 Exercice 1

Dans cet exercice nous traitons la demande de McDonald & Schwing (1973) sur la possibilité de prédire le taux de mortalité à partir des données mesurant la pollution environnementale et à partir des caractéristiques socio-démographiques.

Nous avons calculé alors un estimateur ponctuel ainsi qu'un intervalle de confiance à 95% pour le taux de mortalité à un endroit de 60 localités.

## 1.1 Chargement des données

Les données sont disponibles dans un fichier TXT contenant respectivement dans l'ordre 15 mesures de nature pollution environnementale et de nature caractéristiques socio-démographiques que nous allons les déclarer comme des variables explicatives :  $A_i$ ,  $i = \{1..15\}$  et des mesures sur le taux de mortalités de 60 localités que nous allons la déclarer comme variable de réponse : B. Un échantillon des données est présenté dans la figure 1.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
1	36	27	71	8.1	3.34	11.4	81.5	3243	8.8	42.6	11.7	21	15	59	59	921.870
2	35	23	72	11.1	3.14	11.0	78.8	4281	3.6	50.7	14.4	8	10	39	57	997.875
3	44	29	74	10.4	3.21	9.8	81.6	4260	0.8	39.4	12.4	6	6	33	54	962.354
4	47	45	79	6.5	3.41	11.1	77.5	3125	27.1	50.2	20.6	18	8	24	56	982.291
5	43	35	77	7.6	3.44	9.6	84.6	6441	24.4	43.7	14.3	43	38	206	55	1071.289
6	53	45	80	7.7	3.45	10.2	66.8	3325	38.5	43.1	25.5	30	32	72	54	1030.380
7	43	30	74	10.9	3.23	12.1	83.9	4679	3.5	49.2	11.3	21	32	62	56	934.700
8	45	30	73	9.3	3.29	10.6	86.0	2140	5.3	40.4	10.5	6	4	4	56	899.529
9	36	24	70	9.0	3.31	10.5	83.2	6582	8.1	42.5	12.6	18	12	37	61	1001.902
10	36	27	72	9.5	3.36	10.7	79.3	4213	6.7	41.0	13.2	12	7	20	59	912.347
11	52	42	79	7.7	3.39	9.6	69.2	2302	22.2	41.3	24.2	18	8	27	56	1017.613
12	33	26	76	8.6	3.20	10.9	83.4	6122	16.3	44.9	10.7	88	63	278	58	1024.885
13	40	34	77	9.2	3.21	10.2	77.0	4101	13.0	45.7	15.1	26	26	146	57	970.467
14	35	28	71	8.8	3.29	11.1	86.3	3042	14.7	44.6	11.4	31	21	64	60	985.950
15	37	31	75	8.0	3.26	11.9	78.4	4259	13.1	49.6	13.9	23	9	15	58	958.839
16	35	46	85	7.1	3.22	11.8	79.9	1441	14.8	51.2	16.1	1	1	1	54	860.101
17	36	30	75	7.5	3.35	11.4	81.9	4029	12.4	44.0	12.0	6	4	16	58	936.234
18	15	30	73	8.2	3.15	12.2	84.2	4824	4.7	53.1	12.7	17	8	28	38	871.766
19	31	27	74	7.2	3.44	10.8	87.0	4834	15.8	43.5	13.6	52	35	124	59	959.221
20	30	24	72	6.5	3.53	10.8	79.5	3694	13.1	33.8	12.4	11	4	11	61	941.181

FIGURE 1 – Échantillon des données du taux de mortalité, de la pollution environnementale et des caractéristiques socio-démographiques

## 1.2 Traitement des données

Afin d'identifier les variables explicatives pertinentes pour la prédiction demandée, nous avons commencé par ajuster le modèle. Nous avons choisi de travailler avec un modèle linéaire simple. Nous avons remarqué d'après les figures 2 et 3 (qui résument les résultats de cet ajustement) qu'au moins une des variables est dans le modèle (puisque la sig=0) et que plusieurs variables sont reliées à des p\_value supérieurs à 0.05. Nous avons obtenu un  $R^2_{ajus} = 0.73$ .

```
Call:
lm(formula = B ~ A1 + A2 + A3 + A4 + A5 + A6 + A7 + A8 + A9 +
    A10 + A11 + A12 + A13 + A14 + A15, data = ex1)

Residuals:
    Min       1Q   Median       3Q      Max
-75.285 -14.640   0.694  14.790  75.586

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.863e+03  4.108e+02   4.535  4.4e-05 ***
A1           2.072e+00  8.418e-01   2.462  0.01781 *
A2          -2.178e+00  6.752e-01  -3.225  0.00238 **
A3          -2.834e+00  1.771e+00  -1.600  0.11670
A4          -1.404e+01  7.746e+00  -1.813  0.07670 .
A5          -1.154e+02  6.200e+01  -1.862  0.06933 .
A6          -2.425e+01  1.121e+01  -2.163  0.03605 *
A7          -1.146e+00  1.467e+00  -0.781  0.43871
A8           1.004e-02  4.123e-03   2.435  0.01899 *
A9           3.533e+00  1.282e+00   2.755  0.00850 **
A10          5.229e-01  1.551e+00   0.337  0.73760
A11          2.671e-01  2.565e+00   0.104  0.91755
A12          -8.890e-01  4.524e-01  -1.965  0.05574 .
A13          1.866e+00  9.345e-01   1.997  0.05201 .
A14          -3.447e-02  1.423e-01  -0.242  0.80968
A15          5.331e-01  1.052e+00   0.507  0.61474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.33 on 44 degrees of freedom
Multiple R-squared:  0.7985,    Adjusted R-squared:  0.7298
F-statistic: 11.63 on 15 and 44 DF,  p-value: 9.56e-11
```

FIGURE 2 – Ajustement du modèle

ANOVA					
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	182306.880	15	12153.792	11.625	0.0000
Residual	46000.764	44	1045.472		
Total	228307.644	59			

FIGURE 3 – Tableau ANOVA

### 1.3 Sélection de variables

#### 1.3.1 Résolution de la multicolinéarité

Afin d'améliorer le modèle, nous avons identifié à l'aide du facteur VIF la possibilité de la multicolinéarité entre les variables. Le résultat obtenu présenté dans la figure 4 montre que les VIFs des variables A13 et A14 sont supérieurs à 10 (respectivement 97,71 et 105,91). Nous avons déduit alors l'existence de la multicolinéarité entre les variables.

Tolerance and Variance Inflation Factor			
	Variables	Tolerance	VIF
1	A1	0.250851616	3.986420
2	A2	0.271047883	3.689385
3	A3	0.248719632	4.020591
4	A4	0.137672419	7.263619
5	A5	0.251962606	3.968843
6	A6	0.197278383	5.068979
7	A7	0.311694307	3.208272
8	A8	0.492963984	2.028546
9	A9	0.135413520	7.384787
10	A10	0.344079304	2.906307
11	A11	0.155615067	6.426113
12	A12	0.010233894	97.714514
13	A13	0.009441763	105.912420
14	A14	0.217838328	4.590560
15	A15	0.537419693	1.860743

FIGURE 4 – Les Facteur VIFs du modèle complet

Par la suite, nous avons effectué le diagnostic de multicolinéarité en calculant les indices de conditionnement et nous avons considéré les proportions de variabilités relatives au indice le plus élevé (présenté dans la figure 5 et 6) :

# Eigenvalue and Condition Index

	Eigenvalue	Condition Index	intercept	A1	A2
1	1.315218e+01	1.000000	5.883098e-07	8.789732e-05	1.506476e-04
2	1.696400e+00	2.784420	2.282526e-07	2.682867e-04	8.518035e-07
3	5.138375e-01	5.059248	8.092784e-07	5.946238e-07	1.347225e-03
4	3.821914e-01	5.866221	2.430126e-06	1.551024e-04	1.769459e-03
5	1.028532e-01	11.308107	1.432259e-05	1.972008e-04	7.175117e-02
6	5.830242e-02	15.019505	6.133447e-06	1.105823e-02	1.602379e-01
7	4.550580e-02	17.000651	8.482885e-06	1.451222e-01	1.064420e-01
8	2.004698e-02	25.613826	2.026393e-05	2.627654e-01	1.025463e-01
9	9.576441e-03	37.059269	3.168904e-06	4.092820e-02	2.177752e-03
10	7.672103e-03	41.403943	5.096156e-04	4.168061e-02	9.812957e-03
11	4.681268e-03	53.005038	6.461459e-05	7.679297e-02	4.035946e-02
12	3.329269e-03	62.852772	5.065435e-04	3.470984e-01	2.382908e-01
13	1.545281e-03	92.256110	1.609595e-03	1.875540e-04	8.481267e-03
14	1.032305e-03	112.874232	3.208481e-04	1.431233e-02	2.590945e-02
15	7.620106e-04	131.376719	1.033771e-03	4.217022e-03	1.990702e-01
16	8.194659e-05	400.620776	9.958986e-01	5.512801e-02	3.165258e-02
	A3	A4	A5	A6	A7
1	5.655870e-06	2.009837e-05	2.413696e-06	6.468984e-06	6.998851e-06
2	3.727652e-06	7.466005e-06	1.331057e-06	1.509647e-06	1.274881e-06
3	5.232961e-06	4.099305e-05	2.013138e-06	2.485936e-05	1.269616e-05
4	3.227799e-06	6.811361e-04	5.561299e-06	3.677993e-05	6.217567e-05
5	8.175540e-05	4.261551e-04	8.706697e-05	3.350682e-04	1.354348e-04
6	6.438263e-06	4.484201e-04	5.403552e-05	4.965704e-04	3.523046e-04
7	8.881180e-05	1.137334e-02	4.032700e-05	1.192725e-03	6.321201e-04
8	3.236831e-04	1.646006e-03	1.542046e-04	8.962365e-05	5.850062e-04
9	1.076446e-02	1.256287e-02	1.408041e-07	3.039526e-03	4.831371e-04
10	8.771955e-03	1.234565e-01	1.141953e-02	1.192619e-03	7.377538e-03
11	1.911777e-04	5.595012e-03	8.940513e-04	2.917776e-03	5.332484e-05
12	1.021517e-02	3.924956e-01	2.073497e-03	1.999129e-02	1.208550e-02
13	7.344215e-03	4.614191e-04	6.978473e-02	1.614685e-01	2.645772e-01
14	5.964660e-01	1.505232e-05	7.488148e-02	4.855221e-03	1.302338e-01
15	4.138283e-02	5.089220e-02	2.648086e-02	6.934332e-01	5.004563e-01
16	3.243456e-01	3.998777e-01	8.141188e-01	1.109183e-01	8.294518e-02

FIGURE 5 – Les indices de conditionnement et les proportions de variabilités

	A8	A9	A10	A11	A12
1	3.120280e-04	0.0001951391	1.911421e-05	6.395469e-05	1.021285e-05
2	1.459012e-06	0.0001371357	4.587867e-06	6.612858e-05	2.306696e-03
3	2.281221e-03	0.0038202198	5.020705e-05	5.468802e-05	1.097545e-03
4	6.098644e-03	0.0873703614	7.562655e-05	2.116277e-03	9.604443e-05
5	3.425979e-01	0.0012268442	4.783708e-04	2.626341e-04	4.770025e-06
6	1.994777e-01	0.1091485500	1.174009e-03	3.488991e-02	9.384848e-04
7	5.219299e-02	0.0001711039	5.083593e-03	8.993041e-03	7.019606e-05
8	1.074513e-01	0.0172169260	6.646663e-04	3.030123e-01	1.865947e-04
9	3.503060e-05	0.0448922857	4.354352e-02	1.270835e-03	7.409712e-04
10	5.634835e-03	0.1102431767	1.440542e-01	2.371290e-04	6.888840e-03
11	6.078570e-04	0.0332942041	1.987757e-02	4.806728e-02	7.879080e-01
12	2.595807e-01	0.5626445849	9.822929e-02	2.631358e-01	4.904781e-02
13	4.110470e-03	0.0069562504	2.904780e-01	1.559153e-01	1.344037e-03
14	2.538418e-03	0.0006021579	5.660107e-02	9.977022e-02	1.864662e-03
15	2.810093e-03	0.0213433318	3.309586e-01	6.216322e-02	1.359548e-01
16	1.426943e-02	0.0007377284	8.707534e-03	1.998127e-02	1.154036e-02
	A13	A14	A15		
1	1.149458e-05	3.565405e-04	2.661937e-05		
2	1.933070e-03	3.700879e-03	1.118481e-05		
3	9.589206e-05	1.950754e-01	5.548793e-05		
4	5.938116e-05	9.569153e-03	1.411789e-04		
5	1.983041e-04	1.532825e-02	8.058002e-04		
6	2.306309e-04	9.035933e-02	1.643583e-04		
7	4.529733e-03	4.536080e-02	1.653742e-03		
8	1.156322e-05	4.639720e-03	1.914656e-03		
9	1.751033e-03	3.638742e-04	3.315349e-01		
10	1.213255e-03	5.512278e-05	6.647577e-05		
11	7.342891e-01	3.604687e-01	1.375831e-03		
12	8.204322e-02	1.098667e-02	1.522739e-01		
13	8.910217e-03	2.172156e-02	1.370855e-02		
14	3.043911e-05	5.307635e-06	2.473642e-01		
15	1.593085e-01	2.298087e-01	1.094360e-01		
16	5.384116e-03	1.220006e-02	1.394671e-01		

FIGURE 6 – Les indices de conditionnement et les proportions de variabilités (suite)

Nous remarquons que seulement la variable qui correspond à l'ordonnée à l'origine et la variable A5 sont supérieurs à 60%.

Nous avons effectué alors 3 tests (à chaque test nous avons enlevé une variable respectivement A5, A13, et A14 et nous avons réajusté le modèle). Nous avons constaté que seulement le test qui correspond à enlever A13 règle le problème (comme montré dans la figure 7 qui contient le résultat des VIFs obtenue avec ce dernier modèle) :



Tolerance and Variance Inflation Factor

	Variables	Tolerance	VIF
1	A1	0.2514659	3.976683
2	A2	0.2788609	3.586017
3	A3	0.2490257	4.015649
4	A4	0.1460044	6.849110
5	A5	0.2544764	3.929637
6	A6	0.2177289	4.592868
7	A7	0.3223219	3.102489
8	A8	0.4965025	2.014088
9	A9	0.1375238	7.271468
10	A10	0.3686655	2.712486
11	A11	0.1556645	6.424074
12	A12	0.3158762	3.165797
13	A14	0.5289508	1.890535
14	A15	0.5537073	1.806008

FIGURE 7 – Facteur VIF du modèle sans la variable A13

Par la suite, nous avons enlevé cette variable de notre modèle.

### 1.3.2 Sélection du meilleur AIC

Afin de sélectionner le meilleur modèle, nous avons procédé à modéliser tous les sous modèles possibles à l'aide de la commande `ols_step_all_possible` et nous avons sélectionné le modèle avec le meilleur AIC (c'est à dire le modèle avec le plus petit AIC) (présenté dans la figure 8).

	mindex	n	predictors							rsquare	adjr	predrsq	cp	aic	sbic	sbc	msep	fpe	apc	hsp
<b>6704</b>	6476	7	A1	A2	A3	A6	A8	A9	A14	0.7604988	0.7282582	0.6774085	5.044079	597.1671	430.3689	616.0162	63221.47	1191.743	0.3131939	20.61839

FIGURE 8 – Modèle avec le meilleur AIC

Finalement, les variables sélectionnées avec un  $AIC = 597.17$  sont les variables A1 A2 A3 A6 A8 A9 A14. L'ajustement du modèle avec seulement ces variables (présenté dans la figure 9) a eu comme  $R_{ajus}^2 = 0.73$  qui est la même valeur que celui du modèle initial.

```

Call:
lm(formula = B ~ A1 + A2 + A3 + A6 + A8 + A9 + A14, data = ex1)

Residuals:
    Min       1Q   Median       3Q      Max
-84.67 -18.46   0.40  17.76  93.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.167e+03  1.174e+02   9.945 1.26e-13 ***
A1           1.682e+00  5.773e-01   2.913 0.005261 **
A2          -1.482e+00  4.063e-01  -3.646 0.000616 ***
A3          -2.144e+00  1.197e+00  -1.791 0.079164 .
A6          -1.573e+01  6.172e+00  -2.548 0.013825 *
A8           8.078e-03  3.471e-03   2.328 0.023865 *
A9           4.529e+00  6.447e-01   7.024 4.53e-09 ***
A14          1.724e-01  8.373e-02   2.059 0.044563 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.43 on 52 degrees of freedom
Multiple R-squared:  0.7605,    Adjusted R-squared:  0.7283
F-statistic: 23.59 on 7 and 52 DF,  p-value: 4.61e-14

```

FIGURE 9 – Ajustement du modèle sélectionné

## 1.4 Prédiction demandée

Nous avons calculé l'estimer ponctuel ainsi que l'intervalle de confiance à 95% pour le taux de mortalité à l'endroit pour lequel les variables explicatives valent respectivement (40, 30, 80, 9, 3, 10, 77, 4100, 13, 46, 15, 25, 26, 145, 55) en se basant sur le modèle final sélectionné (le résultat est présenté dans la figure 10)

fit	lwr	upr
978.2874	909.5171	1047.058

FIGURE 10 – L'estimer ponctuel et l'intervalle de confiance à 95% pour le taux de mortalité

Nous avons obtenu le taux de mortalité  $B = 978.29$  qui correspond à un intervalle de confiance à 95% [ 909.52,1047.06]

## 2 Exercice 2

L'objectif de cet exercice est de construire un modèle de régression qui estime la probabilité de diagnostic de maladie coronarienne positif, et de trouver les facteurs qui semblent associés à une hausse du risque d'un diagnostic positif de maladie coronarienne.

## 2.1 Chargement des données

Le jeu de données `processed.cleveland.data` est un fichier contenant la valeur de 13 variables explicatives (`age`, `sex`, `cp`, `trestbps`, `chol`, `fbs`...), et la probabilité de diagnostic de maladie coronarienne positif comme variable réponse.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	Y
1	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	
2	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	
3	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	
4	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	
5	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	
6	56	1	2	120	236	0	0	178	0	0.8	1	0.0	3.0	
7	62	0	4	140	268	0	2	160	0	3.6	3	2.0	3.0	
8	57	0	4	120	354	0	0	163	1	0.6	1	0.0	3.0	
9	63	1	4	130	254	0	2	147	0	1.4	2	1.0	7.0	
10	53	1	4	140	203	1	2	155	1	3.1	3	0.0	7.0	
11	57	1	4	140	192	0	0	148	0	0.4	2	0.0	6.0	
12	56	0	2	140	294	0	2	153	0	1.3	2	0.0	3.0	
13	56	1	3	130	256	1	2	142	1	0.6	2	1.0	6.0	
14	44	1	2	120	263	0	0	173	0	0.0	1	0.0	7.0	
15	52	1	3	172	199	1	0	162	0	0.5	1	0.0	7.0	
16	57	1	3	150	168	0	0	174	0	1.6	1	0.0	3.0	
17	48	1	2	110	229	0	0	168	0	1.0	3	0.0	7.0	
18	54	1	4	140	239	0	0	160	0	1.2	1	0.0	3.0	
19	48	0	3	130	275	0	0	139	0	0.2	1	0.0	3.0	
20	49	1	2	130	266	0	0	171	0	0.6	1	0.0	3.0	
21	64	1	1	110	211	0	2	144	1	1.8	2	0.0	3.0	
22	58	0	1	150	288	1	2	162	0	1.0	1	0.0	3.0	

FIGURE 11 – Echantillon des données

## 2.2 Définir les variables

nous allons les déclarer les variables explicatives comme :  $X_i$ ,  $i = \{1..13\}$ , et la variable réponse  $Y$ .

```

> ex2<- processed_cleveland
> age<-ex2$X1
> sex<-ex2$X2
> cp<-ex2$X3
> cp_1 <- ifelse(test = cp==1,yes = 1,no = 0)
> cp_2 <- ifelse(test = cp==2,yes = 1,no = 0)
> cp_3 <- ifelse(test = cp==3,yes = 1,no = 0)
> trestbps<-ex2$X4
> chol<-ex2$X5
> fbs<-ex2$X6
> restecg<-ex2$X7
> restecg_1 <- ifelse(test = restecg==1,yes = 1,no = 0)
> restecg_2 <- ifelse(test = restecg==2,yes = 1,no = 0)
> thalach<-ex2$X8
> exang<-ex2$X9
> oldpeak<-ex2$X10
> slope<-ex2$X11
> slope_1 <- ifelse(test = slope==1,yes = 1,no = 0)
> slope_2 <- ifelse(test = slope==2,yes = 1,no = 0)
> ca<-ex2$X12
> thal<-ex2$X13
> thal_1 <- ifelse(test = thal==1,yes = 1,no = 0)
> thal_2 <- ifelse(test = thal==2,yes = 1,no = 0)
> Y<-ifelse(test = ex2$X14 > 0,yes = 1,no = 0)

```

FIGURE 12 – Déclaration des variables

## 2.3 Multicolinéarité

### 2.3.1 Détection du problème

Afin de détecter le problème de multi colinéarité, nous avons utilisé le facteur VIF. Le résultat obtenu montre que les VIFs des variables explicatives ca et thal ont un vif supérieur à 10 donc on doit corriger le problème de multicolinéarité.

```
> ols_vif_tol(modele_complet)
  variables  Tolerance      VIF
1      age 0.64662965  1.546480
2      sex 0.82726373  1.208804
3     cp_1 0.80527893  1.241806
4     cp_2 0.66032672  1.514402
5     cp_3 0.62798614  1.592392
6  trestbps 0.82196615  1.216595
7     chol 0.86126467  1.161083
8     fbs 0.87966485  1.136797
9  restecg_1 0.91836487  1.088892
10 restecg_2 0.88758910  1.126647
11  thalach 0.57349995  1.743679
12    exang 0.67983408  1.470947
13  oldpeak 0.52240989  1.914206
14  slope_1 0.16680822  5.994908
15  slope_2 0.20166343  4.958757
16    ca0.0 0.05138771 19.459908
17    ca1.0 0.07033994 14.216674
18    ca2.0 0.09991135 10.008873
19    ca3.0 0.16641999  6.008894
20  thal_1 0.84092932  1.189161
21  thal_2 0.77452273  1.291118
```

FIGURE 13 – Les Facteur VIFs du modèle complet

### 2.3.2 Correction de la multicolinéarité

Nous devons enlever la variable ca pour remédier le problème de multicolinéarité.

```
> modele_complet_lm<- lm (Y~age+sex+cp_1+cp_2+cp_3+trestbps+chol+fbs+restecg_1+restecg_2+thalach+exang+oldpeak+slope_1+slope_2+thal_1+thal_2,data = ex2,x = TRUE, y = TRUE)
> ols_vif_tol(modele_complet_lm)
  variables  Tolerance      VIF
1      age 0.7140059  1.400549
2      sex 0.8512151  1.174791
3     cp_1 0.8224429  1.215890
4     cp_2 0.6705958  1.491211
5     cp_3 0.6666992  1.499927
6  trestbps 0.8342976  1.198613
7     chol 0.8677725  1.152376
8     fbs 0.9046732  1.105371
9  restecg_1 0.9276409  1.078003
10 restecg_2 0.8952542  1.117001
11  thalach 0.5894318  1.696549
12    exang 0.6938731  1.441186
13  oldpeak 0.5667862  1.764334
14  slope_1 0.1727959  5.787175
15  slope_2 0.2071274  4.827945
16  thal_1 0.8614864  1.160784
17  thal_2 0.7996331  1.250574
```

FIGURE 14 – Facteur VIF du modèle sans les variables ca

Nous remaquons que les VIFs de toutes les variables sont inférieurs à 10 par

la suite le problème de la multicolinéarité est résolu.

## 2.4 Sélection du modèle : Modèle GLM

Nous avons choisi de travailler avec le modèle linéaire généralisés sans les variables ca.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.08429  -0.18898  -0.02770   0.00003   2.56277

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.088e+00  4.664e+00  -1.734  0.082933 .
age          1.681e-02  3.711e-02   0.453  0.650557
sex          2.060e+00  7.890e-01   2.611  0.009020 **
cp_1        -5.215e+00  1.523e+00  -3.424  0.000618 ***
cp_2        -1.047e+00  9.771e-01  -1.071  0.284043
cp_3        -3.219e+00  9.561e-01  -3.367  0.000759 ***
trestbps     1.107e-02  1.869e-02   0.592  0.553668
chol         2.763e-03  6.324e-03   0.437  0.662157
fbs          1.902e+00  8.793e-01   2.163  0.030565 *
restecg_1    2.385e+00  3.070e+00   0.777  0.437222
restecg_2    1.226e+00  6.364e-01   1.927  0.054018 .
thalach     -9.994e-03  1.585e-02  -0.631  0.528267
exang        1.375e+00  6.771e-01   2.030  0.042336 *
oldpeak      1.667e+00  4.148e-01   4.020  5.83e-05 ***
slope_1      9.184e-01  1.523e+00   0.603  0.546621
slope_2      2.170e+00  1.384e+00   1.568  0.116858
thal_1       2.470e+01  1.784e+03   0.014  0.988954
thal_2       2.203e+01  2.242e+03   0.010  0.992160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.982  on 302  degrees of freedom
Residual deviance:  81.137  on 285  degrees of freedom
AIC: 117.14

Number of Fisher Scoring iterations: 19
```

FIGURE 15 – Ajustement du modèle

Nous remarquons que les variables explicatives sex, cp, fbs, exang, oldpeak ont des p\_value inférieurs à 0.05 donc elles sont significatives. Elles expliquent la probabilité de diagnostic de maladie coronarienne positif.

## 2.5 Sélection du meilleur modèle

Afin de sélectionner les variables les plus pertinentes dans notre modèle, nous allons utiliser la méthodes algorithmique (méthode d'inclusion) et la méthode de sélection pour tous les sous-modèles .

```
Initial Model:
Y ~ 1

Final Model:
Y ~ thal_1 + thal_2 + oldpeak + exang + cp_1 + cp_3 + fbs + slope_2 +
sex + restecg_2
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				302	417.98214	419.9821
2	+ thal_1	1	100.454890	301	317.52725	321.5272
3	+ thal_2	1	90.726153	300	226.80110	232.8011
4	+ oldpeak	1	57.894440	299	168.90666	176.9067
5	+ exang	1	25.830236	298	143.07642	153.0764
6	+ cp_1	1	13.618704	297	129.45772	141.4577
7	+ cp_3	1	16.220018	296	113.23770	127.2377
8	+ fbs	1	8.674904	295	104.56279	120.5628
9	+ slope_2	1	7.296278	294	97.26651	115.2665
10	+ sex	1	7.354557	293	89.91196	109.9120
11	+ restecg_2	1	4.730566	292	85.18139	107.1814

FIGURE 16 – Méthode d'inclusion

```
formula
Y ~ sex + cp_1 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_2 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + fbs + restecg_2 + thalach + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + fbs + restecg_1 + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + trestbps + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_2 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + chol + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_1 + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_2 + cp_3 + fbs + restecg_1 + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_2 + cp_3 + fbs + restecg_2 + thalach + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_2 + cp_3 + trestbps + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_2 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_1 + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_3 + fbs + restecg_1 + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_2 + cp_3 + chol + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_1 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + trestbps + fbs + restecg_1 + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + fbs + restecg_1 + restecg_2 + thalach + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_3 + trestbps + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + fbs + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + trestbps + fbs + restecg_2 + thalach + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_3 + fbs + restecg_2 + thalach + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_3 + fbs + restecg_2 + exang + oldpeak + slope_1 + slope_2 + thal_1 + thal_2
Y ~ age + sex + cp_1 + cp_3 + chol + fbs + restecg_2 + exang + oldpeak + slope_2 + thal_1 + thal_2
Y ~ sex + cp_1 + cp_3 + chol + fbs + restecg_2 + thalach + exang + oldpeak + slope_2 + thal_1 + thal_2
```

FIGURE 17 – Tous les sous-modèles

Nous remarquons que les facteurs qui semblent associés à une hausse du risque d'un diagnostic positif de maladie coronarienne sont les variables : thal\_1, thal\_2, oldpeak, exang, cp\_1, cp\_3, fbs, slope\_2, sex, restecg\_2.

D'où le meilleur modèle est :

$$Y \sim \text{thal\_1} + \text{thal\_2} + \text{oldpeak} + \text{exang} + \text{cp\_1} + \text{cp\_3} + \text{fbs} + \text{slope\_2}, \text{sex} + \text{restecg\_2}$$

### 3 Exercice 3

L'objectif de cet exercice est de construire un modèle à partir du jeu de données `ausprivauto0405` disponible dans le package `R CASdatasets` (qui contient des données d'une compagnie d'assurance auto) afin de voir s'il y a une association entre les caractéristiques de véhicule et les caractéristiques des clients et le nombre de réclamations des clients et le nombre de réclamations (qui présente la variable réponse).

#### 3.1 Notations

Les notations adoptées pour l'ajustement du modèle sont :

- $X_1$  la variable explicative «Expoure» qui est la proportion de l'année pendant laquelle l'assuré(e) est couvert(e). C'est la variable offset.
- $X_2$  la variable explicative «VehValue» qui est la valeur relative du véhicule (mesure continue).
- $X_3$  la variable explicative «VehAge» qui est l'âge du véhicule sous la forme de variable qualitative à 4 modalités.
- $X_4$  la variable explicative «VehBody» qui est le type de véhicule sous la forme de variable qualitative à 13 modalités.
- $X_5$  la variable explicative «Gender» qui est le : sexe de l'assuré(e) sous la forme de variable qualitative à 2 modalités.
- $X_6$  la variable explicative «DrivAge» qui est l'âge de l'assuré(e) sous la forme de variable qualitative à 6 modalités.
- $\beta$  le vecteur de paramètres.
- $X$  la matrice des variables explicatives  $X_2, X_3, X_4, X_5$  et  $X_6$
- $Y$  la variable réponse «ClaimNb» qui est le nombre de réclamations faites par l'assuré pendant sa période pendant laquelle il est couvert par la compagnie d'assurance.

#### 3.2 Distribution de la variable réponse et fonction de lien

La variable réponse suit une loi de poisson. Soit  $Y_i|X_i$  suit une loi de poisson de moyenne  $\mu_i$  où  $Y_i$  est la valeur de la variable réponse de l'individu  $i$  et  $X_i$  est le vecteur contenant les valeurs des variables explicatives du même individu  $i$ . Nous avons choisi d'utiliser le lien log.

#### 3.3 Modèle adopté

Avant de choisir un modèle nous commençons par tester la présence de la multicolinéarité entre les variables explicatives.



	Variables	Tolerance	VIF
1	vehValue	0.411716794	2.428854
2	vehAgeoldest cars	0.622386690	1.606718
3	vehAgeyoung cars	0.654490850	1.527905
4	vehAgeyoungest cars	0.587715016	1.701505
5	vehBodyConvertible	0.364069873	2.746725
6	vehBodyCoupe	0.058608716	17.062309
7	vehBodyHardtop	0.030192263	33.121069
8	vehBodyHatchback	0.003502661	285.497215
9	vehBodyMinibus	0.063353288	15.784500
10	vehBodyMotorized caravan	0.274275691	3.645967
11	vehBodyPanel van	0.060655688	16.486500
12	vehBodyRoadster	0.638469240	1.566246
13	vehBodySedan	0.003200728	312.428899
14	vehBodyStation wagon	0.003866379	258.639934
15	vehBodyTruck	0.027394372	36.503848
16	vehBodyUtility	0.011105773	90.043263
17	genderMale	0.915099814	1.092777
18	drivAgeolder work. people	0.522449745	1.914060
19	drivAgeoldest people	0.681455059	1.467448
20	drivAgeworking people	0.522904659	1.912395
21	drivAgeyoung people	0.555427121	1.800416
22	drivAgeyoungest people	0.704405592	1.419637

FIGURE 18 – Calcul des VIFS

On trouve qu'il n'y a pas de multicolinéarité entre les variables explicatives. Pour l'ajoutement, nous commençons par un modèle de poisson ordinaire :

$$\ln(\mu_i) = \beta X_i$$

$$\mu_i = \exp(\beta X_i)$$

```

Call:
glm(formula = ClaimNb ~ VehValue + VehAge + VehBody + Gender +
    DriveAge, family = poisson(link = "log"), data = donnees_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7726  -0.3985  -0.3759  -0.3478   5.0230

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.8576419   0.3211834  -5.784 7.31e-09 ***
VehValue      0.0445342   0.0165866   2.685 0.007254 **
VehAgeoldest cars -0.0437363   0.0408224  -1.071 0.283998
VehAgeyoung cars  0.1002390   0.0396325   2.529 0.011432 *
VehAgeyoungest cars -0.0567281   0.0481184  -1.179 0.238427
VehBodyConvertible -2.0263830   0.6684146  -3.032 0.002432 **
VehBodyCoupe     -0.7644921   0.3369607  -2.269 0.023281 *
VehBodyHardtop   -0.8847524   0.3278269  -2.699 0.006958 **
VehBodyHatchback -1.0545383   0.3183277  -3.313 0.000924 ***
VehBodyMinibus   -1.1653113   0.3500375  -3.329 0.000871 ***
VehBodyMotorized caravan -0.5204386   0.4091001  -1.272 0.203318
VehBodyPanel van -0.8178548   0.3388334  -2.414 0.015790 *
VehBodyRoadster  -0.7609566   0.6596484  -1.154 0.248673
VehBodySedan     -1.0166085   0.3177285  -3.200 0.001376 **
VehBodyStation wagon -1.0252982   0.3179549  -3.225 0.001261 **
VehBodyTruck     -1.0460988   0.3283044  -3.186 0.001441 **
VehBodyUtility   -1.2463074   0.3220488  -3.870 0.000109 ***
GenderMale       -0.0109144   0.0300771  -0.363 0.716693
DriveAgeolder work. people 0.1974219   0.0489145   4.036 5.44e-05 ***
DriveAgeoldest people -0.0004901   0.0643249  -0.008 0.993921
DriveAgeworking people  0.2238705   0.0490393   4.565 4.99e-06 ***
DriveAgeyoung people  0.2579782   0.0506895   5.089 3.59e-07 ***
DriveAgeyoungest people  0.4239130   0.0590581   7.178 7.08e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26768  on 67855  degrees of freedom
Residual deviance: 26617  on 67833  degrees of freedom
AIC: 36098

Number of Fisher Scoring iterations: 6

```

FIGURE 19 – Ajustement du modèle avec poisson ordinaire

Nous ajoutons le log de la variable offset  $X_1$  :

$$\ln(\mu_i) = \beta X_i + \ln(X_{1i})$$

```

Car...
glm(formula = ClaimNb ~ vehValue + vehAge + vehBody + Gender +
    DrivAge + offset(log.offset), family = poisson(link = "log"),
    data = donnees_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9082  -0.4524  -0.3462  -0.2213   4.5123

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.19827    0.32076   -3.736 0.000187 ***
vehValue         0.02390    0.01720    1.390 0.164623
vehAgeoldest cars -0.05933    0.04108   -1.444 0.148706
vehAgeyoung cars   0.11145    0.03967    2.810 0.004958 **
vehAgeyoungest cars 0.05550    0.04819    1.152 0.249471
vehBodyconvertible -1.67029    0.66784   -2.501 0.012383 *
vehBodycoupe      -0.51094    0.33695   -1.516 0.129432
vehBodyHardtop    -0.83353    0.32785   -2.542 0.011009 *
vehBodyhatchback  -0.97543    0.31821   -3.065 0.002174 **
vehBodyminibus    -0.98426    0.35005   -2.812 0.004927 **
vehBodyMotorized caravan -0.38795    0.40942   -0.948 0.343344
vehBodyPanel van  -0.85291    0.33883   -2.517 0.011829 *
vehBodyRoadster   -0.56485    0.65987   -0.856 0.391995
vehBodySedan      -0.92396    0.31764   -2.909 0.003628 **
vehBodyStation wagon -0.91190    0.31806   -2.867 0.004143 **
vehBodyTruck      -0.96325    0.32836   -2.933 0.003352 **
vehBodyutility    -1.12006    0.32208   -3.478 0.000506 ***
GenderMale        -0.02289    0.03009   -0.761 0.446929
DrivAgeolder work. people 0.21907    0.04891    4.479 7.50e-06 ***
DrivAgeoldest people  0.01580    0.06433    0.246 0.806052
DrivAgeyoung people  0.24752    0.04904    5.048 4.47e-07 ***
DrivAgeyoungest people 0.30747    0.05068    6.067 1.30e-09 ***
DrivAgeyoungest people 0.47769    0.05904    8.091 5.93e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 25507  on 67855  degrees of freedom
Residual deviance: 25343  on 67833  degrees of freedom
AIC: 34824

Number of Fisher Scoring iterations: 6

```

FIGURE 20 – Ajustement du modèle avec poisson ordinaire en ajoutant une variable offset

Afin de corriger la surdispersion, nous avons ajusté le modèle par deux méthodes :

En ajustant un modèle quasi-poisson :

```

Call:
glm(formula = ClaimNb ~ vehvalue + vehAge + vehBody + Gender +
  DriveAge + offset(log.offsets), family = quasipoisson(link = "log"),
  data = donnees_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9082  -0.4524  -0.3462  -0.2213   4.5123

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.19827    0.38062   -3.148  0.00164 **
vehvalue         0.02390    0.02041    1.171  0.24156
vehAgeoldest cars  -0.05933    0.04875   -1.217  0.22361
vehAgeyoung cars    0.11145    0.04707    2.368  0.01790 *
vehAgeyoungest cars  0.05550    0.05719    0.971  0.33180
vehBodyconvertible -1.67029    0.79249   -2.108  0.03506 *
vehBodyCoupe      -0.51094    0.39984   -1.278  0.20131
vehBodyHardtop    -0.83353    0.38904   -2.143  0.03215 *
vehBodyHatchback  -0.97543    0.37760   -2.583  0.00979 **
vehBodyMinibus    -0.98426    0.41538   -2.370  0.01781 *
vehBodyMotorized caravan -0.38795    0.48583   -0.799  0.42456
vehBodyPanel van  -0.85291    0.40207   -2.121  0.03390 *
vehBodyRoadster   -0.56485    0.78303   -0.721  0.47068
vehBodySedan      -0.92396    0.37693   -2.451  0.01424 *
vehBodystation wagon -0.91190    0.37743   -2.416  0.01569 *
vehBodyTruck      -0.96325    0.38964   -2.472  0.01343 *
vehBodyutility    -1.12006    0.38220   -2.931  0.00338 **
genderMale        -0.02289    0.03571   -0.641  0.52157
driveAgeolder work. people 0.21907    0.05804    3.774  0.00016 ***
driveAgeoldest people  0.01580    0.07634    0.207  0.83608
driveAgeworking people  0.24752    0.05819    4.254  2.10e-05 ***
driveAgeyoung people   0.30747    0.06014    5.113  3.18e-07 ***
driveAgeyoungest people  0.47769    0.07006    6.818  9.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.408105)

Null deviance: 25507  on 67855  degrees of freedom
Residual deviance: 25343  on 67833  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6

```

FIGURE 21 – Ajustement du modèle avec quasi-poisson

En ajustant une loi binomiale négative :

```

Call:
glm.nb(formula = claimNb ~ vehvalue + vehAge + vehBody + Gender +
  DriveAge + offset(log.offsets), data = donnees_1, link = "log",
  init.theta = 2.259193458)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8702  -0.4478  -0.3443  -0.2209   4.0965

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.20475    0.33898   -3.554 0.000379 ***
vehvalue         0.02510    0.01761    1.425 0.154129
vehAgeoldest cars -0.05695    0.04191   -1.359 0.174202
vehAgeyoung cars  0.11057    0.04056    2.726 0.006410 **
vehAgeyoungest cars 0.05229    0.04926    1.062 0.288415
vehBodyConvertible -1.66779    0.68212   -2.445 0.014485 *
vehBodyCoupe     -0.50418    0.35526   -1.419 0.155852
vehBodyHardtop   -0.82954    0.34608   -2.397 0.016532 *
vehBodyHatchback -0.96779    0.33647   -2.876 0.004024 **
vehBodyMinibus   -0.98261    0.36790   -2.671 0.007566 **
vehBodyMotorized caravan -0.38191    0.42929   -0.890 0.373659
vehBodyPanel van -0.85100    0.35711   -2.383 0.017173 *
vehBodyRoadster  -0.57352    0.68854   -0.833 0.404874
vehBodySedan     -0.91782    0.33591   -2.732 0.006289 **
vehBodyStation wagon -0.90735    0.33633   -2.698 0.006980 **
vehBodyTruck     -0.96101    0.34654   -2.773 0.005552 **
vehBodyutility   -1.11621    0.34026   -3.280 0.001036 **
GenderMale       -0.02269    0.03074   -0.738 0.460470
DriveAgeolder work. people 0.21975    0.04983    4.410 1.04e-05 ***
DriveAgeoldest people  0.01496    0.06548    0.228 0.819310
DriveAgeworking people  0.24843    0.04997    4.971 6.65e-07 ***
DriveAgeyoung people   0.30730    0.05169    5.945 2.76e-09 ***
DriveAgeyoungest people 0.48138    0.06035    7.977 1.50e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2.2592) family taken to be 1)

    Null deviance: 23577  on 67855  degrees of freedom
Residual deviance: 23420  on 67833  degrees of freedom
AIC: 34786

Number of Fisher Scoring iterations: 1

```

FIGURE 22 – Ajustement du modèle avec une loi binomiale négative

Afin de choisir le meilleur modèle nous avons effectué le Test de vraisemblance :

$H_0 : Y_i$  suit une loi de poisson VS  $H_1 : Y_i$  suit une loi binomiale négative

La log vraisemblance maximisée sous  $H_0$  est :

$$l_0 = -17388.76$$

La log vraisemblance maximisée sous  $H_1$  est

$$l_1 = -17369.15$$

La statistique du test est :

$$\xi = 2(l_1 - l_0) = 19.60403$$

Alors :

$$P\_value = 0.5P(\chi_1^2 > \xi)$$

```
test= 2*(logLik(m1.binneg)-logLik(m1.pois.ofs))
p_value <- 0.5*pchisq(test, df=1, lower.tail = FALSE)
p_value
log Lik.' 1.904828e-10 (df=24)
```

FIGURE 23 – Test d’hypothèse

D’après la figure 23 nous remarquons que  $P\_value = 1.904828e - 10 < 0.5$  alors nous rejettons l’hypothèse  $H_0$ . Nous pouvons dire que le modèle final est le modèle ajusté par une loi binomiale négative.

Choisissons alors le meilleur modèle par les trois méthodes de sélection.

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
ClaimNb ~ 1

Final Model:
ClaimNb ~ DrivAge + VehAge + VehBody + VehValue
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				67855	23362.91	36101.36
2	+ DrivAge	5	11.896577	67850	23374.81	36043.60
3	+ VehAge	3	4.923112	67847	23379.73	36024.46
4	+ VehBody	12	13.958246	67835	23393.69	36005.49
5	+ VehValue	1	4.101449	67834	23389.59	36000.95

FIGURE 24 – Méthode de sélection 1

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
ClaimNb ~ vehValue + VehAge + VehBody + Gender + DrivAge + offset(log,ofs)

Final Model:
ClaimNb ~ VehAge + VehBody + DrivAge + offset(log,ofs)
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				67833	23419.71	34784.31
2	- Gender	1	0.3878525	67834	23419.32	34782.85
3	- VehValue	1	3.6303710	67835	23422.95	34782.67

FIGURE 25 – Méthode de sélection 2

```

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
ClaimNb ~ 1

Final Model:
ClaimNb ~ DriveAge + VehAge + VehBody + VehValue

      Step Df  Deviance Resid. Df Resid. Dev      AIC
1              5 11.896577    67855   23362.91 36101.36
2 + DriveAge   3  4.923112    67847   23379.73 36024.46
3 + VehAge     1  4.101449    67834   23389.59 36000.95
4 + VehBody    1  4.101449    67834   23389.59 36000.95
5 + VehValue   1  4.101449    67834   23389.59 36000.95

```

FIGURE 26 – Méthode de sélection 3

Le modèle avec la plus petite valeur de l'AIC est notre modèle finale. C'est le modèle choisi par la deuxième méthode où les variables explicatives sont :  $\log(X_1), X_3, X_4$  et  $X_6$

La variable  $X_2$  n'est pas significative au seuil 5% car sa p\_value du test de  $\chi^2_1$  est 0.154129 alors la valeur relative du véhicule n'est pas associée au nombre espéré de réclamations.

## 4 Annexe

### Variables du fichier processed.cleveland.data

Les variables du fichier sont entrées dans le même ordre que celui qui suit.

- age : âge en années
- sex : 1 = homme, 0 = femme
- cp : nature des douleurs à la poitrine, variable qualitative à 4 modalités, où 1 dénote l'angine typique, 2 l'angine atypique, 3 une douleur non anginienne et 4 une douleur asymptomatique
- trestbps : tension artérielle au repos (en mm Hg) à l'admission à l'hôpital
- chol : cholestérol sanguin en mg/dl
- fbs : indicatrice qui vaut 1 si le taux de sucre sanguin à jeun > 120 mg/dl et qui vaut 0 sinon
- restecg : résultat de l'électrocardiogramme au repos, variable qualitative à 3 modalités, où 0 signifie normal, 1 signifie anomalie des ondes ST-T et 2 signifie hypertrophie probable du ventricule gauche
- thalach : pouls maximum atteint
- exang : indicatrice indiquant la présence d'angine induite par l'exercice (1 pour oui, 0 pour non)
- oldpeak : baisse dans ST induite par l'exercice par rapport au repos
- slope : pente du segment de ST lors de l'exercice maximal, variable qualitative à 3 modalités soit 1 pour ascendante, 2 pour plate et 3 pour descendante
- ca : nombre de vaisseaux sanguins majeurs colorés par fluoro-scopie
- thal : variable qualitative à 3 modalités où 3 = normal, 6 = défaut réparé, 7 = défaut réparable
- num : la variable réponse que nous cherchons à prédire est  $Y = 1$  si num > 0

et  $Y = 0$  si  $\text{num} = 0$

**Variables du jeu de données CASdatasets\_1.0-6.tar.gz du package CASdatasets**

- Exposure : proportion de l'année pendant laquelle l'assuré(e) est couvert(e)
- VehValue : valeur relative du véhicule (mesure continue)
- VehAge : âge du véhicule sous la forme de variable qualitative à 4 modalités
- VehBody : type de véhicule sous la forme de variable qualitative à 13 modalités
- Gender : sexe de l'assuré(e) sous la forme de variable qualitative à 2 modalités
- DrivAge : âge de l'assuré(e) sous la forme de variable qualitative à 6 modalités
- ClaimNb : nombre de réclamations, variable réponse