Table des matières

Liste des tableaux

1 Introduction

Le but de ce projet est d'appliquer les méthodes statistiques abordées dans le cours du modèle de biostatistique sur un cas concret.

Dans un premier temps, nous allons pré-traiter le jeu de données *(recodage des variables, choix de la méthode d'imputation des données manquantes)* et faire des analyses descriptives afin de collecter le maximum d'informations sur les possibles relations entre les variables.

Dans un second temps, nous allons élaborer des régressions logistiques dans le but de soumettre des hypothèses.

Notre table ne comporte ni des covariables qui varient en fonction du temps, ni des données appariées, alors on ne pourra pas faire une analyse de survie.

2 Présentation de la table

Le jeu de données à été récolté en **1990** en **Inde** après une grande manifestation. A la suite de cas d'intoxications alimentaires constatées, l'enquête a été menée sur un grand échantillon d'individus (*1094 individus*) ayant participé à cette manifestation

2.1 Statistiques descriptives

Le jeu de données contient **4** variables explivatives qualitatives. Pour les variables du type alimentaire, l'individu devait répondre à la question suivante : **Avez-vous manger cet aliment lors de l'événement ?**.

La table suivante présente un résumé de ces variables :

Table 1 – Résumé des variables qualitatives

Sexe	Boeuf	Oeuf	Eau
Femme :373	**Non : 91**	**Non : 87**	**Non : 25**
Homme :721	**Oui :998**	**Oui :1002**	**Oui :1063**
NA	**NA's : 5**	**NA's : 5**	**NA's : 6**

Le jeu de données contient en plus, **2** variables explivatives quantitatives. L'**âge** *(en années)* et l'**éclairs** *(Nombre d'éclairs mangés par chanque individu)*

Table 2 – Résumé des variables quantitatives

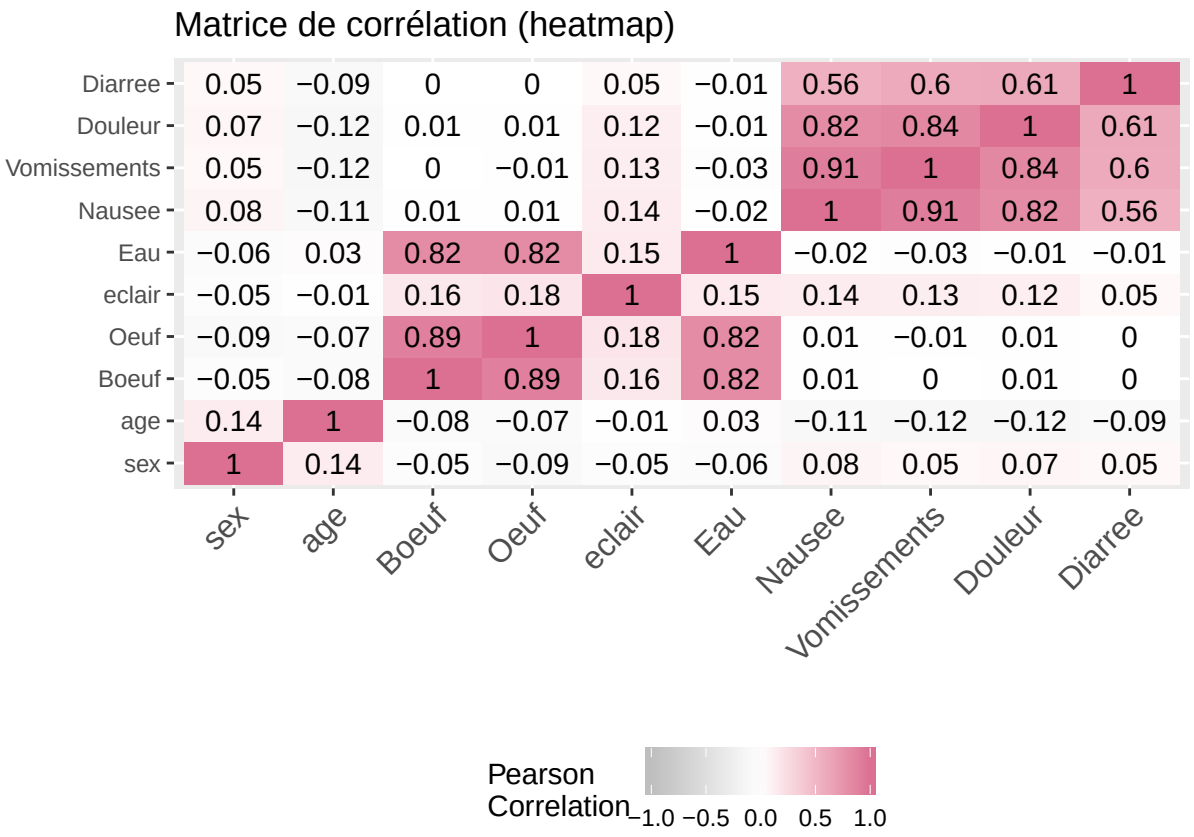
variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
age	**1.00**	**14.00**	**17.00**	**19.32**	**22.00**	**58.00**	**60**
eclair	**0.000**	**0.000**	**2.000**	**1.673**	**2.000**	**20.000**	**117**

La table ci-dessous présente un résumé des variables contenant l'information sur les symptomes chez les individus.

Table 3 – Résumé des variables de symptôme

Nausée	Vomissements	Douleur	Diarrée
Non :658	**Non :679**	**Non :711**	**Non :859**
oui :436	**Oui :415**	**Oui :383**	**Oui :235**

2.1.1 Corrélation



Nous observons dans le graphe ci-dessus que les variables **Eau**, **Boeuf** et **Oeuf** sont très corrélées (*positivement*) entre elles, aussi pour les **variables de symptôme**. Cependant nous constatons que les variables de symptôme ont une faible corrélation avec le reste des variables.

2.2 Données manquantes

Une donnée incomplète est une donnée pour laquelle la valeur de certain attribut est inconnue, ces valeurs sont dites manquantes.

Les valeurs manquantes peuvent être de deux types :

— Valeur manquante totale [1]
— Valeur manquante partielle [2]

En général, des valeurs manquent dans un jeu de données parce qu'**elles n'ont pas pu être observées**, **elles ont été perdues** ou **elles n'étaient pas enregistrées**.

2.2.1 Mécanisme des valeurs manquantes

Avant d'entamer le traitement des données manquantes, nous allons tout d'abord évaluer le mécanisme de ces dernières.

Selon **Little, R.J.A.**, and **Rubin, D.B.** (2002), il y a trois mécanismes distincts de valeurs manquantes :

— **MCAR** [3] : le fait de ne pas avoir la valeur pour une variable, est indépendant des autres variables.

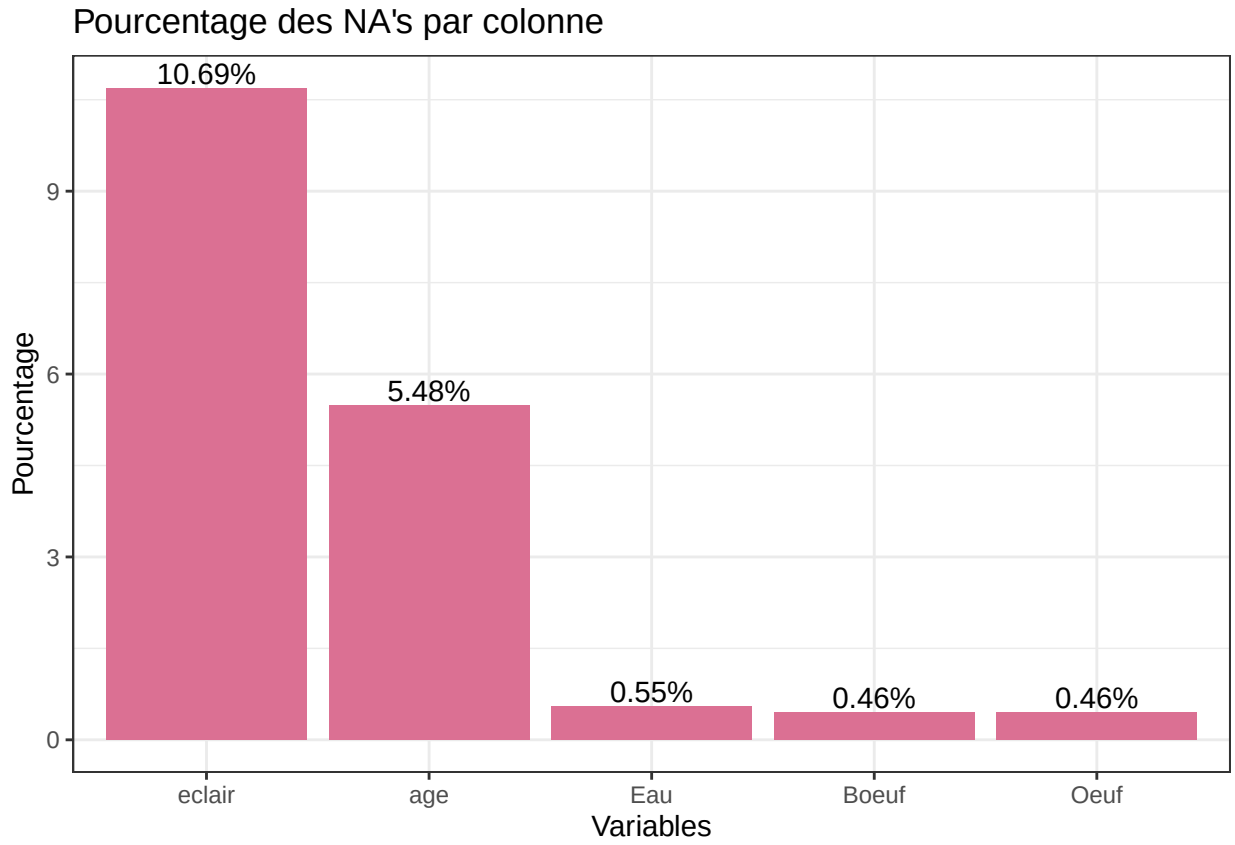
1. C'est-à-dire que toute l'observation manque
2. C'est-à-dire que l'observation est présente mais il manque certaines valeurs de cette observation
3. Missing completly at random

— **MAR**[4] : le fait de ne pas avoir la valeur pour une variable, est dépendant seulement des valeurs observées.
— **MNAR**[5] : le fait de ne pas avoir la valeur pour une variable ne dépendant que des valeurs manquantes.

Selon **Simon et Smonoff** (1986) et **Little** (1988), il est difficile de traiter des données de type MNAR et MAR, ce qui nous incite à faire l'hypothèse que le manque de données est de nature **MCAR**, dans la pratique, et comme l'expliquent **Schafer et Graham** (2002), il est quasiment impossible de déterminer lequel des trois mécanismes est à l'œuvre à partir des données.

2.2.2 Comportement des données manquantes

Comme indiqué dans la partie descriptive, certaines variables contiennent des valeurs manquantes. Le graphe ci-dessous nous montre le pourcentage des valeures manquantes pour chaque variable.



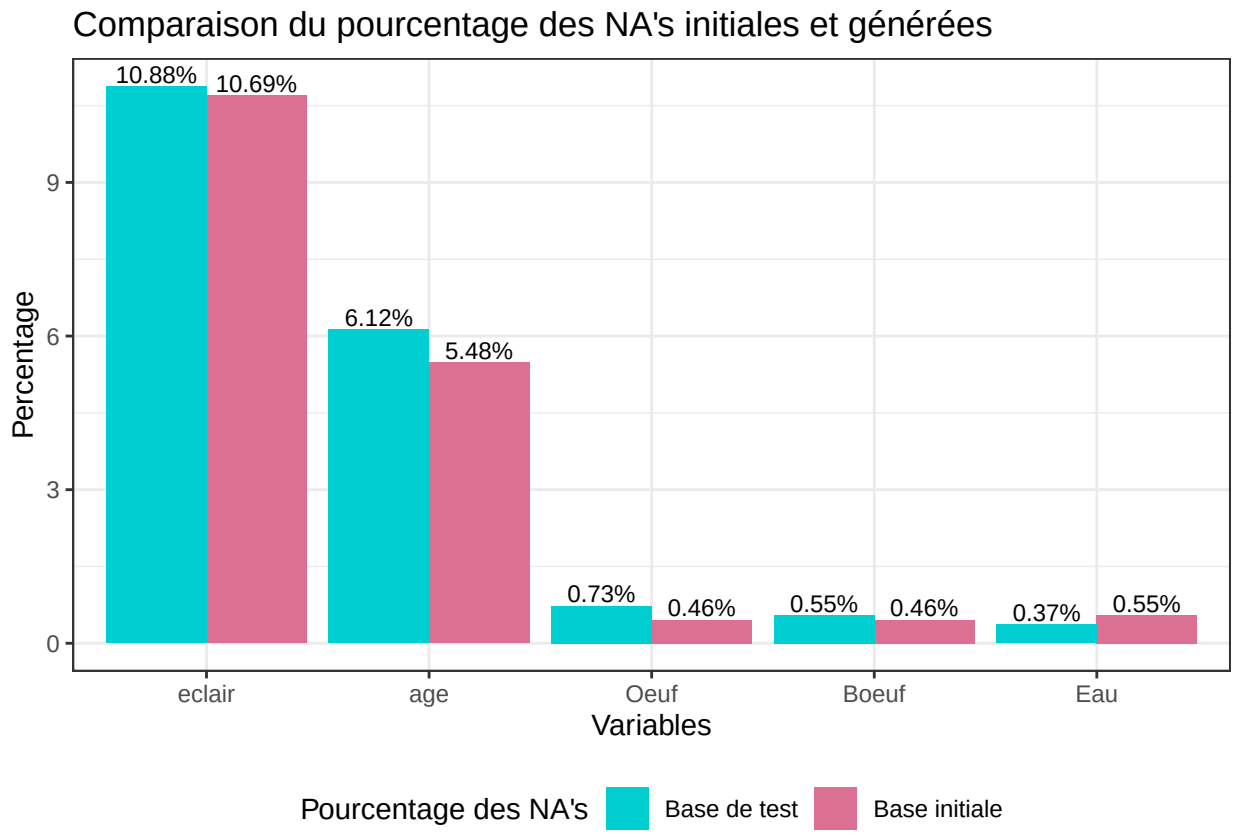
Rappelons que les valeurs qui faisaient référence aux données manquantes dans la variable **Eclair** sont **80** et **90** correspondent respectivement à *l'individu a mangé des éclairs sans se souvenir combien* et *données manquantes*. Cela veut dire que si nous remplaçons les valeurs **80** de cette variable par des **NA's**, à l'imputation, ces dernières peuvent être remplacées par la valeur **0**, ce qui est impossible. Pour cela nous avons choisi de les imputées par la **médianne**.

Pour les autres variables*(age, Eau, . . .)*, nous n'avons pas ce problème, Donc, nous avons décidé de les imputées par différentes méthodes d'*imputation* afin de choisir la méthode la plus adaptée à notre jeu de données.

Comment peut-on faire ce choix ?

4. Missing at random
5. Missing Not at Random

Nous allons commencer par imputer toutes les **NA's** par une méthode d'imputation multiple (*EX : KNN*), ensuite, nous allons considérer **la table imputée** comme notre table de test (*table complète*), nous allons nous servir de la fonction **ampute()** pré-définie sur **R** pour générer des valeurs manquantes dans notre table de test, cette fonction nous permet de choisir le pourcentage des NA's que nous souhaitons générer aussi le mécanisme (**MCAR**, **MAR**, ou bien **MNAR**) et finalement, nous allons les imputées par différentes méthodes et choisir la plus adaptée.



Interprétation :

Le graphe ci-dessus est une comparaison du pourcentage des valeurs manquantes par colonne dans notre jeu de données initial et la base de test (imputée par KNN) après avoir générer des NA's avec les mêmes pourcentages et avec le mécanisme **MCAR**.

Nous observons dans ce graphe que les pourcentages obtenus dans les deux jeux de données sont très proche.

Conséquence :

Ce résultat est très robuste, car il va nous permettre d'appliquer différentes méthodes d'imputation des NA's sur cette base (base de test) dont on connaît les vraies valeurs et qui va rendre possible le calcul de la précision de chaque méthode utilisée.

2.2.3 Choix de la méthode d'imputation

L'imputation de données manquante réfère au fait qu'on remplace les valeurs manquantes dans le jeu de données par des valeurs artificielles. Pour imputer ces valeurs manquantes, nous disposons de plusieurs méthodes d'imputation. Parmi ces dernières, nous allons tester les méthodes suivantes :

— Médianne/Mode

— rf [6]
— pmm [7]
— cart [8]
— KNN [9]

Pour l'imputation par la Médianne/Mode, nous allons remplacer les données manquantes dans les variables quantitatives par la médiane et dans les variables qualitatives par le mode [10].

Pour les méthodes **rf**, **pmm** et **cart**, nous allons se servir de la fonction **mice()** prédéfinie dans **R**. Cette dernière nous donne la possibilité de choisir la méthode d'imputation ainsi que le nombre maximal d'itérations.

Pour la méthode **KNN**, nous allons utiliser la fonction **kNN()**.

TABLE 4 – Précision des méthodes pour l'imputation des variables qualitatives

Variables	rf	knn	pmm	cart	mod_med
Boeuf	**67%**	83%	**83.0%**	83%	83%
Oeuf	**100%**	100%	**88.0%**	100%	100%
Eau	**100%**	100%	**100.0%**	100%	100%

Le tableau ci-dessus nous montre la précision des méthodes d'imputation utilisées pour les variables qualitatives.

Nous observons que les méthodes **KNN**, **cart** et **Médiane/Mode** sont les plus précises. elles ont pu imputer correctement **100% des valeurs manquantes** pour les variables Oeuf (*6 NA's*) et Eau (*5 NA's*) et **83%** pour la variable Boeuf (*5 NA's*)

TABLE 5 – Précision des méthodes pour l'imputation des variables quantitatives

Variables	rf	knn	pmm	cart	Med_mod
age	**6.604205**	**5.2623400**	**4.5484461**	3.6051188	**3.980804**
eclair	**0.536106**	**0.3105576**	**0.4113346**	0.3724863	**0.261426**

Le tableau ci-dessus nous donne l'erreur quadratique moyenne (*MSE*) obtenue entre les valeurs de la base de test et les valeurs imputées par chaque méthode pour les deux variables quantitatives.

Nous déduisons que la méthode **cart** est la plus précise en moyenne.

Nous avons décidé d'imputer par la méthode **cart**, car elle donne l'erreur la plus faible pour les variables quantitatives et une précision pareille que les méthodes **KNN** et **Médianne/Mode** pour les variables qualitatives.

3 Régression logistique avec Y dichotomique

Après avoir imputer les données manquantes, nous allons maintenant créer une nouvelle variable **Malade** qui sera **Dichotomique**.

3.1 Variable dichotomique :

Définition : Une variable dichotomique est une variable qualitative qui ne peut prendre que 2 modalités : OUI ou NON ; masculin ou féminin ; malade ou sain , etc. . . .

6. Random Forest imputations
7. predictive mean matching
8. Classification and regression trees
9. k-nearest neighbors
10. Mode : La modalité la plus présente.

Pour créer la variable **dichotomique**, nous allons regarder si l'individu n'avait aucun symptôme alors il n'est pas malade (0), sinon il est malade (1).

Le tablteau ci-dessous nous présente un résumé de cette nouvelle variable.

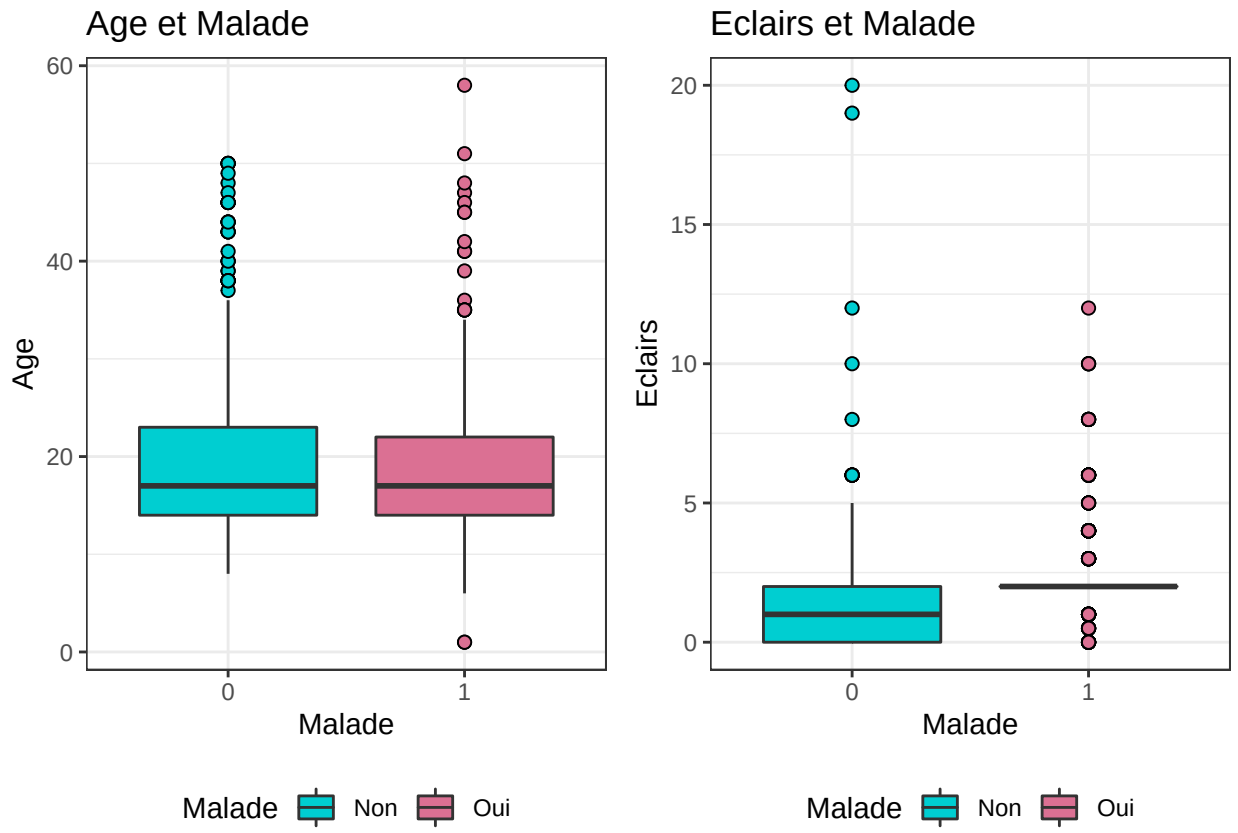
TABLE 6 – Résumé de la variable malade (dichotomique)

Malade	Nombre d'individus
Oui	**469**
Non	**625**

Nous observons que parmi les 1094 individus, **469** *(42%)* individus sont malade (*Représente au moins un symptôme*) et **625** *(57%)* ne le sont pas.

Statistiques descriptives :

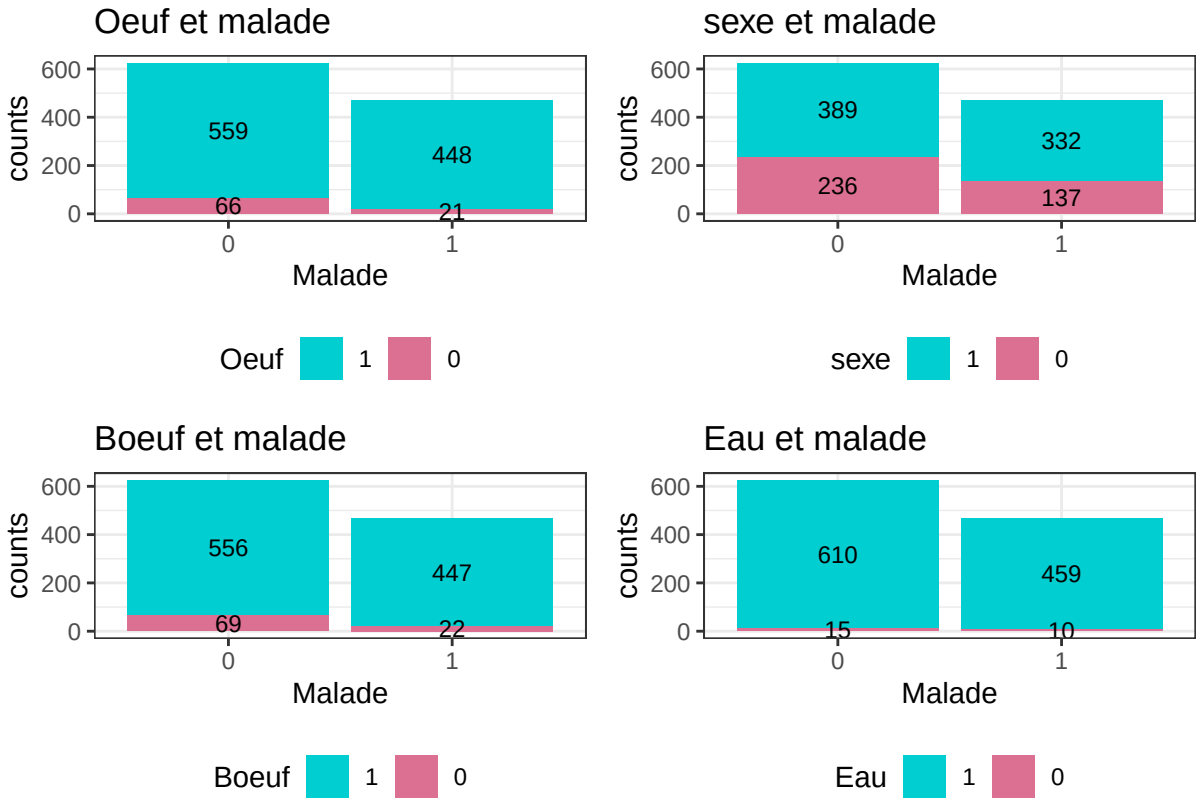
Le graphe ci-dessous est une représentation des variables explcatives quantitatives (*Age, Eclairs*) en fonction de la variable dichotomique **Malade** :



Pour l'age, nous abservons que les médianes sont au même niveau. Cela veut dire que l'age de l'individu n'a pas d'effet sur le fait que celui-ci tombe malade.

Pour les éclairs, nous constatons que la médiane pour les personnes malades (= *2*) est en dessus de celui des personnes non-malades (= *1*). Nous observons aussi des individus toxiqués sans avoir mangé d'éclairs et deux individus non-malades ont mangé 19 et 20 éclairs.

le graphe ci-dessous représente les variables explicatives qualitatives en fonction de la variable malade.



D'après les trois premiers graphes, nous constatons qu'on a très peu de personnes malades après avoir bu de l'eau ou mangé du boeuf ou des oeufs. Cependant nous observons dans le dernier graphe que les femmes ont plus de chance de tomber malade par rapport aux hommes.

3.2 Régression logistique

Nous allons maintenant mener une première étude, en prenant la variable malade (**0 = non-malade, 1 = malade**). Comme celui-ci est dichotomique, nous allons effectuer une **régression logistique** en prenant en compte les variables explicatives (*sexe* et *eclair*) et les intéractions entre les variables (*Eau, age, Boeuf, Oeuf*).

TABLE 7: Premier modèle logistique pour la variable dichotomique

Variables	Coefficients	Erreur standard	Statistique	Valeur p
(Intercept)	**1.1465952**	**1.3316409**	**0.8610393**	**0.3892164**
sex1	0.3708490	0.1446867	2.5631167	0.0103737
eclair	0.6174017	0.0588467	10.4917033	0.0000000
age	-0.1482911	0.0512224	-2.8950425	0.0037911
Boeuf1	**-18.0835119**	**6995.8148589**	**-0.0025849**	**0.9979375**
Oeuf1	**183.3884771**	**6439.7196091**	**0.0284777**	**0.9772811**
Eau1	-3.3336657	1.9221761	-1.7343186	0.0828616
age :Boeuf1	**0.1482911**	**282.7890614**	**0.0005244**	**0.9995816**
age :Oeuf1	**-7.2274750**	**249.8256718**	**-0.0289301**	**0.9769204**
Boeuf1 :Oeuf1	**-163.8036541**	**9508.4921819**	**-0.0172271**	**0.9862555**
age :Eau1	**0.1201493**	**0.0811339**	**1.4808770**	**0.1386394**
Boeuf1 :Eau1	**20.5407196**	**6995.8153214**	**0.0029361**	**0.9976573**

TABLE 7: Premier modèle logistique pour la variable dichotomique
(continued)

Variables	Coefficients	Erreur standard	Statistique	Valeur p
Oeuf1 :Eau1	**-182.7084722**	**6439.7197607**	**-0.0283721**	**0.9773654**
age :Boeuf1 :Oeuf1	**6.9811528**	**377.3361119**	**0.0185012**	**0.9852391**
age :Boeuf1 :Eau1	**-0.2609393**	**282.7891071**	**-0.0009227**	**0.9992638**
age :Oeuf1 :Eau1	**7.2433875**	**249.8256824**	**0.0289938**	**0.9768696**
Boeuf1 :Oeuf1 :Eau1	**161.8689574**	**9508.4925412**	**0.0170236**	**0.9864178**
age :Boeuf1 :Oeuf1 :Eau1	**-6.8825066**	**377.3361481**	**-0.0182397**	**0.9854476**

Seules les variables **sexe**, **Eclairs**, **Age** sont significatives au risques de se tromper de **5%**. Nous observons aussi que la variable **Eau** est significative au risque de se tromper de **9%**.

Définition AIC(*(Akaike Information Criterion)* :

$$AIC = -2\ln(V) + 2k$$

Où :

-k est le nombre de paramètres

-2k représente la pénalité

-V est la vraisemblance.

On va maintenant utiliser une méthode de séléction de variables pour nous permettre d'avoir un meilleur modèle avec seulement les variables suceptibles d'expliquer notre variable d'intérêt **Malade**. Nous allons prendre la méthode qui permet de trouver le meilleur modèle en minimisant le critère **AIC**. Le résultat obtenu est le suivant :

TABLE 8 – Modèle logistique choisi par le critère AIC pour la variable dichotomique

Variables	**Coefficients**	**Erreur standard**	**Statistique**	**Valeur p**
(Intercept)	**0.9421477**	**1.0938952**	**0.8612779**	**0.3890850**
sex1	**0.3734300**	**0.1424069**	**2.6222745**	**0.0087345**
eclair	**0.6371943**	**0.0585093**	**10.8904795**	**0.0000000**
age	**-0.1772541**	**0.0444604**	**-3.9867862**	**0.0000670**
Oeuf1	**0.5529342**	**0.2996125**	**1.8454978**	**0.0649652**
Eau1	**-2.5173437**	**1.1190521**	**-2.2495322**	**0.0244787**
age :Eau1	**0.1504192**	**0.0452147**	**3.3267727**	**0.0008786**

Nous constatons qu'avec ce nouveau modèle, les variables **Eclairs**, **Sexe**, **Age**, **Eau** et l'intéraction **Age :Eau** sont significatives au risques de se tromper de **5%**. Nous observons aussi que la variable **Oeuf** est significative au risque de se tromper de **8%**.

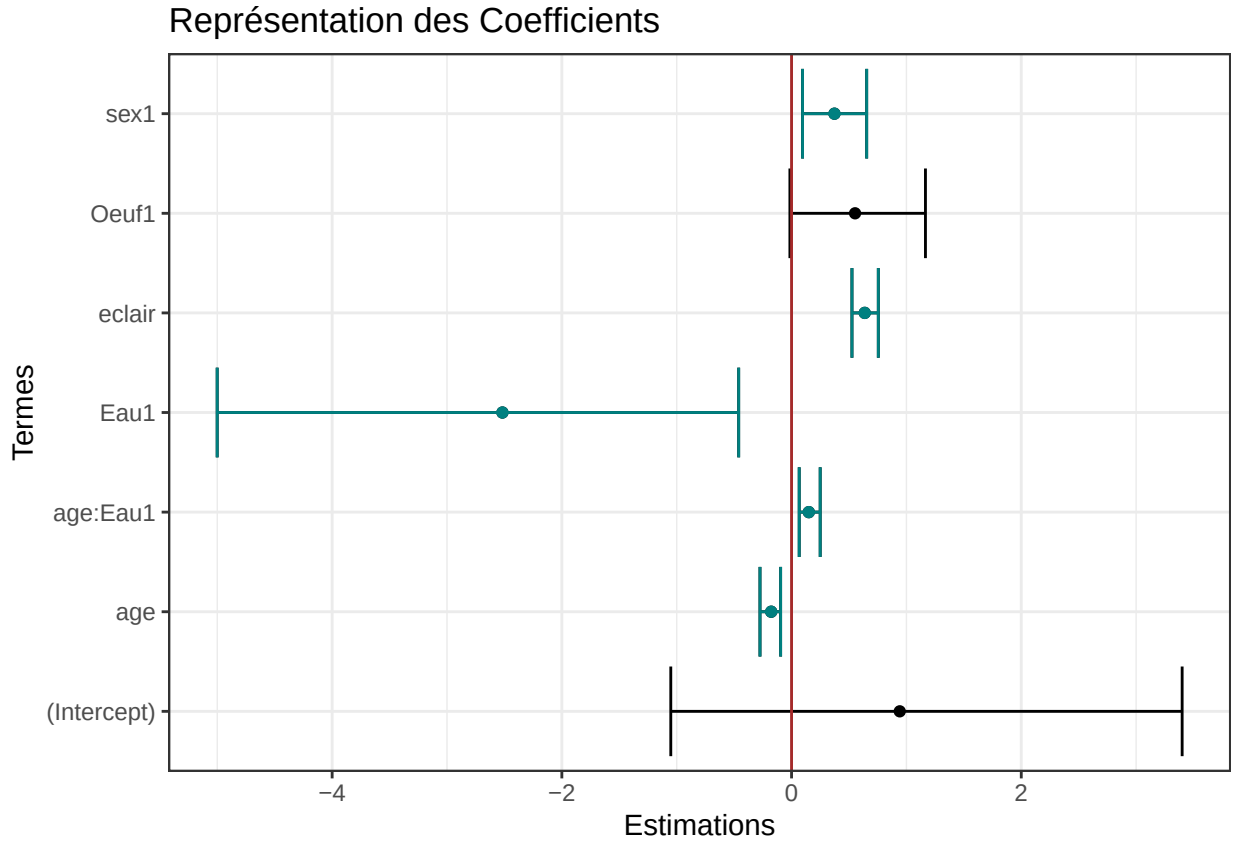
Intéressons nous maintenant aux coefficients et leurs intervalles de confiance.

TABLE 9 – Coefficients et leurs intervalles de confiance pour la variable dichotomique

Variables	Coefficients	2.5%	97.5%
(Intercept)	**0.9421477**	**-1.0515885**	**3.4006440**
sex1	**0.3734300**	**0.0955459**	**0.6541131**
eclair	**0.6371943**	**0.5257943**	**0.7552467**
age	**-0.1772541**	**-0.2749331**	**-0.0952238**
Oeuf1	**0.5529342**	**-0.0145045**	**1.1657906**
Eau1	**-2.5173437**	**-5.0007875**	**-0.4603600**
age :Eau1	**0.1504192**	**0.0667349**	**0.2492459**

Le graphe ci-dessus nous donne les coefficients obtenus pour chaque variable, intéraction et leurs intervalles de confiance.

Pour nous permettre de repérer les coefficients significatifs, on représente sous forme de graphique les résultats obtenus précédemment :



Nous pourrons rien dire sur les coefficients avec un intervalle de confiance contenant la valeur **0**, autrement dit, tous les segments croisant la droite **rouge**.

L'odds de tomber malade losrque la personne est un homme sera multiplier par au moins **exp(0.0955459) = 1.1003** et par au plus **1.9234**.

L'odds de tomber malade losrque la personne a bu de l'eau sera multiplier par au moins **0.0067** et par au plus **0.6311**.

En augmentant l'âge de la personne par une année, l'odds d'être malade sera multiplier par au moins **0.7596** et par au plus **0.9092**.

En augmentant le nombre d'éclairs mangées par une eclair, l'odds d'être malade sera multiplier par au moins **1.6918** et par au plus **2.1281**.

Pour l'intéraction entre **age** et **Eau1**, on estime :

$$\beta = log(\frac{odds(Y = 1|Age = x + 1, Eau = 1)}{odds(Y = 1|Age = x, Eau = 1)}) - log(\frac{odds(Y = 1|Age = x + 1, Eau = 0)}{odds(Y = 1|Age = x, Eau = 0)})$$

Nous nous intéressons à la différence entre l'effet de l'age sur la maladie parmi les buveurs et non-buveurs d'eau.

Nous constatons que l'effet de l'age chez les buvreurs d'eau est plus élevé que l'effet chez les non-buveurs. Parmis les buveaurs d'eau, l'effet de l'augmentation de l'âge d'une année va être multiplié par au moins **1.0690** et par au plus **1.283058**.

3.3 Résumé

La variable dichotomique (**Malade**) est expliquée par l'**âge**, le **sexe**, le nombre d'**éclairs** mangés et le fait de boire ou non de l'**eau**. Elle est aussi expliqué par l'intéraction entre l'**âge** et le fait de boire de l'**eau** ou pas.

L'odds ratio de l'âge est inférieur à 1. Cela veut dire qu'un jeune a plus de risque de tomber malade que les vieux.

Le fait d'être buveur d'eau est un facteur protecteur, un buveur d'eau diminue l'odds de tomber malade.

l'odds ratio de l'éclair est supérieur à 1. Cela signifie que plus on mange des éclairs plus on risque de tomber malade.

Le fait d'être homme est un facteur de risque, un homme a plus de chance de tomber malade qu'une femme.

Les non-buveurs d'eau parmi les personnes agés ont plus de risque de tomber malade que les buveurs d'eau parmi les personnes ayant le même âge.

4 Régression logistique avec Y polytomique ordonnée

Après avoir étudier le cas d'une variable **dichotomique**, nous allons maintenant créer une nouvelle variable avec un niveau de symptomes dite **Polytomique**.

4.1 Variable polytomique

Définition : Une variable polytomique est une variable qui peut prendre plus de deux modalités qui sont ordonnées entre elles, *petit, moyen, grand* ; *passable, assez bien, bien*,...

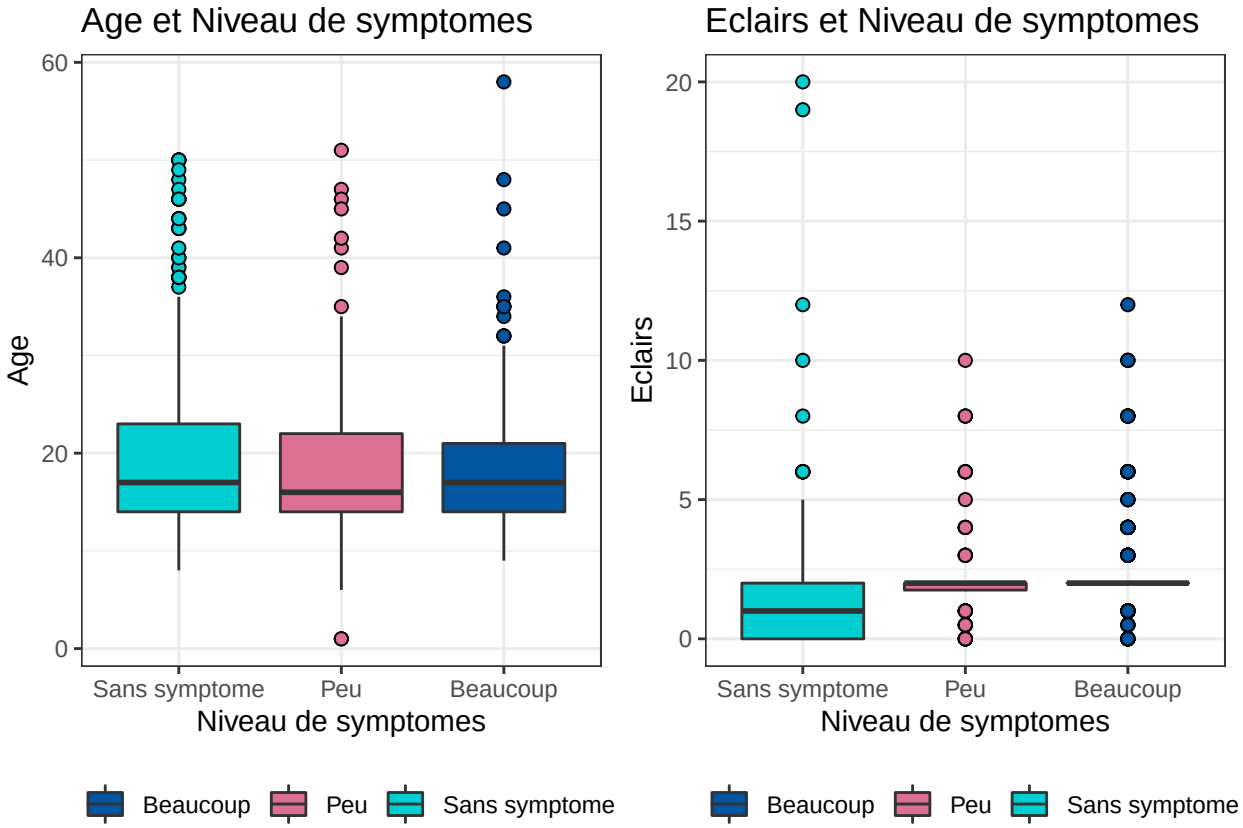
Pour créer cette nouvelle variable, nous regardons si l'individu n'avait pas de symptome la variable prendra la modalité (**Sans symptome**), s'il a un ou deux symptomes la variable prendra (**Peu**), sinon la variable prendra (**Beaucoup**). Cette variable contiendra trois modalités allons de **Pas malade** à **Beaucoup**. Le tableau ci-dessu nous donne le nombre d'individus par modalité :

Table 10 – Nombre d'individus par niveau de symptomes

Niveau de symptomes	Nombre d'individus
Sans symptome	**625**
Peu	**100**
Beaucoup	**369**

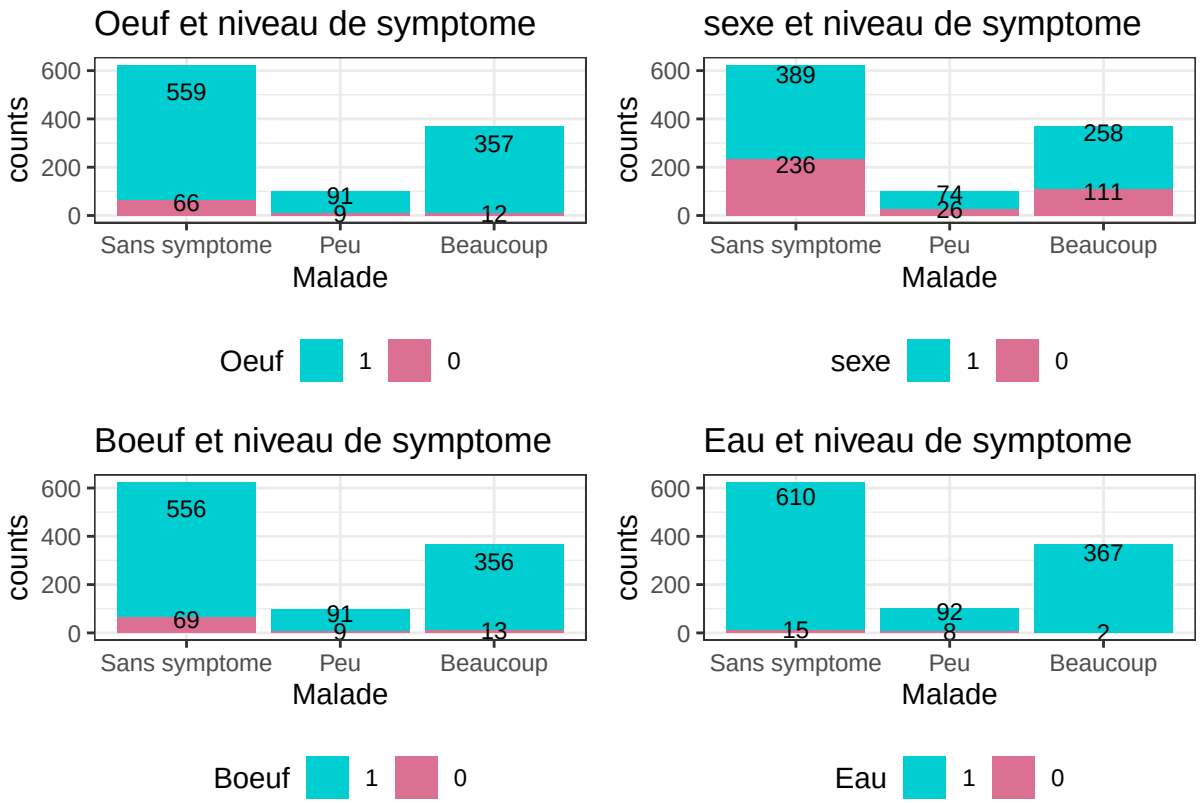
Statistiques descriptives :

Nous reproduisons les mêmes graphiques que pour la variable dichotomique :



Pour l'âge, les trois boxplots semblent identiques . Les personnes avec peu et beaucoup de symptome ont l'air plus jeunes, avec quelques exceptions (*une personne de 58 ans a beaucoup de symptomes*).

Pour les éclairs, Nous constatons que les médianes pour les personnes avec **peu** et **Beaucoup** de symptomes sont identiques (= 2) et en dessus de celui des individus **Sans symptomes** (= 1).



Pour le sexe, nous observons, comme pour la variable dichotomique, qu'un homme a plus de risque de tomber malade qu'une femme.

Pour l'Eau, l'oeuf et le boeuf, nous observons que les graphes sont assez similaires. Nous constatons plus le niveau de symptomes augmente de **Peu** à **Beaucoup** plus l'individu a consommé ces aliments.

4.2 Régression logistique avec Y polytomique ordonnée

Nous allons dans cette partie étudier la variable **Niveau de symptome** en appliquant une régression logistique polytomique ordonnée. Les modalités de cette variable sont ordonnées comme suit : (*Sans symptome* < *Peu* < *Beaucoup*).

Nous prendrons en compte toutes les variables explicatives, l'intéraction entre les variables (*sexe* et *eclair*) et toutes les intéraction de niveau trois maximum entre les variables (*Boeuf, Oeuf, age, Eau*).

TABLE 11: Première régression logistique polytomique ordonnée

Variables	Coefficients	Erreur standard	Statistique
sex1	**1.1283368**	**0.2726141**	**4.1389531**
eclair	**0.9602124**	**0.1255559**	**7.6476865**
Oeuf1	**164.2085373**	**1.0633837**	**154.4207784**
Eau1	**-1.7265838**	**1.5819860**	**-1.0914027**
Boeuf1	**-201.6189695**	**1.4137588**	**-142.6120006**
age	**-0.1265390**	**0.0387355**	**-3.2667497**

sex1 :eclair	**-0.5086172**	**0.1324549**	**-3.8399264**
Oeuf1 :Eau1	**-163.2219297**	**1.3293958**	**-122.7790349**
Oeuf1 :Boeuf1	**40.6363985**	**1.6719470**	**24.3048368**
Eau1 :Boeuf1	**204.8730987**	**1.6111264**	**127.1614030**
Oeuf1 :age	**-6.5247793**	**0.0767431**	**-85.0210989**
Eau1 :age	**0.1008011**	**0.0725980**	**1.3884834**
Boeuf1 :age	**6.2664486**	**0.1226805**	**51.0794082**
Oeuf1 :Eau1 :Boeuf1	**-43.4844222**	**1.4876882**	**-29.2295261**
Oeuf1 :Eau1 :age	**6.5330642**	**0.0700222**	**93.2998935**
Eau1 :Boeuf1 :age	**-6.4249237**	**0.0694167**	**-92.5558943**
Oeuf1 :Boeuf1 :age	**0.1511214**	**0.1614989**	**0.9357422**
Intercepts			
Sans symptome|Peu	**1.1855200**	**0.8337833**	**1.4218562**
Peu|Beaucoup	**1.6542286**	**0.8343655**	**1.9826185**

Afin d'enlever les termes non significatifs et choisir le bon modèle, nous allons réadopter le critère AIC.

Nous obtenons le résultat suivant :

TABLE 12 – Régression polytomique ordonnée choisie par le critère AIC

Variables	Coefficients	Erreur standard	Statistique
sex1	**1.0979042**	**0.2643692**	**4.1529198**
eclair	**0.9522937**	**0.1220440**	**7.8028722**
Oeuf1	**3.7089704**	**2.4601557**	**1.5076161**
Eau1	**-0.2515967**	**1.2383101**	**-0.2031775**
age	**-0.1286778**	**0.0384634**	**-3.3454603**
sex1 :eclair	**-0.4869377**	**0.1278184**	**-3.8096055**
Oeuf1 :Eau1	**-3.8010463**	**2.6476588**	**-1.4356255**
Oeuf1 :age	**-0.1348208**	**0.0916147**	**-1.4716069**
Eau1 :age	**0.0526595**	**0.0698142**	**0.7542811**
Oeuf1 :Eau1 :age	**0.1863599**	**0.1090471**	**1.7089849**
Intercepts			
Sans symptome|Peu	**1.1836012**	**0.8282054**	**1.4291154**
Peu|Beaucoup	**1.6512515**	**0.8287816**	**1.9923844**

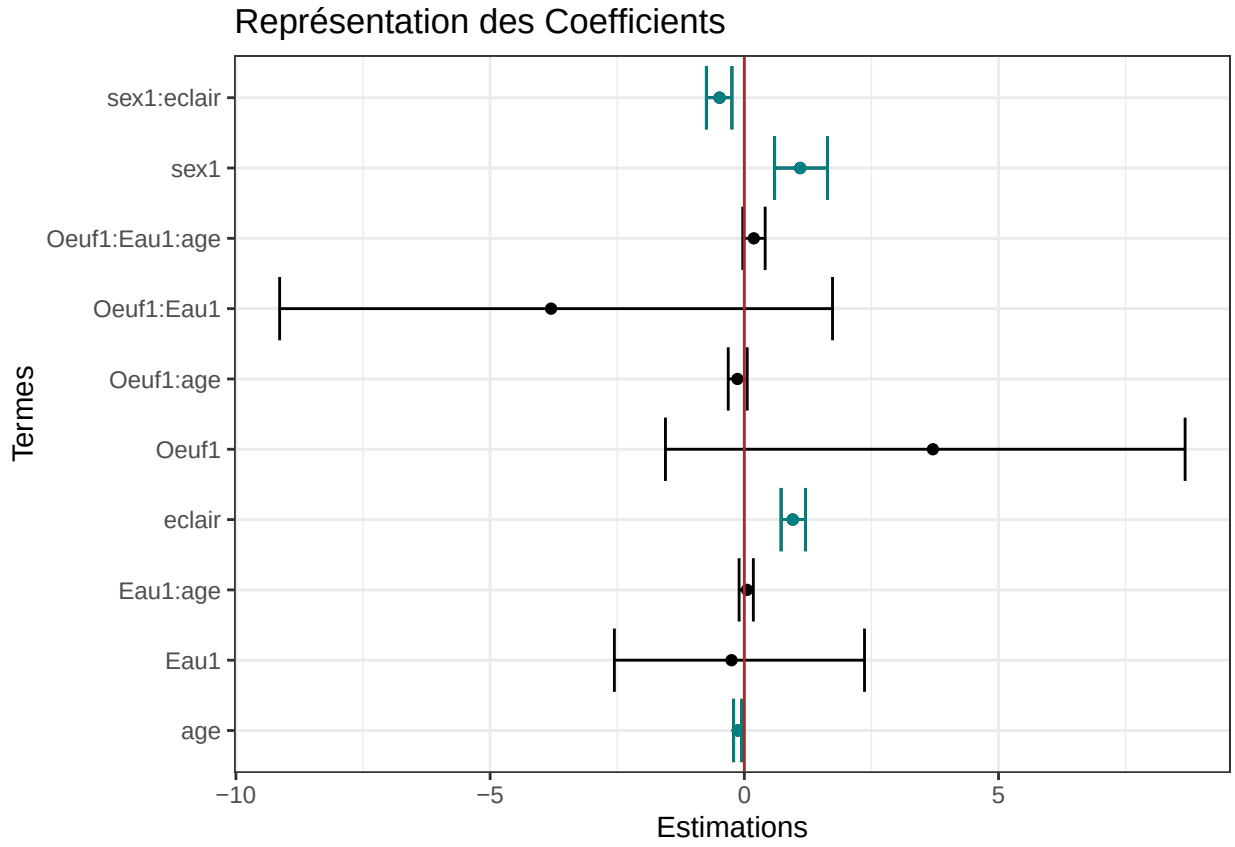
Intéressons-nous maintenant aux coefficients ainsi qu'à leurs intervalle de confiance.

TABLE 13: Coefficients et leurs intervalles de confiance pour la variable polytomique

Variables	Coefficients	2.5%	97.5%
sex1	**1.0979042**	**0.5959438**	**1.6353982**
eclair	**0.9522937**	**0.7233978**	**1.2032009**
Oeuf1	**3.7089704**	**-1.5516729**	**8.6699056**
Eau1	**-0.2515967**	**-2.5558454**	**2.3641730**
age	**-0.1286778**	**-0.2102289**	**-0.0547747**
sex1 :eclair	**-0.4869377**	**-0.7469538**	**-0.2443469**
Oeuf1 :Eau1	**-3.8010463**	**-9.1399683**	**1.7346877**
Oeuf1 :age	**-0.1348208**	**-0.3166991**	**0.0568858**

Eau1 :age	**0.0526595**	**-0.1029067**	**0.1775906**
Oeuf1 :Eau1 :age	**0.1863599**	**-0.0303082**	**0.4097888**

Pour nous permettre de repérer les coefficients significatifs, on représente sous forme de graphique les résultats obtenus précédemment :



• Nous pourrons rien dire sur les coefficients avec un intervalle de confiance contenant **0**.

• Une année de plus sur l'age de l'individu va diviser l'odds de (*Niveau de symptomes* $\leq$ *Sans symptome*) par au moins 0.8104 et au plus 0.9467. Cette augmentation divise aussi l'odds de (*Niveau de symptomes* $\leq$ *Peu de symptome*) par au moins 0.8104 et au plus 0.9467. L'**age** apparaît comme un **facteur protecteur** de tomber beaucoup malade.

• Le fait d'être homme va diviser l'odds de (*Niveau de symptomes* $\leq$ *Sans symptome*) par au moins 1.8147 et au plus 5.1315 par rapport à une femme. Ce fait divise aussi l'odds de (*Niveau de symptomes* $\leq$ *Peu de symptome*) par au moins 1.8147 et au plus 5.1315. Le fait d'être un **homme** apparaît comme un **facteur de risque** de tomber beaucoup malade.

• L'augmentation d'une unité sur le nombre d'éclairs mangés par un individu va diviser l'odds de (*Niveau de symptomes* $\leq$ *Sans symptome*) par au moins 2.0614 et au plus 3.3308. Cette augmentation divise aussi l'odds de (*Niveau de symptomes* $\leq$ *Peu de symptome*) par au moins 2.0614 et au plus 3.3308. Le nombre d'éclairs apparaît comme un **facteur de risque** de tomber beaucoup malade.

• Concernant l'interaction entre eclairs et sexe, on estime :

$$\beta = log(\frac{odds(Y=NbSymptomes \leq SansSymptome|Ecalirs=x+1,sex=1)}{odds(Y=NbSymptomes \leq SansSymptome|Ecalirs=x,sex=1)}) - log(\frac{odds(Y=NbSymptomes \leq SansSymptome|Ecalirs=x+1,sex=0)}{odds(Y=NbSymptomes \leq SansSymptome|Ecalirs=x,sex=0)})$$

Nous regardons la différence entre l'effet du nombre d'éclairs mangés sur la maladie par les hommes et les femmes. L'effet du nombre d'éclairs chez les hommes est inférieur à celui chez les femmes sur la maladie.

Parmi les homme, l'effet de l'augmentation des éclairs mangées par une unité va être multiplié par au moins 0.4738077 et par au plus 0.7832159 par rapport à la même augmentation parmis les femmes.

4.3 Résumé

La variable polytomique (**Niveau de symptomes**) est expliquée par le **sexe**, **Nombre d'eclairs** et **age**. Elle est aussi expliqué par l'interaction entre le **nombre d'éclairs** et le **sexe**.

Le fait d'être homme est un facteur de risque.

Un jeune a plus de risque de tomber malade qu'un vieux. L'âge est un facteur protecteur.

Les homme qui ont eu tendance à manger plus d'éclairs vont moins tomber malade que les femmes en ayant mangé le même nombre d'éclairs.

5 Conclusion

Nous avons pu dans ce projet appliquer différentes notions abordées dans le cours de biostatistique, en commençant par la description de la table de données, l'imputation des données manquantes, le recodage des variables et finalement l'apliquation des régressions logistiques, nous avons abouti aux résultats :

Dans ces deux types de régression logistique, nous déduisons que les manifestants sont tomber malade par leur jeune âge, le fait de ne pas boire de l'eau, manger trop d'éclairs et le fait d'être homme.

Les hommes ont été plus intoxiqués que les femmes, toutes caractéristiques égales par ailleurs.

Les buveurs d'eau ont moins de risque de tomber malade que les non buveurs, toutes choses égales par ailleurs.