Université de Strasbourg

Faculté des sciences économiques et de gestion
Université de Strasbourg

Rapport d'économétrie

Le modèle logit multinomial appliqué au covid-19.

Rédigé par

Khadidja Amadou Abdoulaye (M1 DS2E)

Julian Jimenez Mejia (M1 DS2E)

Professeur : Mr Jamel Trabelsi

Année universitaire : 2020- 2021

Table des matières

1. Introduction 3
2. Revue de littérature 5
3. Le modèle logit multinomial 7
    1. Présentation du modèle : 7
    2. Estimation 7
    3. Interprétation 8
4. Analyse empirique 8
    1. Présentation des données 8
    2. Modélisation des données 9
    3. Vérification des résultats 10
    4. Calcul de probabilité 12
    5. Prédiction du modèle 13
    6. Z-test ou le test de significativité des coefficients 14
    7. Interprétation graphique 15
5. Conclusion 18
6. Références bibliographiques 18
7. Annexes 18

1. Introduction

Nous vivons actuellement une crise sanitaire sans précédent due à une maladie virale appelée Covid-19. Cette maladie est apparue précisément le 17 novembre 2019, il y a plus d'une année dans la ville de Wuhan, en Chine et s'est propagée partout dans le monde. Le mode de transmission de cette maladie se fait par voie respiratoire. Depuis son apparition en novembre 2020, la pandémie du covid-19 a bousculé nos habitudes quotidiennes : par la distanciation sociale, le port du masque, la désinfestation ou le lavage des mains avec le gel hydroalcoolique … , sans compter le nombre de victimes dû à cette pandémie. A la date du 05 mars 2021, la pandémie du Covid -19 a ôté la vie à environ 2 570 200 personnes dans le monde entier selon l'Organisation mondiale de la santé (OMS).

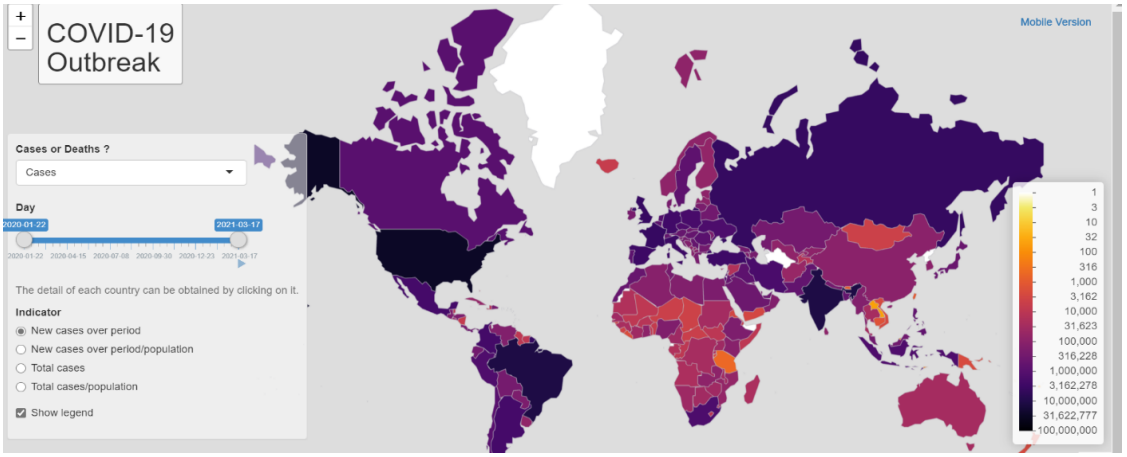


Tableau 1: Evolution du nombre de contamination dans tous les pays recouvrant la période de 22/01/2020 au 18/03/2021

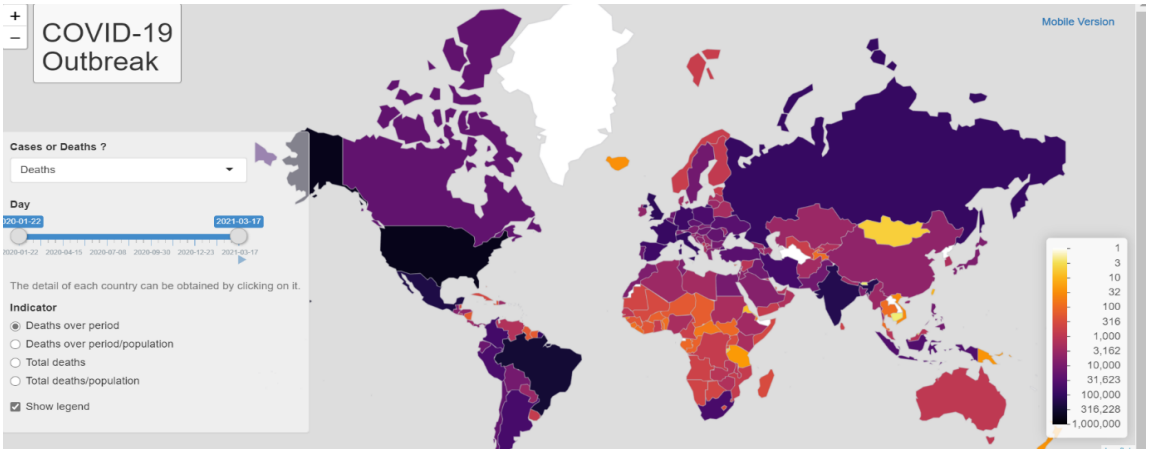


Tableau 2 Evolution du nombre de décès dans tous les pays recouvrant la période de 22/01/2020 au 18/03/2021

Les deux tableaux ci-dessus nous montrent respectivement l'évolution de cas de contamination et du nombre de décès cumulés dès la première apparition du virus (le 22/01/2021) au 18 mars 2021 . Nous pouvons remarquer qu'il y a une forte corrélation entre le nombre de contaminations et le nombre de décès. En effet, les pays où l'on observe un très grand nombre de contamination sont ceux qui ont eu le plus grand nombre de patients décédés à la suite de la Covid-19 ; c'est le cas des Etats-Unis, du Brésil … où le nombre de décès et de contamination ont atteint la barre d'un million de cas contaminés et de décès. La chine (le pays d'origine du virus) comptabilise moins de 10000 cas. Pourtant, elle est le lieu d'origine du virus.

Face à la crise du Covid-19 et aux nombreux inconvénients qu'elle a engendrés, plusieurs actions ont été menées par les gouvernements afin de limiter le nombre de contaminations. Ces actions concernent : la distanciation sociale, la fermeture des lieux publics (bars, restaurants, cinéma, …), un confinement partiel ou total, un couvre-feu qui varie en fonction des villes, des publicités de prévention dans les supports médiatiques, des campagnes de tests massives avec le vaccin Biotech ou Pfizer …. Parmi les mesures prises par les décideurs publics, la plus importante est celle de la distanciation sociale. Ainsi, la majorité des autorités conseillent aux personnes âgées de rester chez elles car elles sont considérées comme des personnes à risque. En effet, les adultes de plus de 65 ans et les personnes ayant des problèmes de santé seraient les plus vulnérables face à cette maladie.

Notre dossier s'articule de la manière suivante. Dans un premier temps nous allons faire une revue de la littérature qui nous permettra de présenter le modèle multinomial logit, son estimation et son interprétation et expliquer pourquoi ce modèle nous semble pertinent pour analyser nos données. Dans la deuxième partie, nous allons faire une analyse empirique des données dans laquelle nous allons présenter les données, les modéliser grâce au modèle logit multinomial, calculer et interpréter les différentes probabilités obtenues appliqué au genre (homme, femme) puis de donner une conclusion grâce à la partie prediction.la troisième partie de notre dossier sera consacré à l'interprétation graphique. La dernière partie nous permettra de faire une conclusion et quelques commentaires économiques de nos résultats.

2. Revue de littérature

Plusieurs études ont été réalisées sur le sujet. Les chercheurs et scientifiques du monde entier ont souhaité révéler les déterminants de la contamination. Une étude de la direction de la recherche, des études, de l'évaluation et de la statistiques (**Drees)** du ministère de la santé publique français a analysé le parcours de 90000 personnes hospitalisées entre le 1er mars 2020 et le 15 juin 2020. Cette étude a permis de faire un point sur l'évolution de l'épidémie entre le 1 mars et le 15 juin et d'analyser le portrait type des malades. Les résultats ont été sans surprise. Sur l'ensemble de ces personnes testées, ce sont les personnes âgées (de plus de 65 ans) qui représentent les malades hospitalisés. L'âge médian des personnes décédées est de 81 ans. Lors de la première vague, les résultats ont montré que les hommes étaient les premières

victimes de la Covid -19 puisqu'ils représentaient 60% des personnes décédées contre environ 40% chez les femmes.

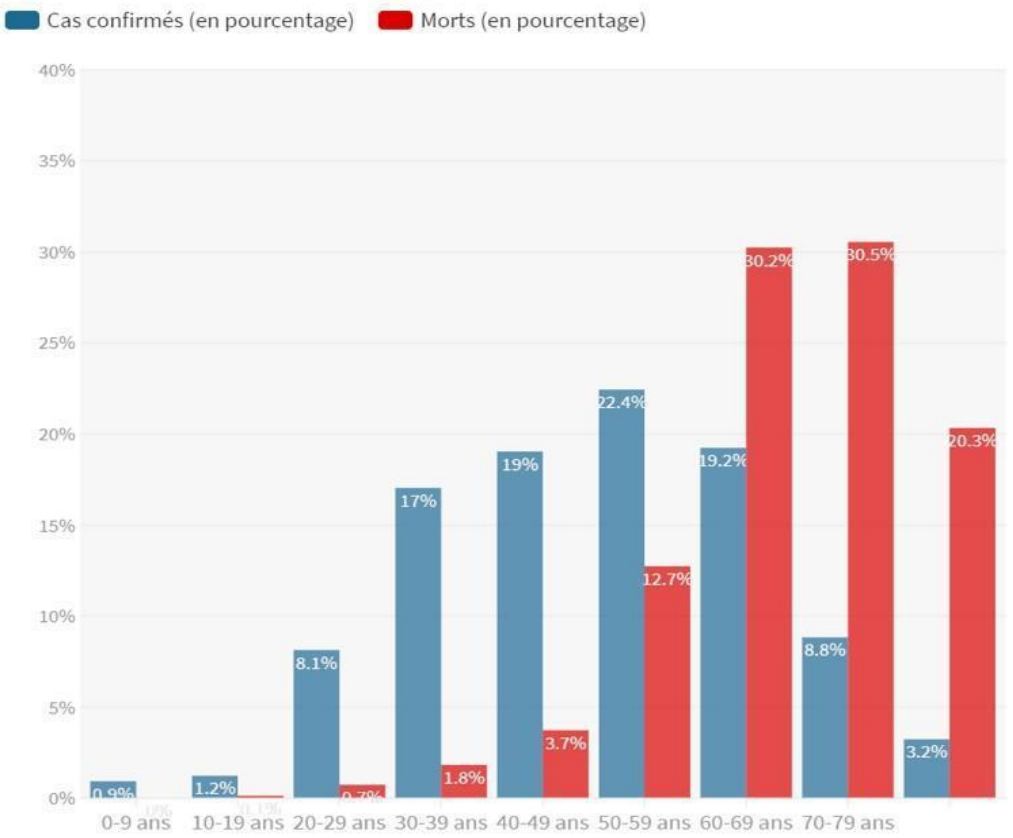


Figure 3 taux de contamination et de décès dû au covid -19 en fonction de l'âge des patients.

De même, Yanez et all (2020) dans leur article *"COVID-19 mortality risk for older men and women" présentent* leur étude, où les informations obtenues de 16 pays corroborent à l'idée que les taux de mortalité sont fortement liés aux groupes de population plus âgés. Il est important de noter que bien que ces pays suivent un comportement similaire en termes de taux de mortalité dans cette tranche de la population, ils présentent également des différences marquées qui peuvent aller des systèmes de santé locaux, aux caractéristiques (comorbidités) des patients affectés dans chaque pays, ceci étant un aspect fondamental pour déterminer l'impact de covid-19 chez les patients ayant des maladies préexistantes telles que l'hypertension, le diabète et l'obésité. Ces résultats sont aussi contrastés par Inam Z et all (2020).

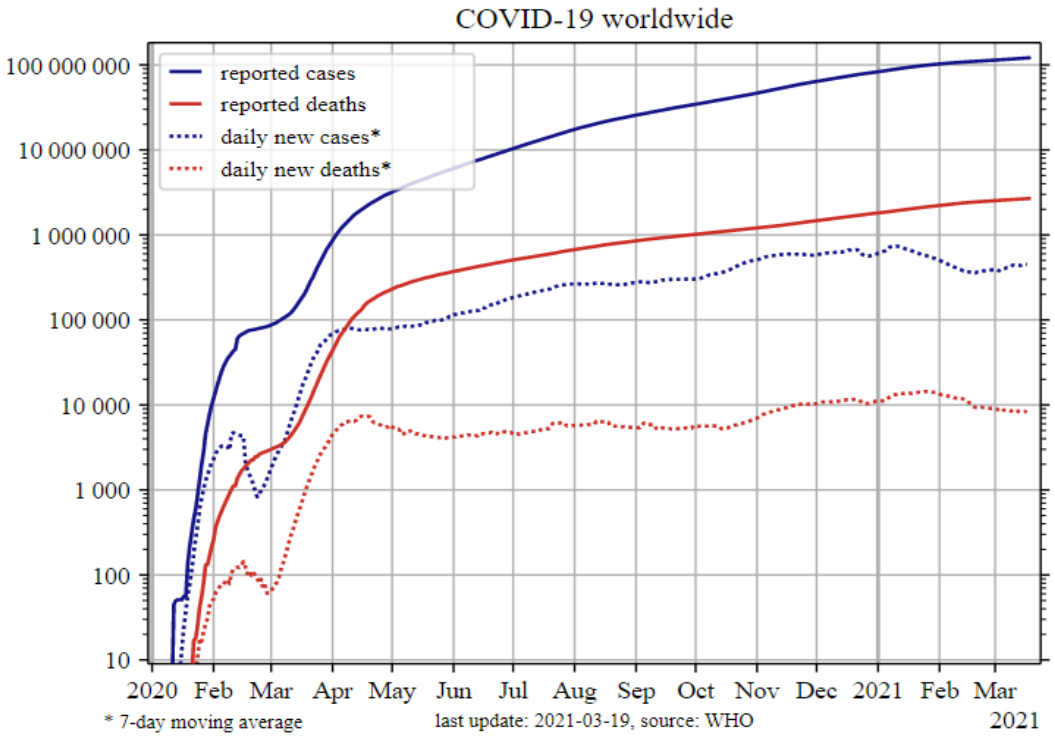


Figure 4 Evolution du nombre de décès dans tous les pays recouvrant la période de 22/01/2020 au 18/03/2021

Le modèle logit multinomial

1. Présentation du modèle :

Les modèles de type multinomial logit sont des modèles probabilistes non linéaires utilisés lorsque la variable explicative peut prendre plusieurs valeurs. C'est un modèle stochastique qui peut prendre en compte plusieurs variables explicatives.

Le modèle que nous allons estimer peut prendre plusieurs valeurs, y ∈ 1,2, 3, k avec k un entier positif. Y définit une catégorie, donc il n'y a pas d'ordre à respecter. C'est donc un modèle polytomique non ordonné. Par exemple si l'on veut déterminer le mode de transport des étudiants de la FACULTÉ DES SCIENCES ECONOMIQUES ET DE GESTION DE STRASBOURG, la variable Y1 peut prendre plusieurs valeurs. Y1 = 1 représente le tram, Y2 = 2 représente le vélo et Y3 = 3 représente la voiture, et X représente la distance par rapport au domicile des étudiants. Cela nous permet donc d'avoir plusieurs catégories, ou un menu de moyens de déplacement qui sont influencés par une ou plusieurs variables explicatives qui peuvent être la distance, le niveau financier … Étant donné que Yj prend plusieurs valeurs selon k, on va s'intéresser à P(y = j | x), ∀k avec x une matrice 1 xJ telle que x1 = 1.

Notre modèle suit donc une loi logistique.

Par définition, on dit qu'un modèle suit une loi logistique si sa fonction de distribution est une fonction logique. On a :

$$P(y = j \mid x) = \frac{e^{x\beta_j}}{1 + \sum_{h=1}^{J} (e^{x\beta_h})} \qquad j = 1...J$$

Avec B_j inconnu et de dimension jx1
La somme des probabilités fait 1, ainsi on peut écrire : ($P(y = j \mid x) + P(y = 0 \mid x)$) = 1

Avec

$$P(y = 0 \mid x)) = \frac{1}{1 + \sum_{h=1}^{J} \epsilon}$$

On remarque que lorsque j=1, on a un modèle logit binaire.

2. Estimation

On peut estimer le modèle par la méthode du maximum de vraisemblance. Tout d'abord il faut appliquer le logarithme à la densité vue plus haut.

$$Li(\beta) = X1[yi = j] \,/\, log\,[P(y = j|x)]$$

avec j= 0

Une fois qu'on obtient la log vraisemblance il faut maximiser $X(Li(\beta))$, on obtient alors l'estimateur du maximum de vraisemblance.

$$\widehat{\beta} = \frac{\partial \sum_{i}^{N} (l_i(\beta))}{\partial \beta}$$

3. Interprétation

L'interprétation de ce type de modélisation est plus délicate que pour les modélisations basiques.

L'effet partiel d'un changement de Xk sur y se traduit par :

$$\frac{\partial P(y = j \mid x)}{\partial x_k} = P(y = j \mid x)\{\beta_{jk} - [\frac{\sum_{h=1}^{J} \beta_{hk} e^{x\beta_h}}{g(x,\beta)}]$$

où β_{hk} est le $k^{ieme`}$ élément de β_h et $g(x,\beta) = 1 + \sum_{h=1}^{X} e^{x\beta_h}$

La direction de l'effet partiel n'est pas déterminée par $_{jk}$.

Nous allons maintenant voir comment interpréter les coefficients obtenus. Nous allons montrer dans le cas pratique que

$$log[\frac{P(y = j \mid x)}{P(y = 0 \mid x)}] = x\beta_j$$

et cela représente le ratio de chance ou le ratio de pourcentage.

Si l'on change d'une unité x_k alors on change le ratio de chance de β_{jk}

3. Analyse empirique

1. Présentation des données

Au cours de nos recherches, nous n'avons pas trouvé de données récentes concernant la deuxième vague du covid 19. Nous avons fait notre étude sur les résultats des recherches obtenus lors de la première vague (février 2020 à juin 2020).

Dans le cadre de notre étude, nous avons voulu étudier l'influence de l'âge et du sexe des individus sur le risque d'etre testé positif à la covid -19. Le modèle que nous voulons étudier est un modèle polytomique/ logit multinomial avec une variable dépendante ordinale qui peut prendre plusieurs valeurs. Nous avons pris 3

catégories pour la variable de reponse . Un individu peut être : soit en bonne santé (1), soit malade (2) soit décédé (3) du covid -19. Nous avons pris 2 variables explicatives Xi qui sont l'âge et le sexe.

2. Modélisation des données

Nous commençons d'abord par catégoriser la variable dépendante y. Puis nous modélisons notre fonction grâce à la fonction mlogit du package() en mettant comme catégorie de référence l'individu sain :y=1.

```
# Modelisation multinominal Logit

mydata.logit = mlogit.data(covid19, choice = Outcome,shape = wide, alt.levels = c
    (1,2,3))
mnl.regressio,= mlogit(Outcome ~ 0 | Age + Gender, data = mydata.logit, reflevel =
1 )
summary(mnl.reg)
```

```
Call:
mlogit(formula = Outcome ~ 0 | Age + Gender, data = mydata.mnl,
    reflevel = 1, method = "nr")

Frequencies of alternatives:choice
       0        1        2
0.759091 0.157640 0.083269

nr method
7 iterations, 0h:0m:5s
g'(-H)^-1g = 0.000435
successive function values within tolerance limits

Coefficients :
                  Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept):1 -1.83095729  0.05141487 -35.6114 < 2.2e-16 ***
(Intercept):2 -8.03203598  0.15818744 -50.7754 < 2.2e-16 ***
Age:1          0.00379347  0.00086062   4.4079 1.044e-05 ***
Age:2          0.08382457  0.00194816  43.0275 < 2.2e-16 ***
Gender:1       0.15922670  0.03774083   4.2190 2.454e-05 ***
Gender:2       0.59964542  0.05589405  10.7283 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -13840
McFadden R^2:  0.10738
Likelihood ratio test : chisq = 3329.8 (p.value = < 2.22e-16)
```

Ensuite nous réalisons des prédictions. Ceci permet de voir le pourcentage du modèle qui a été correctement prédit.

```
# Predicted probabilities (1st syntax)

n = NROW(covid19)# number of observations pp.logit
= fitted(mnl.regression, outcome = FALSE)
head(pp.logit)
pp.success = array(NA, c(n ,3) )   # good prediction = 1, bad prediction = 0 for (i in
1:n) {
   pp.success[i,1] = ( covid19 \$Outcome[i]==1 & pp.logit[i,1]==max(pp.logit[i,]))
   pp.success[i,2] = ( covid19 \$Outcome[i]==2 & pp.logit[i,2]==max(pp.logit[i,]))
   pp.success[i,3] = ( covid19 \$Outcome[i]==3 & pp.logit[i,3]==max(pp.logit[i,]))
}
print(Percent correctly predicted, MNL); sum(pp.success)/n

# Predicted probabilities (2nd syntax)
pp.success2 = rep(NA,n)   # predicted status (1, 2 or 3) for (i in 1:n) {
   if ( pp.logit[i,1]==max(pp.logit[i,]) ) { pp.success2[i] = 1 } if (
   pp.logit[i,2]==max(pp.logit[i,]) ) { pp.success2[i] = 2 } if (
   pp.logit[i,3]==max(pp.logit[i,]) ) { pp.success2[i] = 3 }
}
with(covid19, table(covid19\$Outcome ,pp.success2)) # Observed (status) versus
     Predicted (pp.success2)
print(Percent correctly predicted, LOGIT); sum(diag(with(covid19, table(Outcome,
     pp.success2)) ))/n
```

```
> print("Percent correctly predicted"); sum(diag(with(covid_19, table(Outcome,pp.success2))
 ))/n
[1] "Percent correctly predicted"
[1] 0.7466016
```

On constate que 75% du modèle à bien été prédit, et donc 25% du modèle a été mal prédit.

3. Vérification des résultats

L'objectif de cette section est de vérifier les résultats obtenus précédemment. Pour cela au lieu d'utiliser la fonction mlogit nous utiliserons la fonction multinom du package (nnet). Pour se faire, on commence par convertir la variable de réponse en variable de catégorie par la fonction factor. Ensuite, il nous faut choisir un état ou une catégorie de référence. Cette catégorie sera l'individu est saint (1). Cette opération est possible par la fonction relevel de R.

```
# Modelisation multinomial regression

covid19$OutcomeF <- factor(covid19 $Outcome) # Conversion en variable
catégorique
covid19$out <- relevel(covid19 $OutcomeF, ref = 1)
# La référence est le patient sain (1)
```

Afin de réaliser notre modélisation, nous allons utiliser la fonction multinom de R comme argument de notre formule. Cela nous permettra de définir notre modélisation par une fonction linéaire.

```
modelisation <- multinom(out ~ Age + Gender , data = covid19)
```

tableau

On constate que l'erreur très élevée qui se divise par 2 après 10 itérations puis converge vers une valeur finale.

```
summary(modelisation)
```

```
> summary(modelisation)
Call:
multinom(formula = out ~ Age + Gender, data = covid_19)

Coefficients:
  (Intercept)         Age    Gender
1   -1.830874 0.003792423 0.1592337
2   -8.033048 0.083834937 0.6001472

Std. Errors:
  (Intercept)          Age     Gender
1  0.05141438 0.0008606122 0.03774055
2  0.15820759 0.0019483688 0.05589533

Residual Deviance: 27680.21
AIC: 27692.21
```

On constate que les coefficients sont exactement similaires à ceux de la modélisation précédente

4. Calcul de probabilité

Grâce aux coefficients obtenus par la modélisation nous pouvons effectuer le calcul de probabilité du ratio de chance d'être contaminé. Nous avons donc les deux équations suivantes :

$$log[\frac{P(2)}{P(1)}] = -1.830862 + 0.00379220(Age) + 0.1592892(Gender) = y_2$$

$$log[\frac{P(3)}{P(1)}] = -8.033858 + 0.08384524(Age) + 0.5998309(Gender) = y_3$$

Nous constatons que le coefficient Âge est positif sur la probabilité d'être infecté par le covid-19 par rapport à un individu sain.

Ces 2 équations vont nous permettre de calculer les probabilités individuelles. Si on transforme les deux en passant à l'exponentielle des deux côtés de l'égalité, on a:

$$\frac{P(2)}{P(1)} = e^{y_2}$$
$$\frac{P(2)}{P(1)} = e^{y_3}$$

La somme de ces 2 équations nous donne

$$\frac{P(2) + P(3)}{P(1)} = e^{y_2} + e^{y_3}$$

et comme *P*(1) + *P*(2) + *P*(3) = 1 ceci implique

$$\frac{1 - P(1)}{P(1)} = e^{y_2} + e^{y_3}$$

$$\frac{1}{P(1)} = 1 + e^{y_2} + e^{y_3}$$

$$P(1) = \frac{1}{1 + e^{y_2} + e^{y_3}}$$

Ce calcul représente la probabilité pour un individu d'être saint (1). On peut en déduire les autres probabilités qui correspondent respectivement à la probabilité d'être malade et la probabilité de mourir de la covid -19.

$$\frac{P(2)}{P(1)} = e^{y_2}$$
$$P(2) = \frac{e^{y_2}}{1 + e^{y_2} + e^{y_3}}$$
$$P(3) = \frac{e^{y_3}}{1 + e^{y_2} + e^{y_3}}$$

5. Prédiction du modèle

Dans cette section nous réalisons des prévisions sur notre modèle. Cela nous permet d'obtenir les différentes probabilités de chacune des catégories 1,2 et 3. autrement dit, les probabilités d'être atteint de la covid 19 associés respectivement à une personne en bonne santé, malade ou décédée. Pour chaque individu la somme des probabilités est égale à 1.

```
Predict (modelisation, data = covid19, type = prob)
```

```
100    0.7539071 0.1778915 0.0682013627
```

Pour l'individu 100, il a 75,39 % de chances d'être saint, 17,78 % de chance d'être infecté et 6,9 % de chance de mourir du covid-19.

Par ailleurs, il est intéressant de comparer nos prédictions avec les valeurs réelles du modèle. Pour cela, on crée une matrice qui classe les individus sains, infectés et décédés du covid19 pour notre base de données et pour nos prédictions.

```
confusionMatrix <- table(predict(modelisation),covid19$OutcomeF)
print(confusionMatrix)
```

```
     1       2       3
116389     3379    1628
2       0       0       0
3     252      77     197
```

Nous pouvons constater que 16389 individus sont classés saints par la base de données et par les prédictions, 252 individus sont classés saints par la base de données mais décédés par les prédictions.

Il s'ensuit que nous pouvons déterminer le pourcentage de mauvaise classification de notre prédiction.

```
1-sum(diag(confusionMatrix))/sum(confusionMatrix)
```

```
[1]  0.2434 084
```

24% de mauvaise classification c'est à dire 24% du temps le modèle classe mal un individu. En d'autres termes, environ 75% de modèle est bien prédit. Ce qui s'accorde une nouvelle fois avec l'étude précédente.

Pour des soucis de vérification on peut calculer la somme de la probabilité de prédiction du modèle précédente avec la probabilité de mauvaise prédiction du modèle présent.

```
a<-sum(pp.success)/n
b<-1-sum(diag(confusionMatrix))/sum(confusionMatrix)
a+b
```

```
> a+b
[1] 1
```

La valeur est exactement 1 ainsi on peut conclure que les deux études sont exactement les mêmes, mais avec de petites variations de m´méthodes.

6. Z-test ou le test de significativité des coefficients

Nous continuons notre étude en testant la significativité des coefficients grâce à un Z-test.

```
# Two tail Z-Test

Z_test <- summary(modelisation)
$coefficients/summary(modelisation) $standard.errors p <- (1 -
pnorm(abs(z), 0, 1)) * 2 p
```

```
z_test <- summary(modelisation) $coefficients/summary(modelisation) $standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
(Intercept)         Age       Gender
          0 1.04978e-05 2.452057e-05
          0 0.00000e+00 0.000000e+00
```

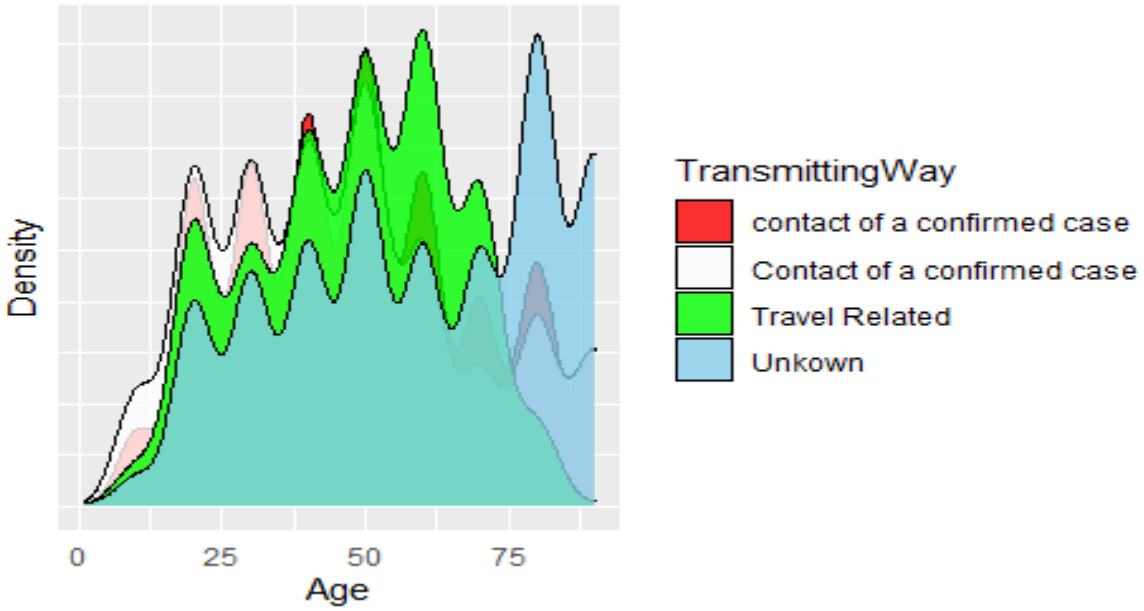
Les p-values sont très proches de 0, ce qui signifie un fort niveau de confiance. Ainsi les variables sont significatives à 99%.

7. Interprétation graphique

Le but de cette section est d'apporter une interprétation graphique de nos résultats. Avant d'expliquer graphiquement le lien entre la transmission du covid 19 par rapport à la tranche d'âge, nous réalisons un histogramme en proportion.

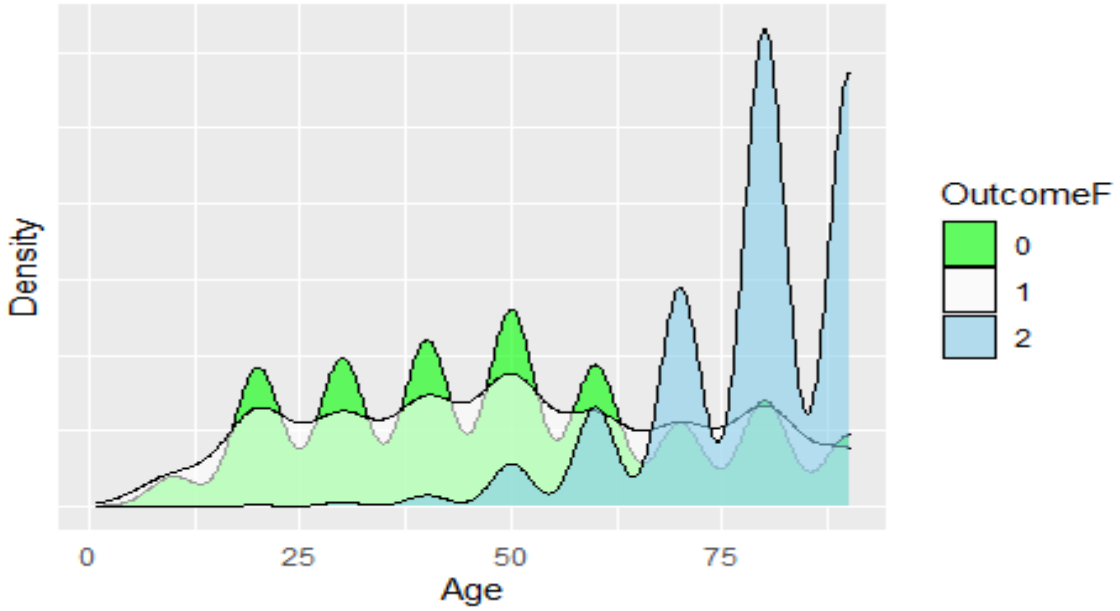
```
#Graphs

A= ggplot(covid19,aes(x = Age, fill = OutcomeF))+ ylab(Density) + geom_density(
    alpha = 0.8)+scale_fill_manual(values = c(scales::alpha(brown,.5),scales::
    alpha(red,.5),skyblue)) + theme(axis.text.y = element_blank(), axis.ticks =
    element_blank())
B= ggplot(covid19, aes(x = Age, fill=TransmittingWay))+ geom_density(alpha =
    0.8)+ ylab(Density) +scale_fill_manual(values =
    c(scales::alpha(brown,.8),scales ::alpha(red,.5),scales::alpha(skyblue,.5)))+
    theme(axis.text.y = element_ blank(), axis.ticks = element_blank())
```

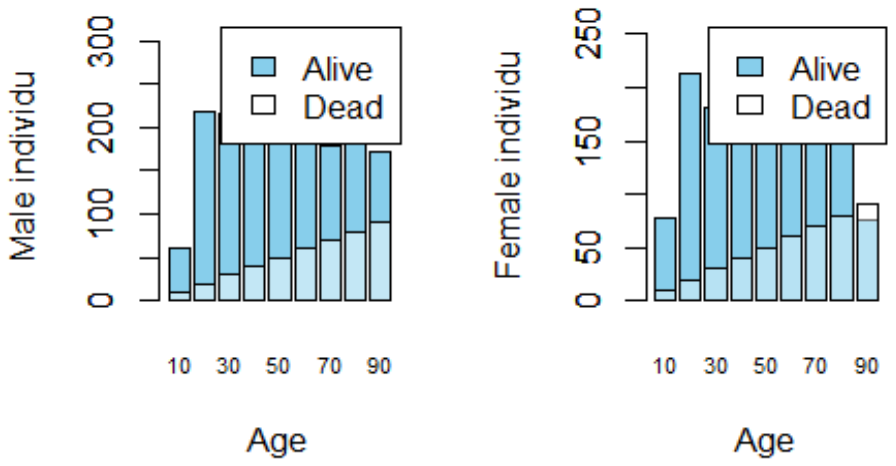


La figure 4 nous montre un graphique qui représente les moyens de transmission du covid-19 par proportion. Nous avons l'âge en abscisse et la densité en ordonnée. La courbe du moyen de transmission est non aléatoire et varie en fonction de l'âge. On peut remarquer que les personnes qui se transmettent le plus le virus se trouvent dans la tranche d'âge de 40 à 50 ans. Un pic de contamination est observé à l'âge 50. De plus, on peut constater que la transmission directe est très forte. On peut notamment voir que la transmission par le voyage est aussi fortement représentée. Par ailleurs elle est très présente pour les personnes assez âgées (plus de 50 ans) ce qui peut s'expliquer par le fait qu'ils ont les moyens de voyager et sont plus libres et contrairement aux jeunes et aux parents.

Dans un second temps, on vérifie l'état de santé des individus selon leur âge une nouvelle fois.



Sur la figure 5, on constate que les personnes qui meurent du covid 19 sont les personnes plus âgées. Ce qui confirme notre hypothèse. On voit aussi que le covid 19 touche presque uniformément toutes les catégories d'âges. Sauf les enfants en bas âge. Enfin les personnes qui guérissent du virus se trouvent dans une tranche d'âge de 15 à 55 ans ce qui semble correspondre à la réalité étant donné que leur système immunitaire est plus fort que les enfants ou les personnes plus âgées. Au vu des résultats de cette étude, nous pouvons conclure que l'Age et le sexe jouent un rôle déterminant quant au taux de contamination et de mortalité due à la Covid-19.



Enfin sur la figure 6, on remarque une nouvelle fois que les personnes âgées sont celles qui meurent le plus du virus. Cependant, on remarque que les femmes ont tendance à mieux résister au coronavirus que les hommes.

4. Conclusion

Ce projet a permis de mettre en lumière les modèles logit multinomiaux et a permis de les appliquer au cas du Covid -19. Les travaux nous ont permis de modéliser le concept de régression multinomial appliqué au cas du covid 19 sur des données publiques du canada.

L'analyse des résultats nous ont permis de conclure que ce sont les personnes âgées qui sont les plus vulnérables à la transmission. Nous avons aussi vu que les hommes sont les plus atteints du covid-19 . Ainsi, l'âge et le sexe jouent un rôle important dans la transmission du virus selon la situation de l'individu. Grâce au test de Z-test nous avons pu tester la significativité de nos coefficients.

Enfin, les interprétations graphiques nous ont permis aussi de confirmer les résultats obtenus.

5. Références bibliographiques

- Coronavirus (COVID-19) tracker, latest cases in Canada. (2021). Retrieved 10 March 2021, from https://www.covid-19canada.com/#latest-news

- Maladie à coronavirus 2019 — Wikipédia. (2021). Retrieved 5 March 2021, from https://fr.wikipedia.org/wiki/Maladie_%C3%A0_coronavirus_2019

- Demagny, X. (2021). Covid-19 : une étude permet de dresser le profil type des malades. Retrieved 6 March 2021, from https://www.franceinter.fr/societe/covid-19-une-etude-permet-de-dresser-le-profil-type-des-malades

- Etude sur le profil des patients du covid en France :Coronavirus: le profil type des malades en Chine enfin connu grâce à une grande étude | Le HuffPost (huffingtonpost.fr)

- Imam Z, Odish F, Gill I, O'Connor D, Armstrong J, Vanood A, Ibironke O, Hanna A, Ranski A, Halalau A. Older age and comorbidity are independent mortality predictors in

a large cohort of 1305 COVID-19 patients in Michigan, United States. J Intern Med. 2020;288(4):469–76. https://doi.org/10.1111/joim.13119.

- O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590,** 140–145 (2021). https://doi.org/10.1038/s41586-020-2918-0

- (2021). Retrieved 13 March 2021, from https://www.univ-orleans.fr/deg/masters/ESA/CH/Qualitatif_Chapitre2.pdf
-
- (2021). Retrieved 9 March 2021, from https://thibautfabacher.shinyapps.io/covid-19/

6. Annexes

```
#####Multinomial Regression : Covid-19
library(readxl)
library(nnet)
library(ggplot2)
library(mlogit)
library(stargazer)
# Importations des donnees
covid_19 <- read_excel("/Users/amado/Downloads/covid19.xlsx")
covid_19
covid_19$TransmittingWay[covid_19$TransmittingWay %in% c("Information
pending")]<-
  "contact of a confirmed case"
covid_19$TransmittingWay[covid_19$TransmittingWay %in%
c("Neither")]<-"Unkown"
covid_19$TransmittingWay[covid_19$TransmittingWay %in% c("Travel
related")]<-"Travel Related"

# Modelisation multinominal Logit
mydata.logit = mlogit.data(covid_19, choice = "Outcome",shape =
"wide",alt.levels = c(1,2,3))
mydata.logit
mnl.regression = mlogit(Outcome ~ 0 | Age + Gender, data = mydata.mnl,
reflevel = 1 )
summary(mnl.regression)

# Predicted probabilities (1st syntax)
n = NROW(covid_19) # number of observations
pp.logit = fitted(mnl.regression, outcome = FALSE)
head(pp.logit)
pp.success = array(NA, c(n ,3) ) # good prediction = 1, bad prediction = 0
for (i in 1:n) {
  pp.success[i,1] = ( covid_19 $Outcome[i]==1 & pp.mnl[i,
1]==max(pp.logit[i,]))
  pp.success[i,2] = ( covid_19 $Outcome[i]==2 & pp.mnl[i,
2]==max(pp.logit[i,]))
  pp.success[i,3] = ( covid_19 $Outcome[i]==3 & pp.mnl[i,
3]==max(pp.logit[i,]))
}
print("percent predicted,LOGIT"); sum(pp.success)/n

# Predicted probabilities (2nd syntax)
pp.success2 = rep(NA,n) # predicted status (1, 2 or 3)
for (i in 1:n) {
  if ( pp.logit[i,1]==max(pp.logit[i,]) ) { pp.success2[i] = 1 }
  if ( pp.logit[i,2]==max(pp.logit[i,]) ) { pp.success2[i] = 2 }
  if ( pp.logit[i,3]==max(pp.logit[i,]) ) { pp.success2[i] = 3 }
}
with(covid_19, table(covid_19 $Outcome ,pp.success2)) # Observed (status)
versus Predicted (pp.success2)
print("Percent correctly predicted"); sum(diag(with(covid_19,
table(Outcome,pp.success2)) ))/n

# Modelisation multionomial regression
covid_19$OutcomeF <- factor(covid_19 $Outcome)
covid_19$OutcomeF
###conversion en variable categorique
covid_19$out <- relevel(covid_19$OutcomeF, ref = 1) ###la reference est le
patient saint (1)
modelisation <- multinom(out ~ Age + Gender, data = covid_19)
summary(modelisation)
```

```
# Prediction
k=predict(modelisation, data = covid_19, type = "prob")
k

# Comparaison
predict(modelisation,dimen=1)
confusionmatrix=table(predict(modelisation),covid_19$OutcomeF)
print(confusionmatrix)
1-sum(diag(confusionmatrix))/sum(confusionmatrix)
sum (pp.success)/n +1- sum (diag(confusionmatrix))/sum (confusionmatrix)

# Two tail Z-Test
z_test <- summary(modelisation) $coefficients/summary(modelisation)
$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p

# Graphs
v=ggplot(covid_19,aes(x = Age, fill = OutcomeF))+ ylab("Density") +
geom_density(
  alpha = 0.6)+scale_fill_manual(values = c(scales::alpha( "green" ,.
3),scales::alpha( "white" ,.3), "skyblue" )) + theme(axis.text.y =
element_blank(), axis.ticks =

element_blank())
v
o=ggplot(covid_19, aes(x = Age, fill=TransmittingWay))+ geom_density(alpha
=0.8)+
  ylab("Density") +scale_fill_manual(values = c(scales::alpha( "red" ,.
8),scales
                                                  :: alpha ( "white" ,.5) ,
scales :: alpha ( "green" ,.5), "skyblue" ) ) + theme ( axis.text.y =
element_blank () , axis.ticks =

element_blank())
o



# Initialisation
age <- c(10, 20, 30, 40, 50, 60, 70, 80, 90)
death <- c(0, 0 ,0 ,0 ,0 ,0, 0, 0, 0)
womenDeathHist <- array(c(age,death), dim = c(9, 2))
womenAliveHist <- array(c(age,death), dim = c(9, 2))
menDeathHist <- array(c(age,death), dim = c(9, 2))
menAliveHist <- array(c(age,death), dim = c(9, 2))
deathHist <- array(c(age,death), dim = c(9, 2))
aliveHist <- array(c(age,death), dim = c(9, 2))

# Function
for(i in 1:length(covid_19 $Age)) {
  if (covid_19 $Outcome[i] == 3)
    deathHist[(covid_19 $Age[i])/10,2] = deathHist[(covid_19 $Age[i])/10,2]
+ 1
  else if (covid_19 $Outcome[i] == 1)
    aliveHist[(covid_19 $Age[i])/10,2] = aliveHist[(covid_19 $Age[i])/10,2]
+ 1

  if (covid_19 $Outcome[i] == 3 & covid_19 $Gender[i] == 0)
```

```
    womenDeathHist[(covid_19 $Age[i])/10,2] = womenDeathHist[(covid_19
$Age[i])
                                                                /10,2] + 1
  else if ( covid_19 $ Outcome [ i ] == 1 & covid_19 $ Gender [ i ] == 0)
    womenAliveHist [( covid_19 $ Age [ i ])/10,2] = womenAliveHist [(
covid_19 $ Age [ i ])/10,2] + 1
  else if (covid_19 $Outcome[i] == 3 & covid_19 $Gender[i] == 1)
    menDeathHist[(covid_19 $Age[i])/10,2] = menDeathHist[(covid_19 $Age[i])/
10,2] + 1
  else if (covid_19 $Outcome[i] == 1 & covid_19$Gender[i] == 1)
    menAliveHist[(covid_19 $Age[i])/10,2] = menAliveHist[(covid_19 $Age[i])/
10 ,2] + 1
}

#Plot
barplot(aliveHist[,2], xlab = "Age", ylab = "Individus",names.arg =
c(10,20,30,40,50,60,70,80,90), cex.names = 0.7,col ="skyblue" )
barplot(deathHist[,1], col=scales::alpha("White" ,.5), add=T)
legend("topright", c("Alive","Dead"), fill = c(col="skyblue"
,col=scales::alpha( "White",.5)))

par(mfrow=c(1,2))
barplot(womenAliveHist[,2], xlab = "Age", ylab = "Male individu",names.arg =
c(10,20,30,40,50,60,70,80,90), cex.names = 0.7,col= "skyblue" )
barplot(womenDeathHist[,1], col=scales::alpha( "White" ,.5), add=T)
legend("topright", c("Alive","Dead"), fill = c( "skyblue"
,col=scales::alpha( "White" ,.4)
))

barplot(menAliveHist[,2], xlab = "Age", ylab = "Female individu",names.arg =
c(10,20,30,40,50,60,70,80,90), cex.names = 0.7,col= "skyblue" )
barplot(menDeathHist[,1], col=scales::alpha( "White" ,.4), add=T)
legend("topright", c("Alive","Dead"), fill = c( "skyblue"
,col=scales::alpha( "White" ,.5)
))
```