Mais qui a des téléphones ?!

Justine Blanchot et Elyass Sayd, Groupe 0

25/04/2022

Introduction

Avant la fameuse conférence *Apple Keynote* de 2007 par Steve Jobs où il annonçait la sortie du premier iPhone, personne n'aurait pu imaginer que les écrans tactiles se retrouveraient dans toutes les mains. L'iPhone était le précurseur du marché des smartphones qui a connu une des expansions les plus fulgurantes avec l'amélioration constante des infrastructures internet. Aurions-nous pu prévoir quels marchés allaient-être les plus réceptifs à l'arrivée de cette technologie? À l'aide de quels critères? Nous allons au fil de ce rapport nous questionner sur les corrélations possibles pour un pays donné entre le nombre de téléphones pour mille habitants et différents critères de développement.

Problématique et objectif

A partir d'un jeu de données issu de Kaggle, publié en 2017 par Fernando Lasso, intitulé "Countries of the world" et rassemblant des données du gouvernement américain (Anciennes données du CIA World Factbook et International Consortium for the Advancement of Academic Publication https://gsociology.icaap.org/dataupload.html), nous allons essayer d'expliquer la déjà existante fougue du téléphone mobile en 2000 en fonction de certains indicateurs.

Nous tenterons donc de répondre à la problématique suivante:

Aurions-nous pu anticiper la réussite du marché des téléphones portables dans un pays à l'aide des indicateurs classiques de développement? Si oui, à l'aide desquels? Si le temps nous le permets, nous comparerons notre analyse avec nouvelles données plus récentes du World Factbook.

I - Analyse descriptive des données

A - Analyse du jeu de données Le jeu de données que nous avons choisi contient les données de 227 pays par rapport à 19 variables. Dans le cadre de cette étude, nous nous limiterons aux variables suivantes: la région d'appartenance du pays, la densité de population (nombre habitants par miles carré), le PIB par habitant et, enfin, la part de la population active travaillant dans l'agriculture, l'industrie et les services.

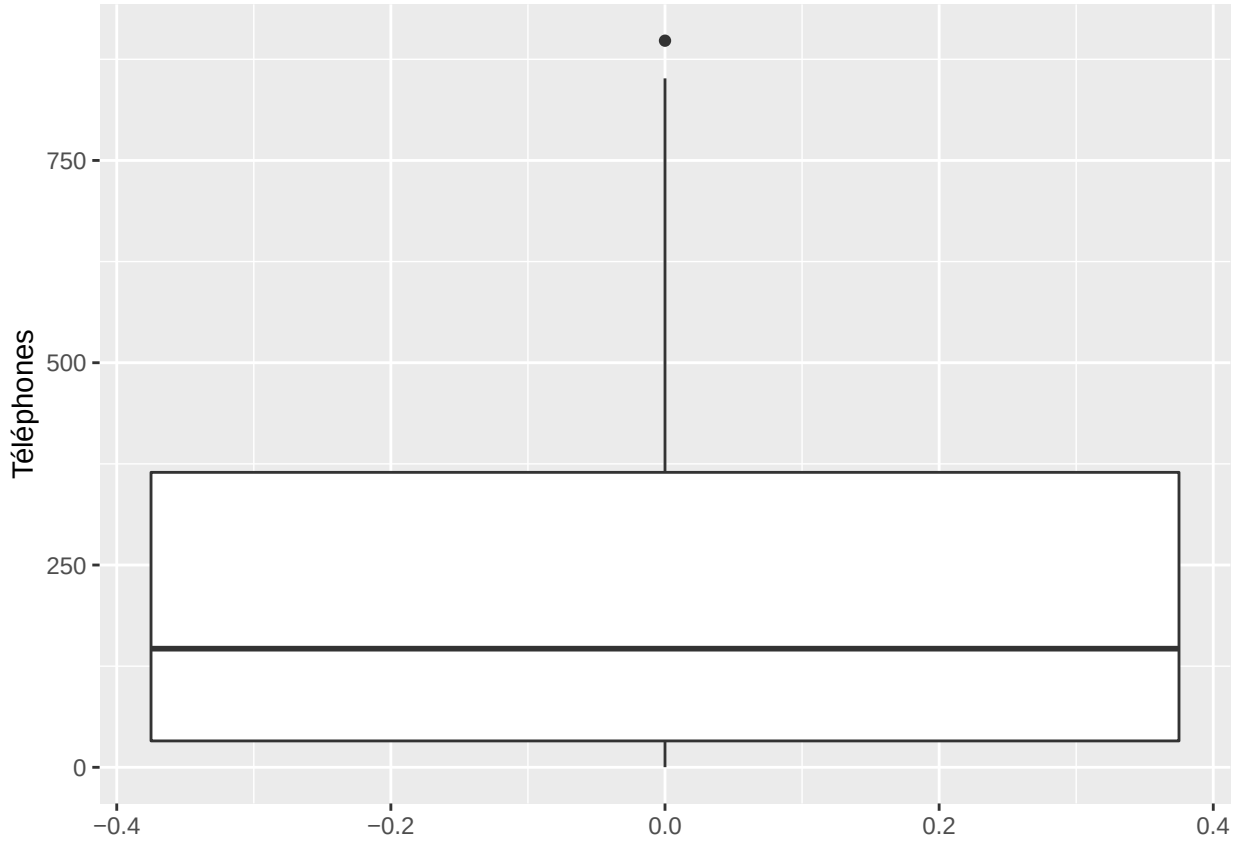
```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

Pour être en adéquation avec les consignes du projet, nous avons choisi 6 variables d'intérêt, puisqu'il était recommandé d'en prendre 3 ou 4 minimum. Nous avons également renommé les noms des colonnes du fichier en français. Nous nous sommes rendu compte que notre variable d'intérêt possédait 4 valeurs manquantes. C'est également le cas pour certaines de nos variables explicatives. Nous avons décidé de retirer les pays où il manque des données de notre étude. Nous avons donc gardé 209 pays. Il était demandé d'avoir au minimum 100 données, ce qui est donc le cas dans notre étude.

B - Analyse univariée de la variable d'intérêt Introduisons des notations: T, pour téléphones, est la variables à expliquer, c'est-à-dire le nombres de téléphones pour mille habitants. T est la réalisation

d'une variable aléatoire X. Dans cette première partie, nous allons analyser les données de T et essayer de déterminer la loi de X. Commençons par regarder quelques caractéristiques de T:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2    32.6   146.6   222.6   364.5   898.0
```



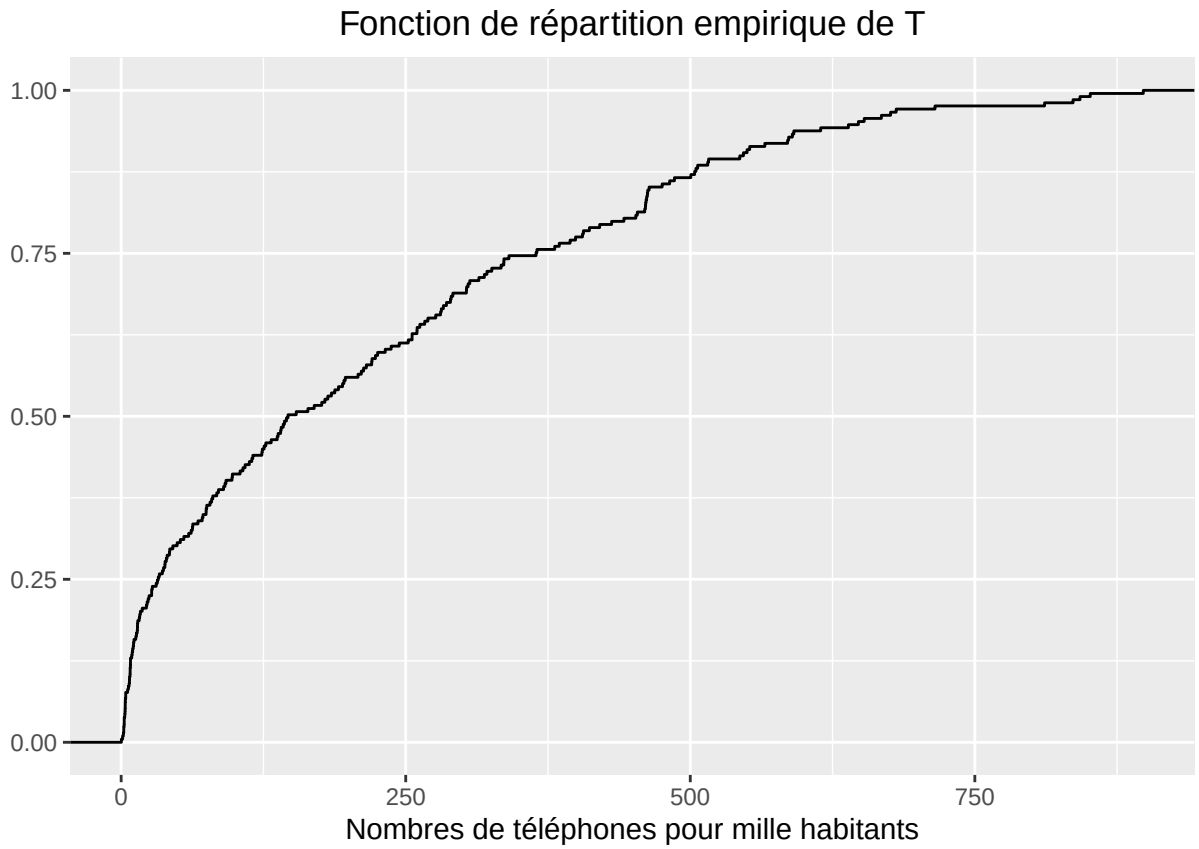
On peut commencer à analyser les indicateurs statistiques de notre échantillon. Par lecture de notre sommaire et de notre box-plot, on on obtient que notre loi n'est pas symétrique, puisque la moyenne, 222,6 est différente de la médiane, 146.6.

Nos deux indicateurs de tendance centrales sont assez bas. Notre premier quartile est à 32.6 mais surtout notre troisième quartile est égal à 364.5. Cela signifie que 75% des pays ont moins de 364.5 téléphones pour mille habitants.

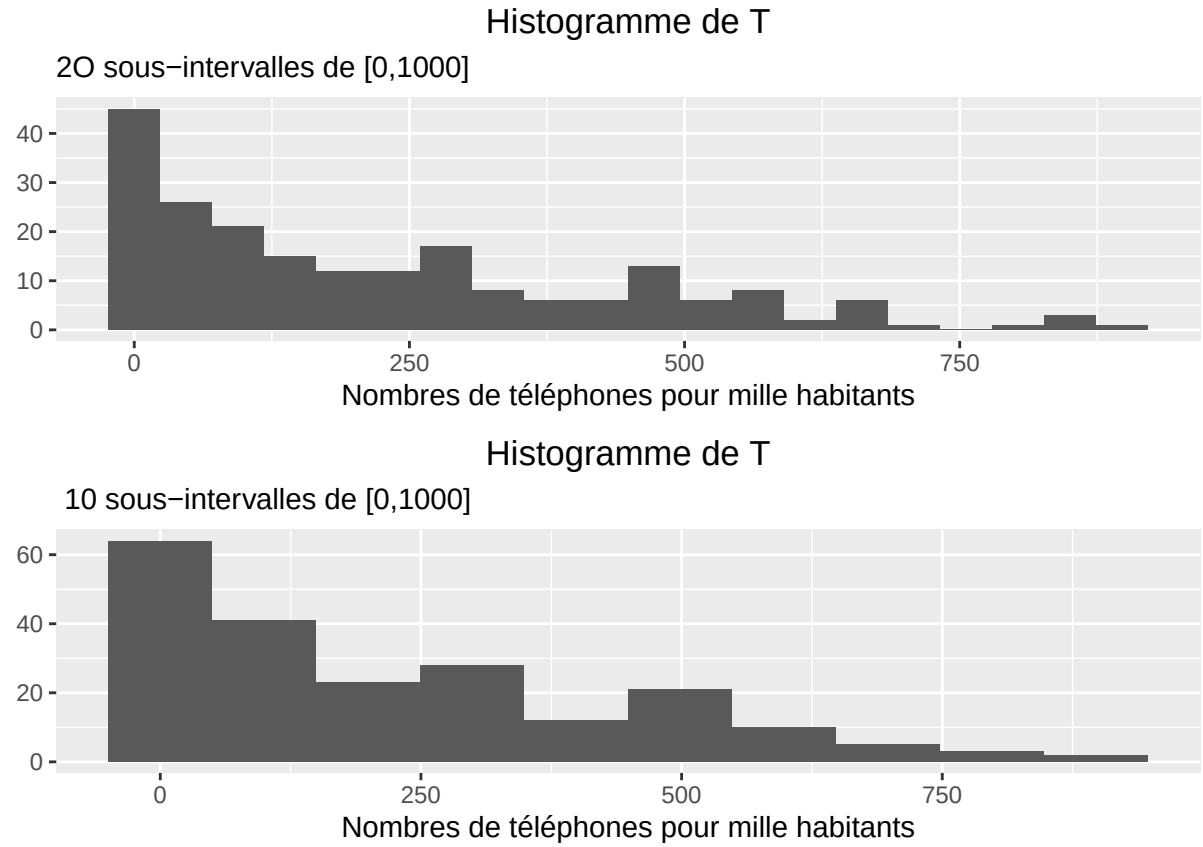
Notre étendue est de 897.8, mais l'étendue est sensible aux valeurs aberrantes et nous voyons dans notre box-plot que nous en avons une. De plus, la longueur d'une des moustaches est assez élevée. L'écart-interquartile, plus robuste que l'étendu, est de 331.9. Ainsi, la moitié des pays ont entre 32.6 et 364.5 téléphones pour mille habitants.

Nous allons pour l'instant garder l'observation "hors norme" mais nous reviendrons possiblement dessus dans d'autres parties du rapport.

Avant toutes recherches plus précises sur notre variable T, nous allons nous convaincre qu'elle est continue. Pour ce faire, traçons sa fonction de répartition empirique.



On voit que la fonction de répartition empirique de T réalise beaucoup de sauts. Elle en réalise exactement 201. 201 est "grand" devant 209 donc notre variable d'intérêt est continue. Nous pouvons alors tracer son histogramme.

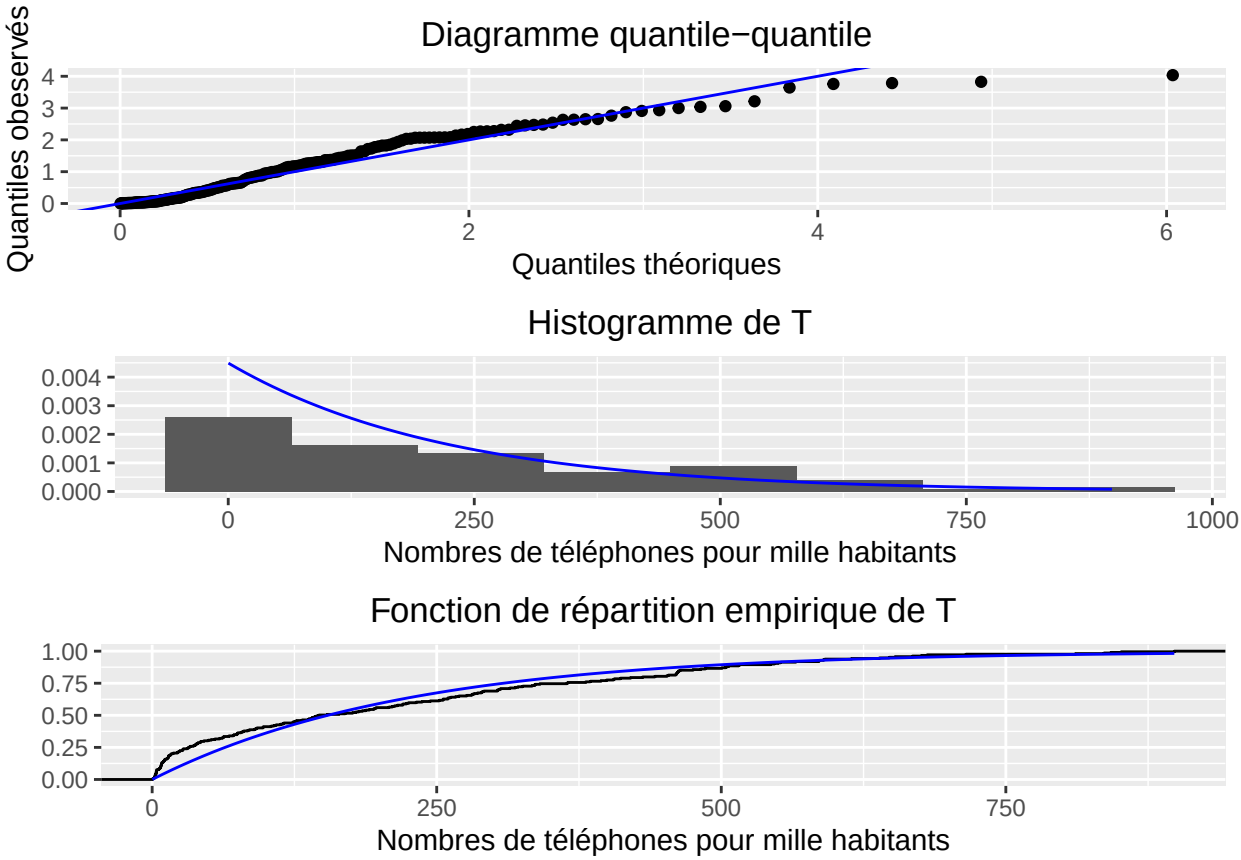


Les histogrammes sont uni-modaux avec un pic dans en 0, puis décroissant. On en déduit que dans notre jeu de données, beaucoup de pays ont moins de 100 téléphones pour mille habitants et à l'inverse, peu de pays on plus de de 700 téléphones pour mille habitants.

Au vue de ces histogrammes, nous pouvons penser que T est un échantillon de loi exponentielle, c'est-à-dire X suit une loi exponentielle de paramètre λ. L'espérance d'une loi exponentielle est $\frac{1}{\lambda}$. La méthode des moments nous propose la moyenne empirique de T comme estimateur sans biais, consistant et asymptotiquement normal de l'espérance.

Nous avons vu dans le sommaire plus haut que la moyenne empirique de T est $\hat{\mu} = 222.5574163$. En prenant $\hat{\lambda} = 0.0044932$, nous voulons vérifier que notre hypothèse est acceptable c'est à dire que la variables Téléphones suit une loi exponentielle de paramètre $\hat{\lambda} = 0.0044932$. Nous allons alors tracer:

1. L'histogramme de T et la densité $\hat{\lambda}e^{\hat{\lambda}x}$
2. La fonction de répartition empirique de T et la fonction de répartition de $e^{\hat{\lambda}x}$.
3. Le QQ-plot entre les quantiles théoriques $e^{\hat{\lambda}x}$ et les quantiles empiriques de T.



Tout d'abord, les points de notre QQ-plot s'alignent bien sur la première bissectrice. C'est encourageant. De même, notre histogramme approche la courbe de densité de notre loi théorique et notre fonction de répartition fait de même. L'adéquation n'est pas parfaite mais elle est suffisamment raisonnable pour retenir l'hypothèse que T suit une loi exponentielle de paramètre $\hat{\lambda}$.

Nous allons finaliser notre analyse uni-variée en réalisant un test de Kolmogorov-Smirnov sur notre échantillon, puisque la variables Téléphones et loi exponentielle sont toutes deux des lois continues.

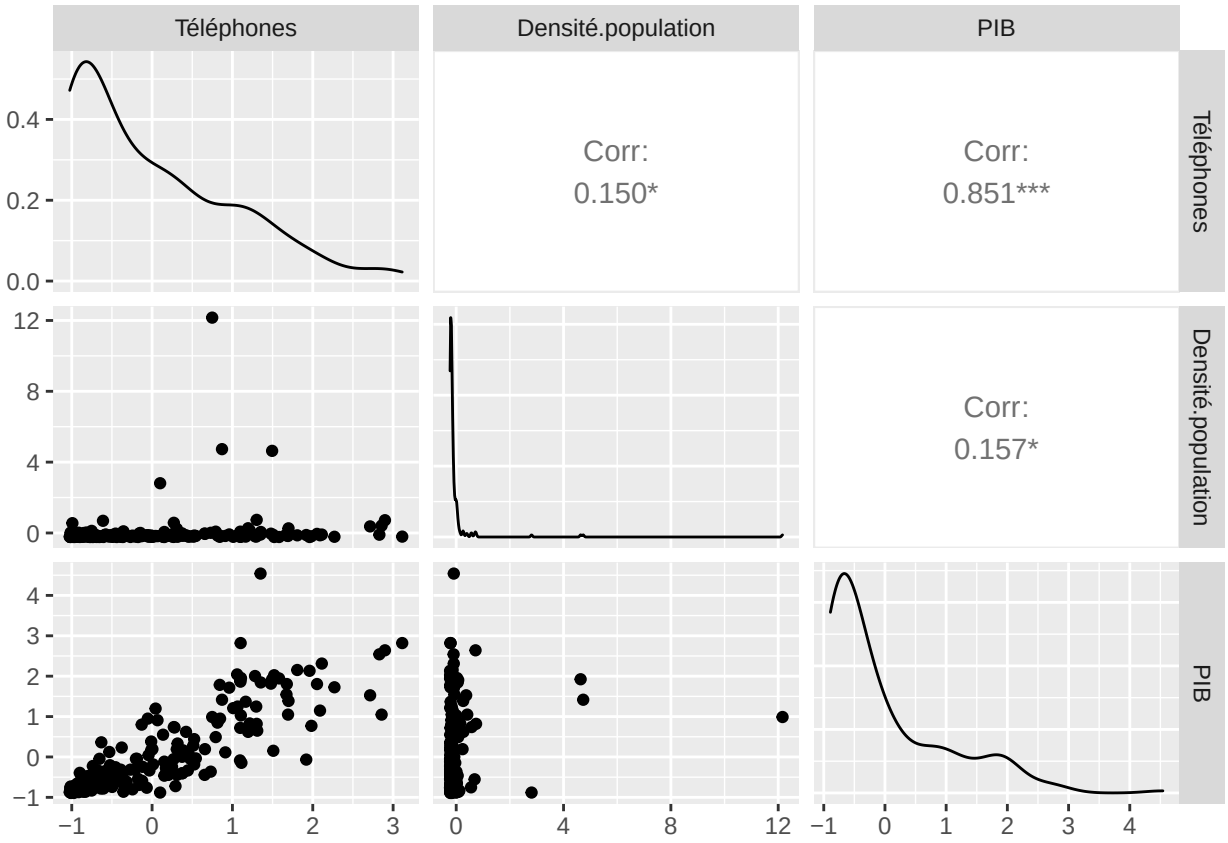
```
Warning in ks.test(df.countries$Téléphones, rexpsample): p-value will be
approximate in the presence of ties

    Two-sample Kolmogorov-Smirnov test

data:  df.countries$Téléphones and rexpsample
D = 0.93301, p-value < 2.2e-16
alternative hypothesis: two-sided
```

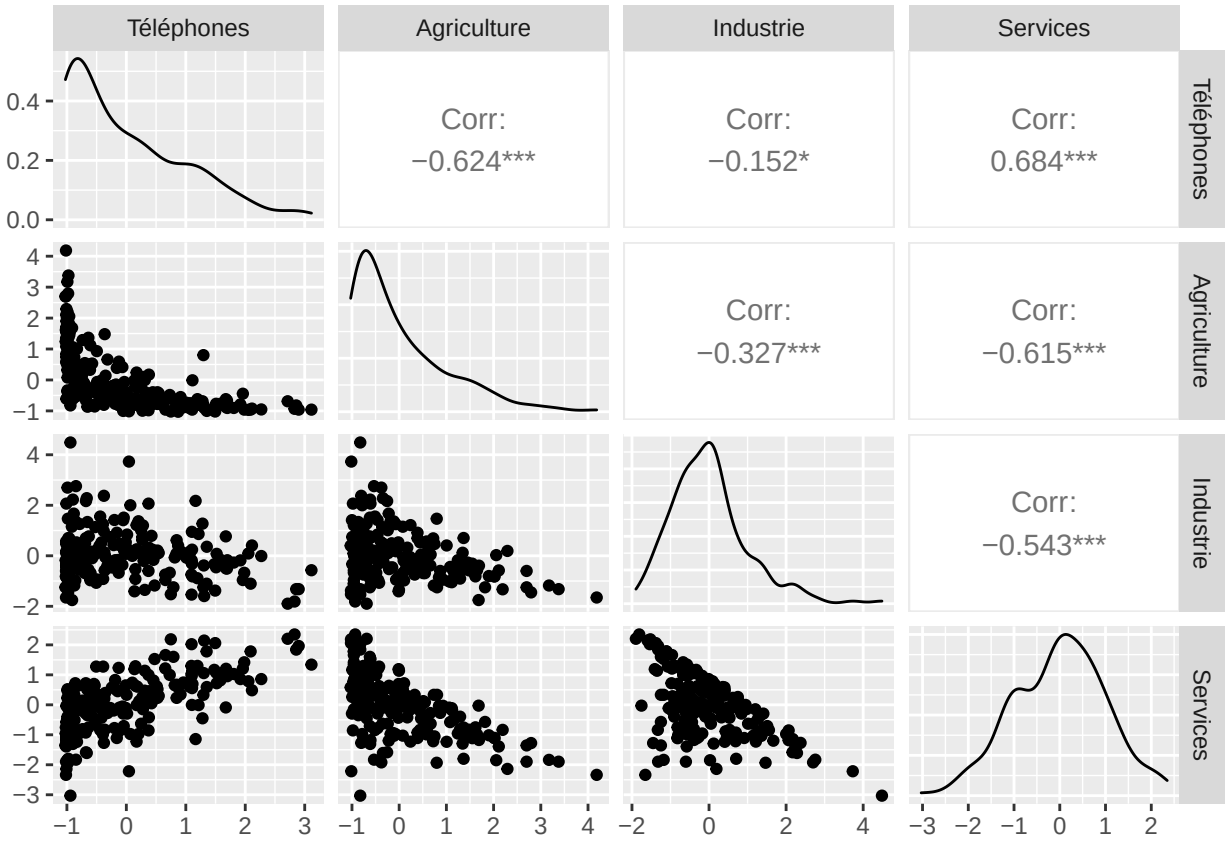
C - Analyse bi-variée Nous allons dans cette partie nous intéressée aux potentielles corrélation entre notre variables d'intérêt T et les autres variables de notre jeu de données.

Commençons par calculer la matrice de corrélation et présentons un diagramme de dispersion de toutes les paires de variables, pour chaque variable quantitative de notre jeu de données.

```
                  Téléphones Densité.population         PIB Agriculture
Téléphones         1.0000000          0.1496463  0.85052424  -0.6238864
Densité.population 0.1496463          1.0000000  0.15709460  -0.1492505
PIB                0.8505242          0.1570946  1.00000000  -0.5940198
Agriculture       -0.6238864         -0.1492505 -0.59401976   1.0000000
```

```
Industrie           -0.1524518         -0.1407955 -0.02783529  -0.3270188
Services             0.6839341          0.2510659  0.55256157  -0.6151106
                      Industrie   Services
Téléphones          -0.15245178  0.6839341
Densité.population -0.14079553  0.2510659
PIB                 -0.02783529  0.5525616
Agriculture         -0.32701879 -0.6151106
Industrie            1.00000000 -0.5429972
Services            -0.54299721  1.0000000
```





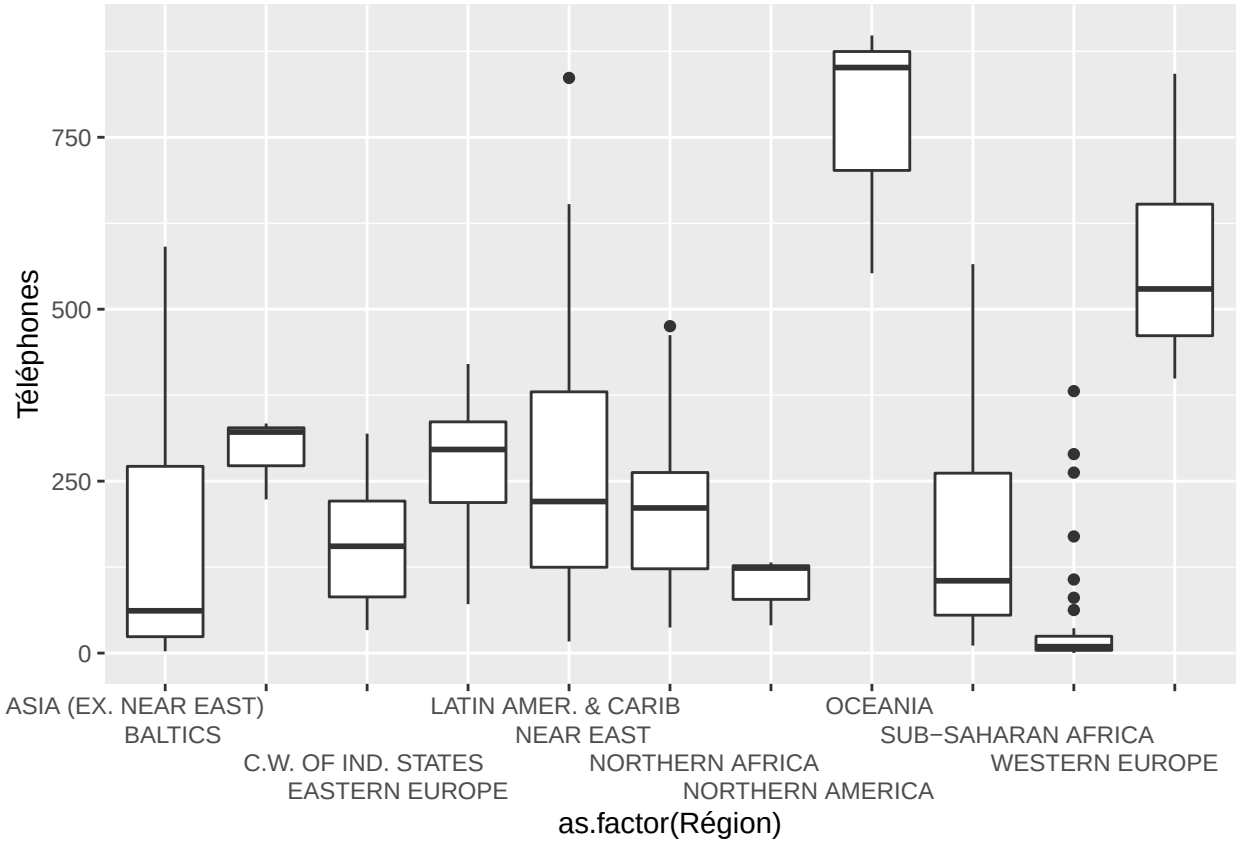
On a un facteur de corrélation d'environ 0.85 pour le couple de variables T et PIB; on en déduit qu'elles sont fortement corrélées linéairement. Il en est de même pour les variables Services et T avec un facteur de corrélation de 0.68. Finalement, la variable T est également corrélée négativement à la variable Agriculture avec un coefficient de corrélation de -0,62.

En analysant maintenant les nuages du points des paires de variables, on en déduit que les couples :

- (T, Densité de population): les points ne sont pas du tout alignés
- (T, PIB): les points relativement alignés ce qui était prévisible par l'analyse du coefficient de corrélation de ces deux variables.
- (T, Agriculture):
- (T, Industrie):
- (T, Services):

Nous avons maintenant une idée plus claire des liens entre notre variables d'intérêt et les variables quantitatives de notre jeu de données.

Nous allons terminer cette partie en nous intéressant à la variables Région, qui est une variable qualitative. Traçons le boxplot de la variable T pour chaque modalité de la variable Région.



Les boxplots sont extrêmement différents, on peut penser que la région d'appartenance du pays joue un rôle important sur le nombre de téléphones pour mille habitants. On remarque qu'en Amérique du nord et en Europe de l'ouest, les médianes sont hautes alors qu'elles sont beaucoup plus basses en Afrique Subsaharienne, en Océanie ou encore en Asie. Il y a parfois des observations hors-norme, notamment en Afrique sub-saharienne. Nous reviendrons sur ces observations plus loin dans le rapport.

II- Régression linéaire multiple

Dans cette partie, nous allons nous intéresser à une régression linéaire permettant d'exprimer le nombre de téléphone pour mille habitants parla densité de population (nombre habitants par miles carré), le PIB par habitant et enfin la part de la population active travaillant dans l'agriculture, l'industrie et les services.

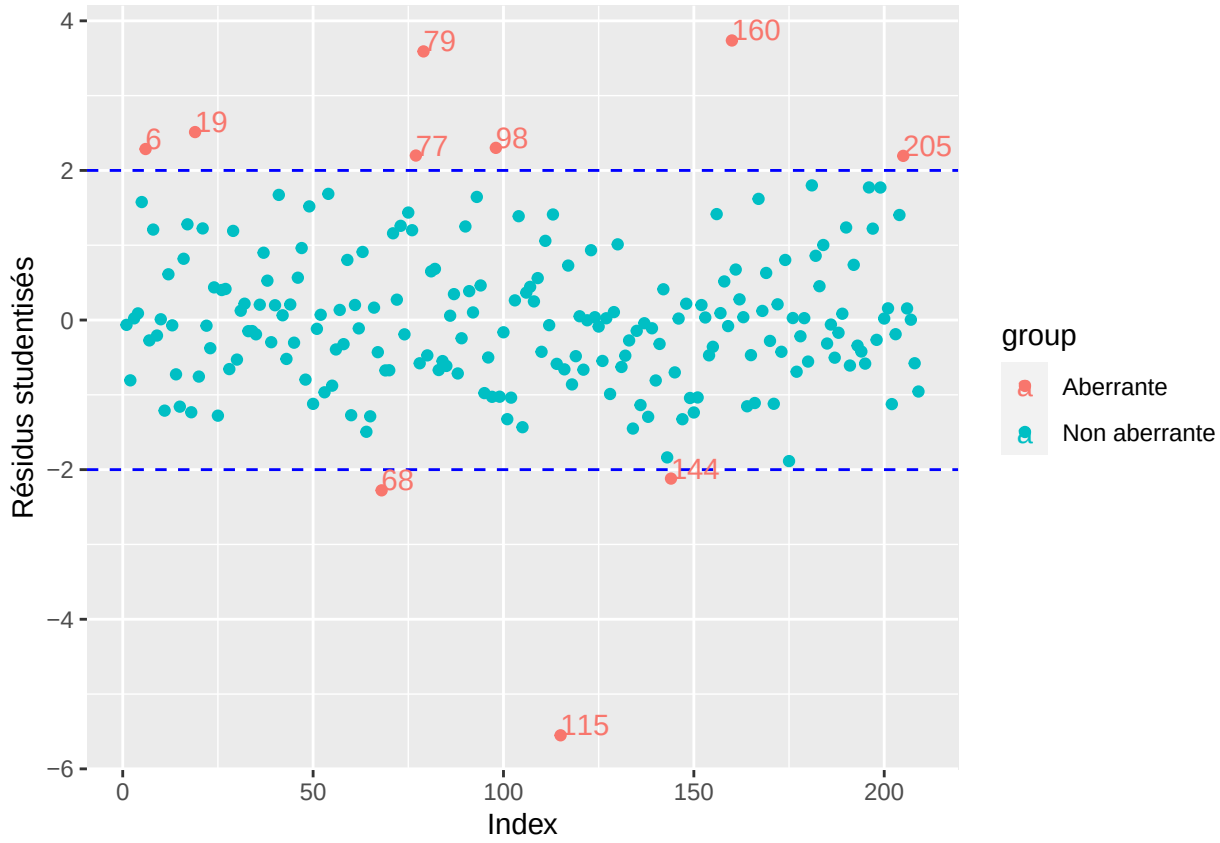
On considère le modèle linéaire suivant: Téléphones $= \beta_0 + \beta_1$PIB $+ \beta_2$Densité $+ \beta_3$ Agriculture $+ \beta_4$Industrie $+ \beta_5$Services $+ \varepsilon$ Commençons par centrer et réduire les données pour qu'elles soient à la même échelle. Puis, réalisons une première régression linéaire.

```
Call:
lm(formula = Téléphones ~ Densité.population + PIB + Agriculture +
    Industrie + Services, data = df.countries)

Residuals:
    Min      1Q  Median      3Q     Max
-483.00  -57.60   -6.37   45.98  359.35

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.225e+03  1.053e+03  -1.162    0.246
```

```
Densité.population -6.134e-03  5.478e-03  -1.120    0.264
PIB                 1.421e-02  8.975e-04  15.831   <2e-16 ***
Agriculture         1.025e+03  1.053e+03   0.974    0.331
Industrie           1.110e+03  1.052e+03   1.054    0.293
Services            1.496e+03  1.057e+03   1.416    0.158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99.82 on 203 degrees of freedom
Multiple R-squared:  0.7934,    Adjusted R-squared:  0.7883
F-statistic: 155.9 on 5 and 203 DF,  p-value: < 2.2e-16
```

Avant d'aller plus loin dans la régression linéaire et la sélection de variables nous allons nous intéresser à notre jeu de données, voir s'il comporte des valeurs aberrantes, des points leviers et analyser les distances de Cook.

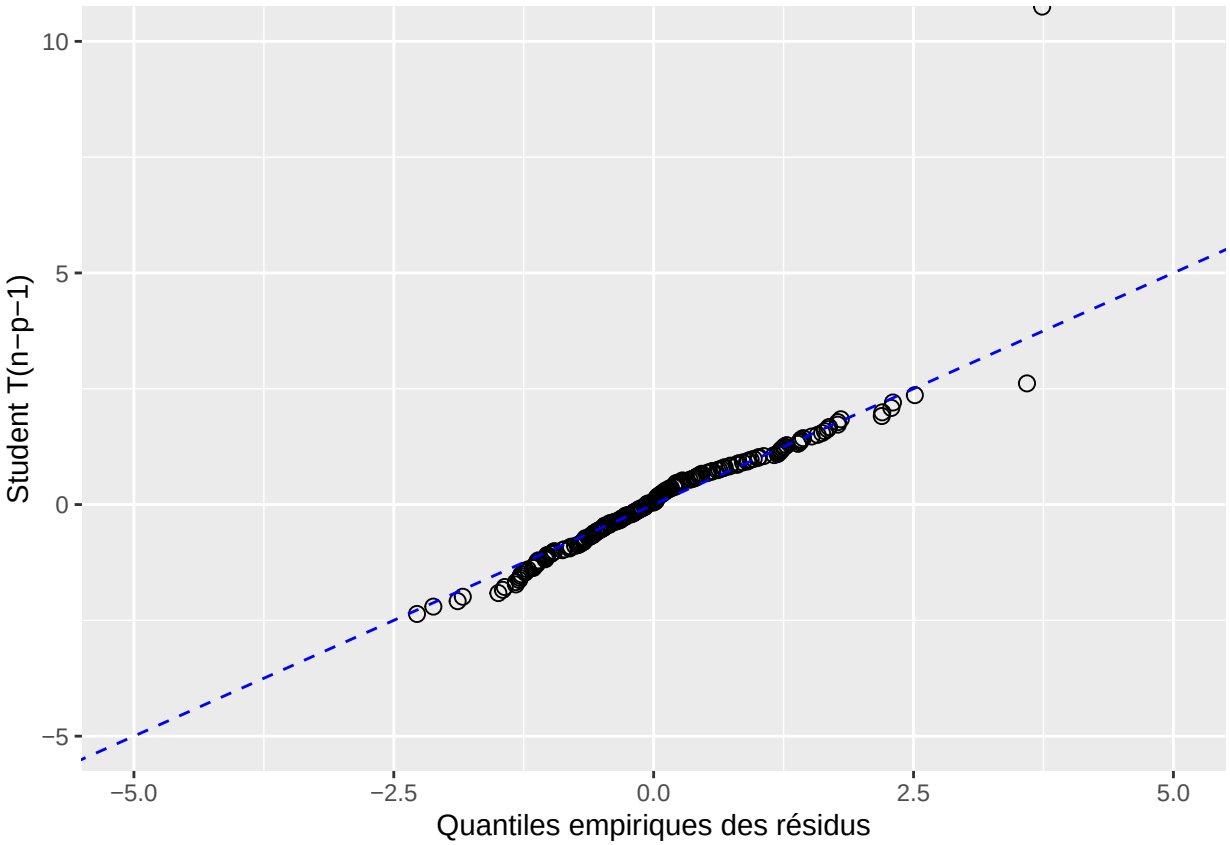
1) Valeurs aberrantes



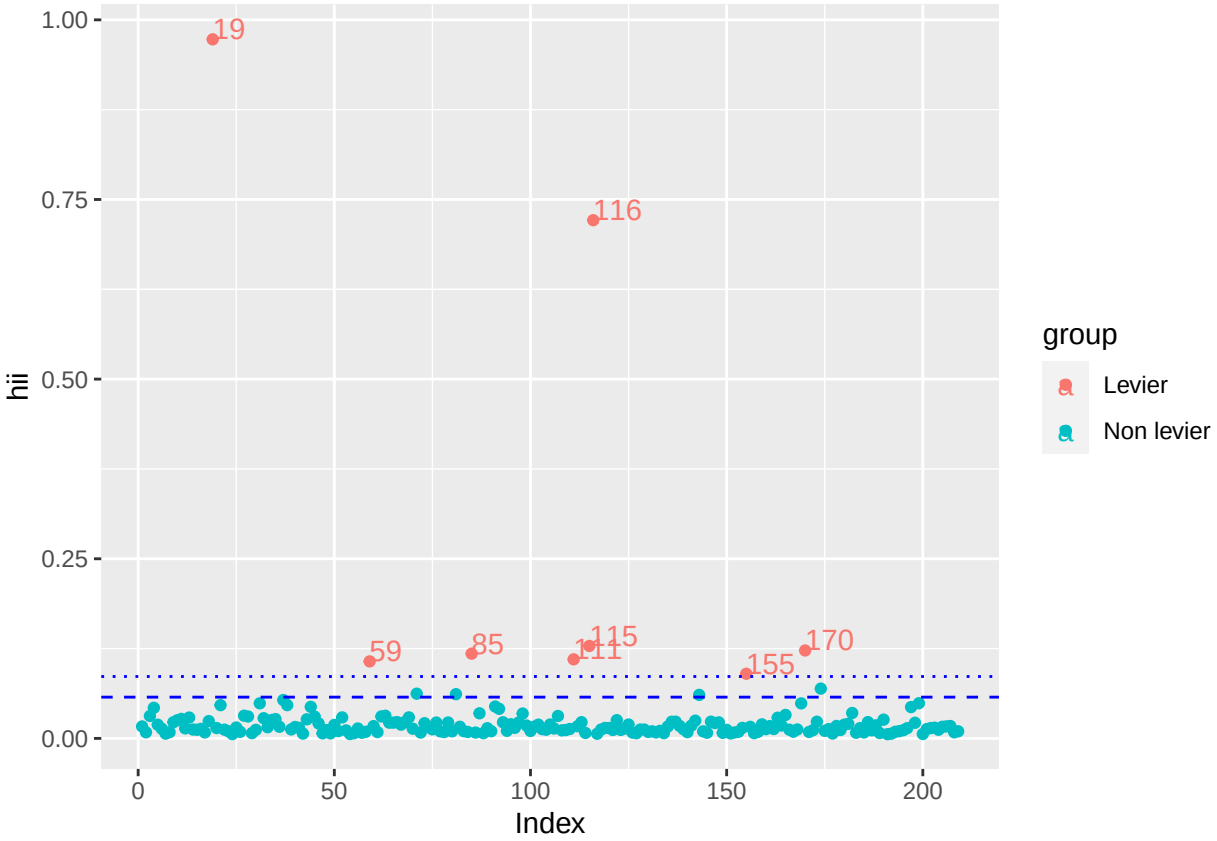
On a 10 valeurs aberrantes dan un échantillon de taille 209.

2) QQ-plot

```
Warning: Removed 1 rows containing missing values (geom_point).
```

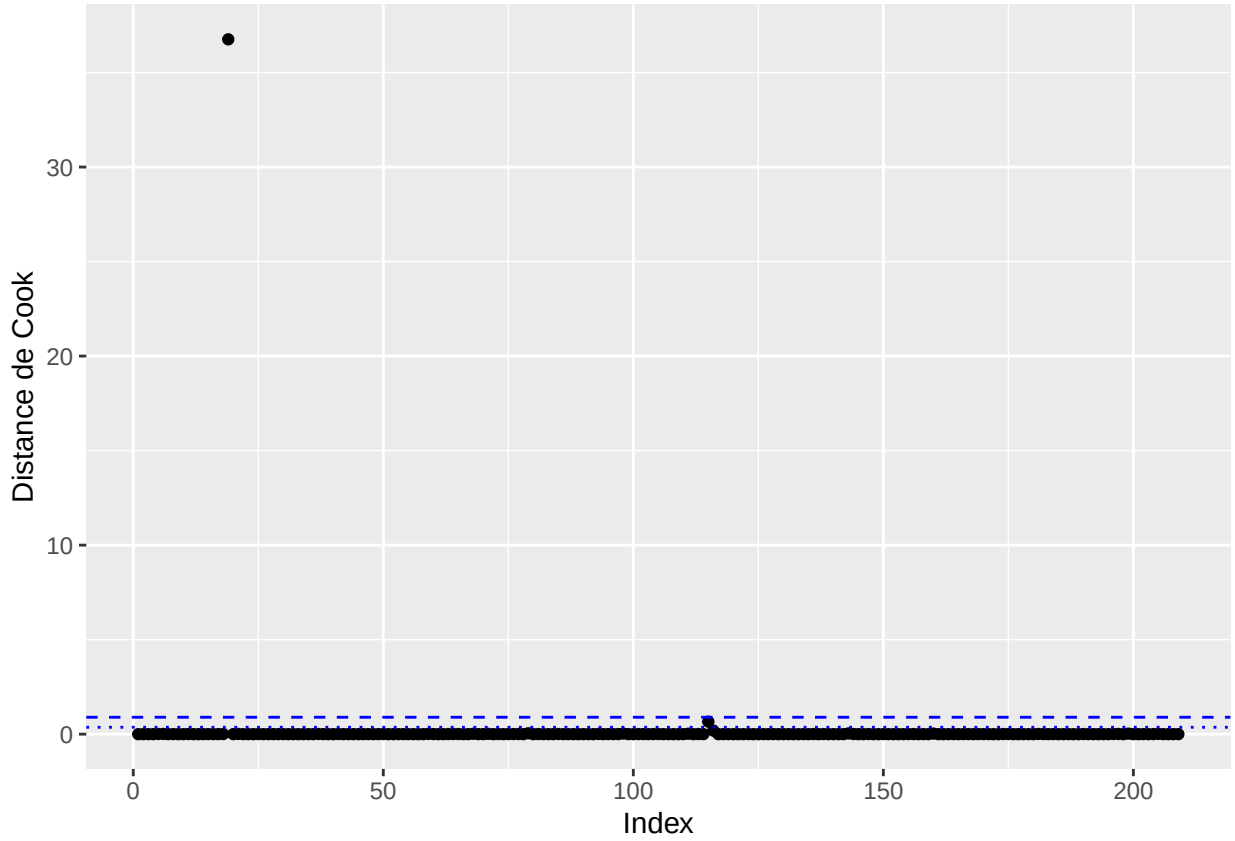


3) Points levier



Il y a 7 valeurs qui dépassent les deux seuils.

4) Distance de Cook

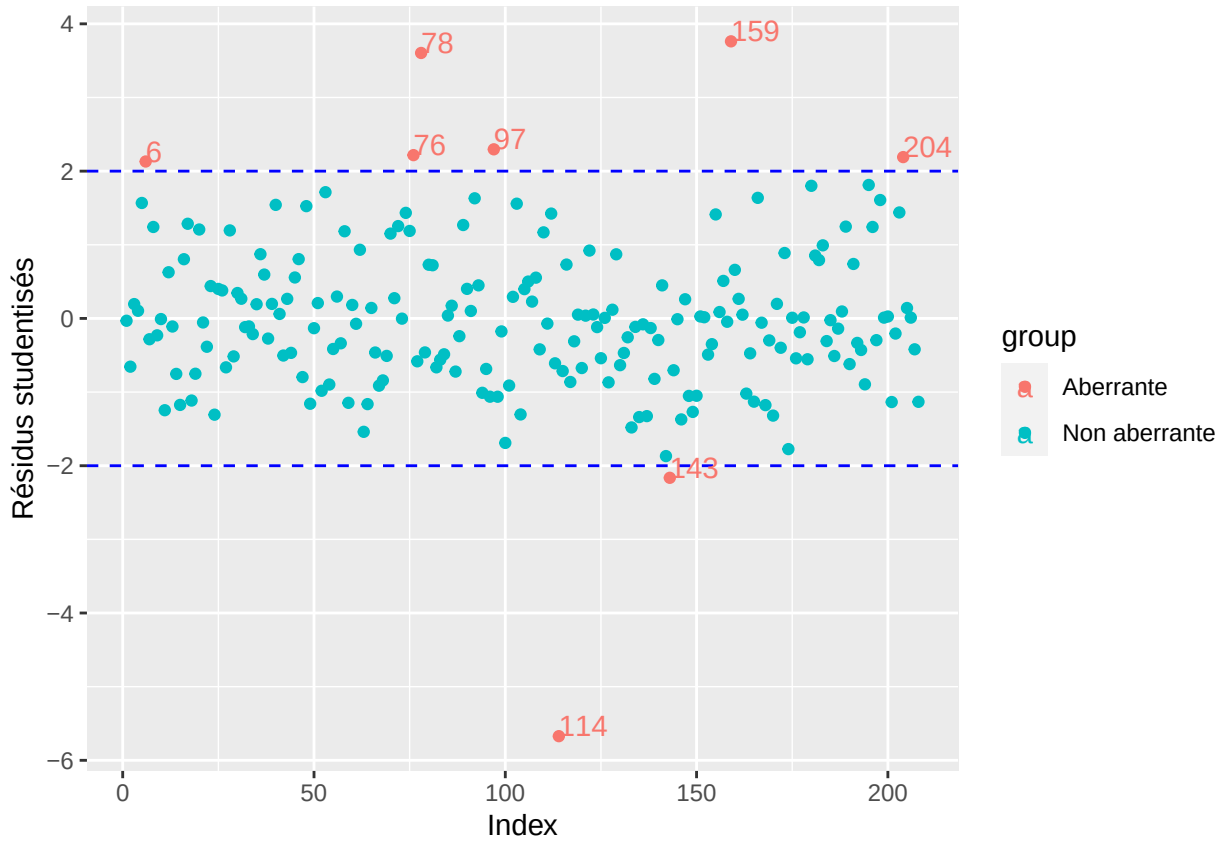


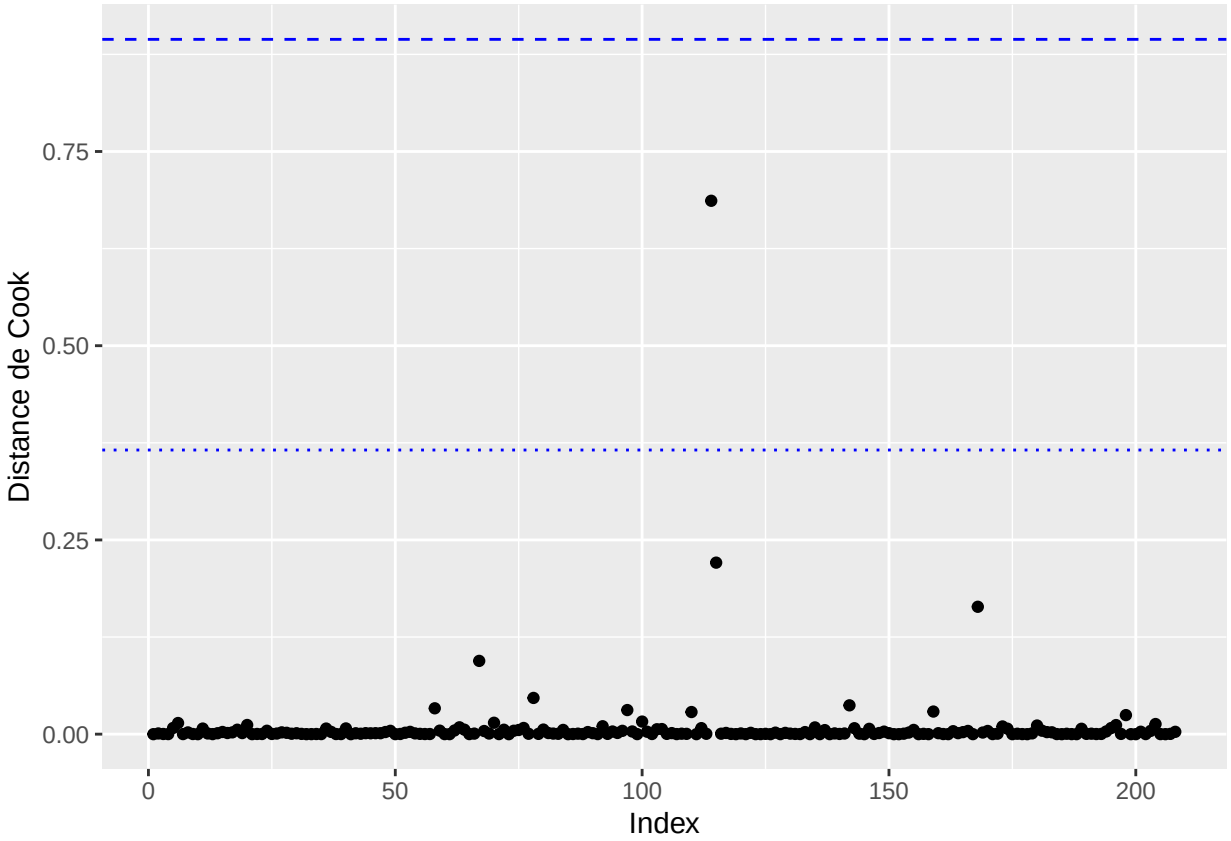
On observe qu'une valeur dépasse le seuil préoccupant du quantile $f^n_{n-p}(0.5)$.

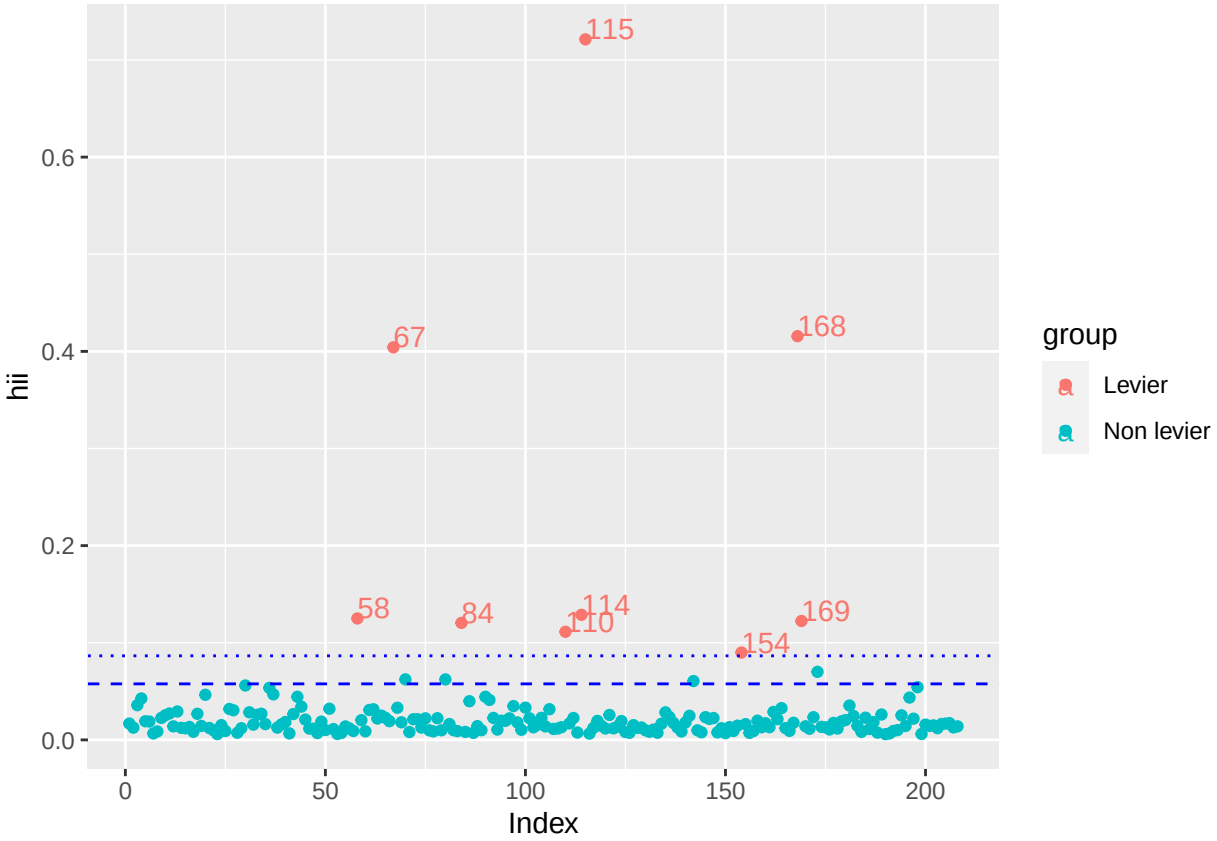
```
[1] 0.9728571
```

```
[1] 36.75434
```

```
Belize
     19
```


Résidus studentisés
Index
group
Aberrante
Non aberrante
78
159
6
76
97
204
143
114
4
2
0
−2
−4
−6
0
50
100
150
200


Distance de Cook
0.75
0.50
0.25
0.00
0
50
100
150
200
Index


115
67
168
58
84
114
110
169
154
0.6
0.4
0.2
0.0
hii
0
50
100
150
200
Index
group
Levier
Non levier