BAUDU APOLLINE - BERTRAND MORGANE - RODRIGUEZ MANUELA - VITET AURELIEN

PROMO A - GROUPE 5-1

Méthodes Statistiques pour l'Ingénieur

Projet MSI « Open data »

UNIVERSITÉ DE LYON

ENTPE

Sommaire

I- Présentation de l'étude générale

L'enquête que nous avons prise pour réaliser ce projet s'intitule « Histoire de vie », réalisée entre février et avril 2003 par le biais d'un questionnaire (cf. lien dans la bibliographie). L'échantillon est constitué de 13 500 personnes de plus de 18 ans vivant en France métropolitaine. Cette enquête, par le biais de critères objectifs et subjectifs a pour but de comprendre comment l'individu s'intègre dans la société grâce aux différents liens sociaux qu'il a créé tout au long de sa vie. Elle porte sur la construction de l'identité de l'individu suivant différents critères sociaux.

a) Historique et auteurs de l'étude

Cette enquête a été réalisée dans la continuité d'enquêtes précédentes. Ainsi, elle suit l'enquête « Mobilité géographique et Insertion sociale » (MGIS) réalisée en 1992 par l'Insee et l'Ined. Elle avait été réalisée pour aborder la question de l'insertion sociale via la méthode de l'enquête statistique par sondage.

Ainsi, l'échantillon de 13 500 personnes, décrit comme aléatoire, de l'enquête « Histoire de vie » de 2003, n'a pas été tiré dans l'échantillon maitre mais dans la base de données ayant servie à cette précédente enquête de 1992. Cela a été fait par choix afin de sur-représenter certaines populations.

Afin de réaliser un sondage le plus qualitatif possible, une enquête pilote a été réalisée en 2002 sur 800 personnes. Le but était d'observer si les enquêtés acceptaient de répondre aux questions posées et de se rendre compte de la praticité du sondage pour les enquêteurs. Les conclusions ont permis d'améliorer le sondage de l'enquête officielle de 2003, réalisée sur les 13 500 personnes.

La responsable de cette enquête est la direction des statistiques démographiques et sociales qui gère le répertoire national d'identification des personnes physiques ainsi que le fichier national des électeurs. Elle s'occupe de réaliser des études concernant les populations et les ménages sur différents critères sociaux (emploi, revenus, patrimoine, consommation…).

Le producteur et le diffuseur de cette enquête sont respectivement l'INSEE et l'ADISP (Archives de Données Issues de la Statistique Publique). On peut d'ailleurs retrouver l'enquête sur le site de l'INSEE via le lien suivant :
https://www.insee.fr/fr/statistiques/2532244#consulter

Cette enquête « Histoire de vie » a également des partenaires comme le ministère des affaires sociales (Drees), le ministère du travail et de l'emploi (Dares), le ministère de la culture (DEP) et la Délégation interministérielle à la ville (Div) ou l'Inserm.

b) Déroulement de l'étude

L'enquête réalisée se base sur un échantillon de 13 000 individus vivants en France métropolitaine et d'au moins 18 ans. L'enquête consiste à répondre à un questionnaire (cf. lien dans la biographie), à domicile en présence de l'enquêteur pour une durée d'environ 80

minutes. Le questionnaire comporte 808 variables portant sur différents thèmes à caractéristiques sociales.

L'enquête s'étend sur 2 mois, de février à avril 2003, et la fréquence des récoltes de données se fait de façon apériodique ou ponctuelle.

Dans 85% des cas, les entretiens ont été réalisés en une seule visite. Malheureusement, 16% des personnes interrogées n'ont pas souhaité répondre au questionnaire. De plus, suite aux entretiens, « seuls » 62% des questionnaires ont été remplis de façon totale et sont exploitables. Cela reste néanmoins un bon résultat pour une enquête sociale de ce type. Ainsi, ce sont 8 403 questionnaires qui sont utilisables, ce qui représente un échantillon assez conséquent.

c) Description des données

L'enquête contient 808 variables qui permettent de mettre en avant différents aspects de la vie sociale des personnes : vie familiale, emploi, loisirs, amis, lieu de vie, santé…

Les auteurs de cette enquête sont partis du fait que pour avoir des résultats significatifs, il était nécessaire d'avoir des données objectives telles que la situation professionnelle, l'état de santé, les centres d'intérêts, le lieu de vie, … afin de pouvoir faire des analyses chiffrées. Cependant, il était également important d'avoir des données subjectives comme des ressentis ou des sentiments. Ainsi, la base de données prend par exemple en compte l'importance du travail dans la vie de l'enquêté, son ressenti vis-à-vis d'évènements douloureux (discrimination, mise à l'écart, moqueries…), les liens qu'il entretient avec ses parents ou encore la façon dont il se considère dans la vie.

Si nous devions créer des groupes de variables, nous pourrions dire que les thèmes principaux sont les suivants : situation professionnelle (en activité, chômage, au foyer, ...), parents, conjoints et enfants, origines, lieu de vie, loisirs, discriminations, santé, cercle amical, identité, mobilité et vie associative. Pour chaque domaine, de nombreuses variables permettent de décrire de nombreux critères et d'avoir une vision globale du domaine.

Pour résumer, le questionnaire permet de faire une biographie complète de la personne enquêtée en reprenant année par année sa situation familiale, professionnelle (avec les périodes d'inactivité) ou son lieu de vie. Il prend également en compte les évolutions subjectives avec l'identification au sein de la famille, dans la profession, les loisirs, la santé, le rapport au corps ou encore les discriminations subies.

d) Problèmes et objectif de l'étude

De par les variables présentes dans les données de l'enquête décrite ci-dessus, nous pouvons dire que l'objectif principal de cette étude est d'analyser l'insertion des individus dans la société grâce à l'observation des liens avec leur famille ou leurs collègues de travail, leur adhésion à des associations, leurs loisirs, la présence d'ami(e)s ou encore leur façon de se définir en tant qu'individu. Elle a également pour but de mettre en avant les difficultés que

certaines personnes peuvent rencontrer à cause de moqueries, de l'absence d'un cercle familial stable ou de liens sociaux, d'une situation professionnelle instable ou encore de maladies.

Plus tôt, nous avions mis en avant le fait que l'échantillon de l'enquête « Histoire de Vie » ne provenait pas de la base maitresse, mais de la base de données ayant servie à une autre étude. Le but est de favoriser la sur-représentation de certaines populations que l'on souhaite observer en priorité. C'est notamment le cas des personnes de moins de 60 ans qui souffrent de problème de santé et des personnes nées en France dont un parent est né à l'étranger.

Si les personnes âgées de moins de 60 ans ayant un problème de santé sont bien sur-représentées dans l'échantillon, la proportion de personnes nées en France d'un parent étranger reste inférieure aux attentes.

Nous pouvons donc dire que le but de cette étude est d'observer l'insertion sociale des personnes ayant des problèmes de santés ainsi que des personnes d'origine étrangère.

De notre côté, nous souhaitions nous concentrer sur l'impact de la composition du foyer familial et des liens familiaux sur l'âge du départ de l'individu du domicile familial. Ainsi, notre problématique est la suivante : « Existe-t-il un lien entre la composition du foyer familial et l'âge de départ de l'individu de celui-ci ? ».

Nous souhaitons monter qu'il existe un lien entre le fait d'avoir été élevé par ses parents biologiques ou non, le lien sentimental envers eux mais aussi le nombre de frères et sœurs avec l'âge de départ du domicile familial. Pour cela, nous mettrons en parallèle les enquêtés élevés par leurs parents biologiques et ceux élevés par un couple ne comportant aucun des deux parents biologiques. Nous ferons ensuite varier les autres variables afin de voir leurs impacts.

II- Méthode et données

a) <u>Préparation des données</u>

Du fait des 808 variables différentes, nous avons dû sélectionner les données nécessaires à la résolution de notre problématique.

Pour rappel notre problématique est la suivante : « Existe-t-il un lien entre la composition du foyer familial et l'âge de départ de l'individu de celui-ci ? ».

Nous avons ainsi retenu les variables suivantes :

- BQUITE : Age (révolu au moment du départ) auquel l'enquêté a quitté définitivement le domicile des parents pour plus d'un an. (Y)
- BELEVQ1 : Elevé par les deux parents biologiques en couple jusqu'à 18 ans. (X_1)
- BELEVQ6 : Elevé par un couple ne comprenant aucun des parents biologiques (enfant confié ou adopté) jusqu'à 18 ans. (X_2)
- BPAREQ1 : Considère ses deux parents biologiques, comme ses parents. (X_3)
- BPAREQ6 : Considère le couple par qui l'enquêté a été élevé, comme ses parents. (X_4)
- BNBFS : Nombre de frères et sœurs et demi-frères et demi-sœurs. (X_5)

Afin de garder uniquement les variables nécessaires à notre étude, nous avons dû faire plusieurs manipulations sur le logiciel R. Dans un premier temps, nous avons dû fusionner les données car les 808 variables étaient réparties dans 3 fichiers différents. La seule colonne commune aux trois tableaux est la première intitulée « IDENT ». Voici le script commenté qui a permis de réaliser la fusion :

```
#fusion des trois tableaux de données selon la première colonne qui est commune
DATA1<-merge(hdv1,hdv2, by="IDENT.C.8")
DATA<-merge(DATA1,hdv3, by="IDENT.C.8")
```

Une fois la fusion réalisée, il est nécessaire de garder uniquement les variables dont nous avons besoin. Voici le script qui nous a permis de réaliser cette tâche :

```
#Selection des données sur lesquelles nous allons travailler
A = DATA[, c('BQUITE.N.4.0', 'BELEVQ1.C.1', 'BPAREQ1.C.1', 'BNBFS.N.4.0', 'BELEVQ6.C.1','BPAREQ6.C.1')]

#Suppression des lignes ne comportant pas de données (des NA)
A<-subset(A, A$BQUITE.N.4.0!="NA")
A<-subset(A, A$BELEVQ1.C.1!="NA")
A<-subset(A, A$BPAREQ1.C.1!="NA")
A<-subset(A, A$BNBFS.N.4.0!="NA")
A<-subset(A, A$BELEVQ6.C.1!="NA")
A<-subset(A, A$BPAREQ6.C.1!="NA")
```

Nous créons ensuite un nouveau tableau « DATA » comportant uniquement les colonnes dont nous avons besoin ; c'est-à-dire celles qui correspondent aux variables que nous avons énoncé précédemment. La fonction « SUBSET » permet quant à elle de supprimer les lignes qui n'ont pas de réponse (« NA »), car celles-ci nous gêneraient dans la réalisation de nos régressions linéaires (nécessité de données chiffrées).

b) <u>Statistiques descriptives</u>

Dans un premier temps, nous avons réalisé une analyse descriptive pour notre variable Y, qui est l'âge de départ du domicile familial. Pour cela, nous avons utilisé la fonction « summary ».

Nous avons également fait une analyse descriptive toujours sur l'âge de départ du domicile familial mais selon si l'enquêté a été élevé par ses deux parents biologiques en couple jusqu'à ses 18 ans (X_1), ou par un couple ne comportant aucun de ses parents biologiques (X_2).

Les scripts ainsi que les résultats sont présents dans la partie résultat (III-a).

c) <u>Analyse économétrique</u>

Pour l'analyse économétrique, nous avons réalisé des analyses linéaires primaires et des analyses linéaires multiples. Le but de notre étude est d'observer s'il existe un lien entre l'âge de départ du domicile familial et la composition du foyer familial. Notre variable Y, qui doit être quantitative, est l'âge de départ de la maison (BQUITE).

- **Régressions linéaires primaires**

Nous avons réalisé trois régressions linéaires primaires. La première, est le lien entre l'âge de départ du foyer (Y) et le nombre de frères et sœurs dans le foyer (X_5 : BNBFS). La seconde, étudie le lien entre l'âge de départ du foyer (Y) et le fait d'avoir été élevé par ses deux parents en couple jusqu'à ses 18 ans (X_1 : BELEVQ1). La dernière étudie le lien entre l'âge de départ du foyer (Y) et le fait d'avoir été élevé par un couple qui ne comporte aucun des parents biologiques jusqu'à ses 18 ans (X_2 : BELEVQ6). Les trois régressions ont la même forme suivante :

$$\mathrm{Y} = \mathrm{a_i X_i} + \mathrm{b_i} + \xi \quad \text{, avec i = 1…3}$$

a_i : coefficient de la régression linéaire
$\mathrm{b_i}$: ordonnée à l'origine
ξ : coefficient d'erreur de la régression linéaire

Dans le script ci-dessous, nous pouvons observer le code pour les 3 régressions linéaires primaires :

```
18 #Régression linéaires primaires
19 x<-lm(A$BQUITE.N.4.0~A$BNBFS.N.4.0)
20 y0<-lm(A$BQUITE.N.4.0~A$BELEVQ1.C.1)
21 z0<-lm(A$BQUITE.N.4.0~A$BELEVQ6.C.1)
```

Dans un premier temps, il nous paraissait important de réaliser ces régressions linéaires primaires afin d'observer si ces critères seuls ont un lien avec l'âge de départ du domicile familial. Nous pensions qu'il était pertinent de regarder si le fait d'être élevé par ses parents biologiques ou non a un impact sur la stabilité du foyer et donc sur l'âge de départ du domicile familial. Nous voulions également faire le lien avec le nombre de frères et sœurs qui pourrait parfois causer des tensions familiales et donc favoriser un départ anticipé.

- **Régressions linéaires multiples**

Par la suite, nous avons combiné ces variables afin de voir si la combinaison de certaines variables avait un effet sur l'âge de départ du domicile familial.

Le premier bloque de régressions linéaires multiples combine le fait d'avoir été élevé par ses parents biologiques (X_1)ou non (X_2) et l'entente que l'enquêté a avec le couple qui l'a élevé (parents biologiques (X_3)ou non (X_4)). Ainsi, nous pouvons les écrire de la façon suivante :

$$Y = a_1 X_1 + a_3 X_3 + \mathrm{b} + \xi$$

$$Y = a_2 X_2 + a_4 X_4 + \mathrm{b} + \xi$$

Le script correspondant est le suivant :

```
24 #Régression linéaires multiples
25 y1<-lm(A$BQUITE.N.4.0~A$BELEVQ1.C.1+A$BPAREQ1.C.1)
26 z1<-lm(A$BQUITE.N.4.0~A$BELEVQ6.C.1+A$BPAREQ6.C.1)
```

Le second bloque ajoute à chacune des deux régressions linéaires multiples la variable définissant le nombre de frères et sœurs ainsi que de demi-frères et demi-sœurs dans le foyer (variable X_5).

Nous pouvons ainsi les écrire de la façon suivante :

$$Y = a_1X_1 + a_3X_3 + a_5X_5 + b + \xi$$

$$Y = a_2X_2 + a_4X_4 + a_5X_5 + b + \xi$$

III- Résultats

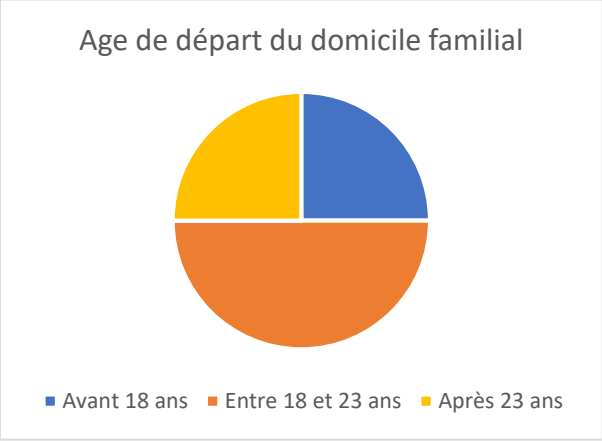
a) Analyse descriptive

Dans un premier temps, il était nécessaire de réaliser une analyse descriptive des variables quantitatives afin d'obtenir la valeur minimale, maximale, la valeur médiane, la moyenne ainsi que les quartiles. La fonction dont nous avons besoin pour obtenir ces données est la fonction « summary ». Voici les scripts que nous avons réalisé pour 3 variables différentes :

- La première variable correspond à l'âge de départ du foyer (variable originale).

```
> summary(A$BQUITE.N.4.0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   18.00   20.00   20.13   23.00   99.00
```

Nous pouvons observer que l'âge moyen de départ du foyer est de 20,13 ans et que la valeur médiane est de 20,0 ans. Ce sont 50% des enquêtés qui quittent le foyer familial entre 18,0 ans et 23 ans. Un quart des personnes ayant répondues au questionnaire quittent donc le foyer familial avant leur majorité et un autre quart après 23 ans.



Afin de déterminer si le contexte familial a un impact sur cet âge de départ, nous avons également réalisé les mêmes calculs suivant si l'enquêté a été élevé par ses parents biologiques ou non. Pour cela, nous avons dû créer deux variables d'après le script suivant :

```
38 #Création de sous-tableaux pour les boxplot
39 C<-subset(A, A$BELEVQ1.C.1 ==1) #sélection des personnes ayant été élevées par leurs parents
40 D<-subset(A, A$BELEVQ6.C.1 ==1) #sélection des personnes ayant été elevées par leurs parents adoptifs
```

- La seconde variable correspond à l'âge de départ du foyer pour les personnes élevées par leurs parents biologiques (cf. le script de création de cette variable).

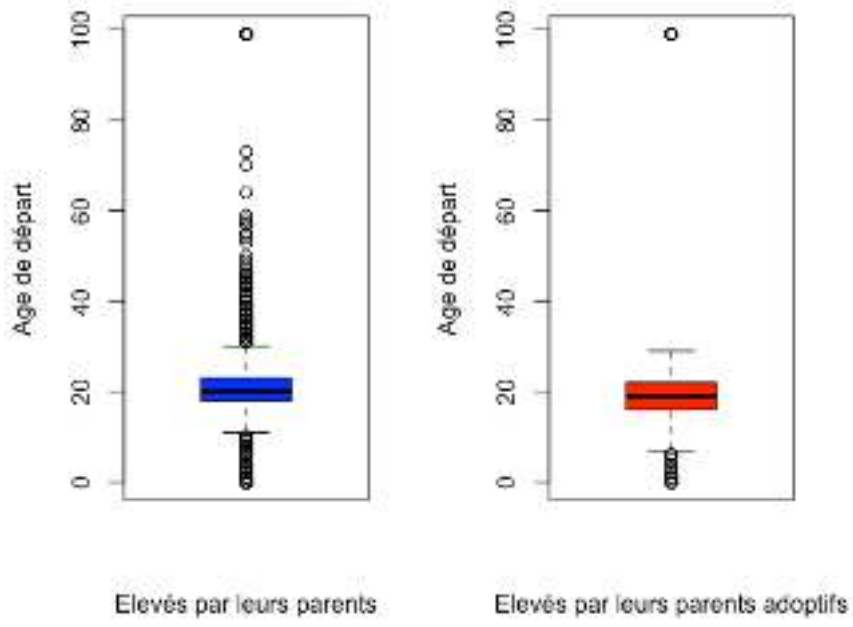
```
> summary(C$BQUITE.N.4.0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   18.00   20.00   19.95   23.00   99.00
```

On peut voir que les résultats sont similaires à ceux de la variable originale. De ce fait, on peut penser que le fait d'avoir été élevé par ses deux parents biologiques n'a pas d'impact sur l'âge de départ du foyer familial, car les enquêtés rentrent dans les valeurs moyennes.

- La dernière variable correspond à l'âge de départ du foyer pour les personnes élevées par un couple ne comportant aucun parent biologique (cf. le script de création de cette variable).

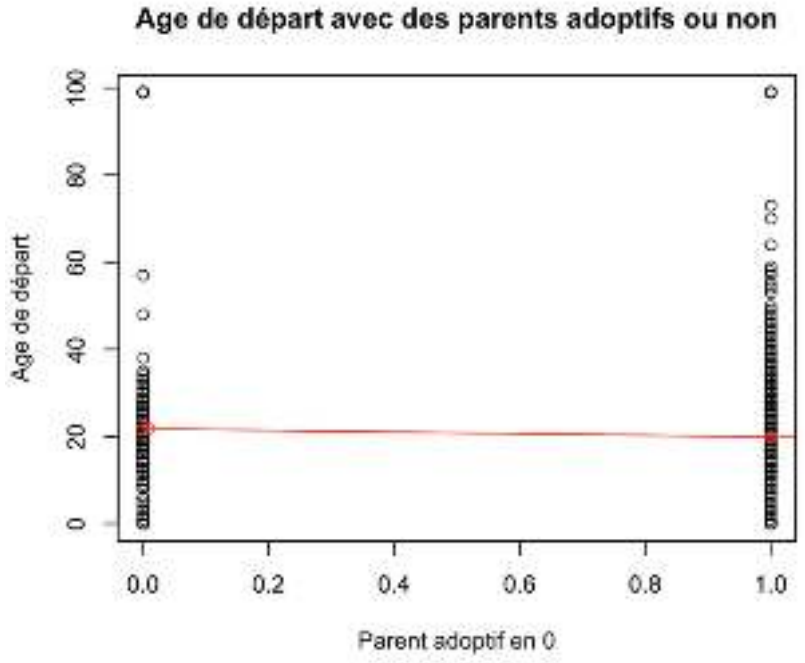
```
> summary(D$BQUITE.N.4.0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   16.00   19.00   20.44   22.00   99.00
```

En comparant aux deux premières variables, on peut remarquer que celle-ci nous donne des résultats bien différents. Ce qui est le plus significatif est la variation des quartiles. En effet, la moitié des enquêtés quittent le domicile familial entre 16 ans et 22 ans. Ainsi, un quart des personnes interrogées quittent le domicile familial avant 16 ans, ce qui est beaucoup plus jeune que pour les deux variables précédentes. Cependant, la moyenne reste dans le même ordre de grandeur. En général, nous pouvons dire que les enquêtés ayant été élevés par un couple ne comportant aucun des deux parents biologiques ont tendance à quitter le foyer familial plus tôt. Les boites à moustache ci-dessous permettent de mieux visualiser les résultats.



Nous pouvons également tracer ce graphe (cf page ci-après), qui est révélateur de la différence entre l'âge de départ des enquêtés élevés par leurs parents biologiques et ceux élevés par un couple ne comportant aucun des deux parents biologiques. Ce graphe nous confirme que lorsque l'enquêté a été élevé par un couple ne comportant aucun des deux parents biologiques, l'âge de départ du domicile familial se fait en moyenne plus tôt avec une

très forte concentration autour des 20ans. Seulement trois personnes ont quitté le domicile familial après 40 ans. Au contraire, les sujets ayant été élevés par leurs deux parents biologiques ont eu une répartition des âges de départ du domicile familiale beaucoup plus étendue. En effet, bon nombre d'entre eux quittent le domicile familial après 40 ans. On peut donc dire qu'au premier regard, le fait d'avoir été élevé par ses parents biologiques entraine un départ plus tardif de l'enfant du domicile familial. Cela peut-il s'expliquer par un cadre familial plus stable ?



La courbe rouge représente la régression linéaire associée. Celle-ci semble presque constante. Cependant, tracer une telle courbe n'a pas de sens car nous avons une variable discrète. Enfin, les points sont trop éloignés de la droite pour donner un sens à cette modélisation.

b) Analyse économétrique

Pour afficher les résultats des régressions linéaires, nous avons utilisé pour chacune d'entre elles le script suivant :

```
#Exploitation des résultats (seulement avec x pour l'exemple)
summary(x) #donne les informations stockées sur x
x$coefficients #donne les coefficients de la droite
summary(x)$r.squared #donne la valeur du R^2
```

Pour l'analyse économétrique, nous prenons $\alpha = 10\%$ afin d'avoir des intervalles de confiance à 90%. Nous allons analyser les résultats obtenus pour les régressions primaires et multiples. Pour chaque étude, il faut regarder la quatrième colonne des coefficients, qui correspond à la p_{value}. Si le coefficient est inférieur à 0,1, cela signifie qu'il y a bien l'existence d'une régression linéaire entre les deux variables. En effet, le test a été fait avec les hypothèses suivantes :

$$\begin{cases} H0 : a_i = 0 \\ H1 : a_i \neq 0 \end{cases}$$

Ainsi, si $p_{va:ue} < 0{,}1$, on rejette H_0 et ainsi $a_i \neq 0$, ce qui montre l'existence d'une régression linéaire et donc d'un lien entre les variables testées.

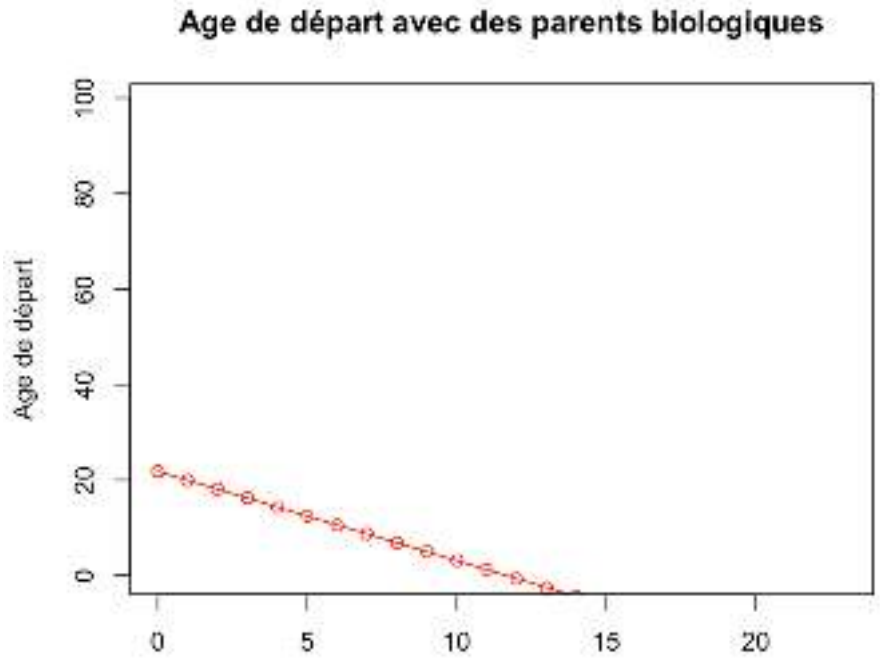
- **Analyse des régressions linéaires primaires :**

- **Par les deux parents biologiques**

Pour celle étude, nous pouvons observer que $p_{value} < 0{,}1$. Ainsi, on rejette l'hypothèse H_0 et nous pouvons supposer que $Y = a_1X_1 + b_1 + \varepsilon$, avec $a_1 = -1{,}8732$ et $b_1 = 21{,}8211$. Cette régression linéaire montre un lien entre l'âge de départ du domicile familial et le fait d'avoir été élevé par ses parents biologiques.

```
Call:
lm(formula = A$BQUITE.N.4.0 ~ A$BELEVQ1.C.1)

Residuals:
    Min      1Q  Median      3Q     Max
-21.821  -1.948   0.052   3.052  79.052

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.8211     0.3607   60.51  < 2e-16 ***
A$BELEVQ1.C.1 -1.8732     0.3800   -4.93 8.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.41 on 8401 degrees of freedom
Multiple R-squared:  0.002884,  Adjusted R-squared:  0.002766
F-statistic:  24.3 on 1 and 8401 DF,  p-value: 8.393e-07
```

Nous pouvons donc tracer la régression linéaire suivante :



Cependant, même si la régression linéaire primaire semble montrer un lien entre l'âge de départ du domicile familial et le fait d'avoir été élevé par ses parent biologiques, comme nous avons une variable discrète qui vaut 0 ou 1, cela n'a pas de sens de tracer une telle droite.

- **Age de départ et parents non biologiques.**

Nous pouvons observer que p_{value}= 0,55 > 0,1 ce qui signifie que l'on accepte l'hypothèse H_0 et on a alors : $Y = 0 * X_2 + b_2 + \varepsilon$, avec b = 20,1185. Ainsi, ce modèle ne permet pas de conclure un lien linéaire entre les deux variables.

Cependant, avec l'analyse descriptive, nous avions plutôt observé un lien entre le fait de ne pas avoir été élevé par ses parents biologiques et l'âge de départ du foyer familial. Ces deux régressions linéaires mettent en avant le contraire, ce qui peut venir de du modèle utilisé.

```
Call:
lm(formula = ASBQUITE.N.4.0 ~ ASBELEVQ6.C.1)

Residuals:
    Min      1Q  Median      3Q     Max
-20.439  -2.119  -0.119   2.881  78.881

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.1185     0.1165 172.704   <2e-16 ***
ASBELEVQ6.C.1  0.3209     0.5366   0.598     0.55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 8401 degrees of freedom
Multiple R-squared:  4.256e-05, Adjusted R-squared:  -7.647e-05
F-statistic: 0.3576 on 1 and 8401 DF,  p-value: 0.5499
```

- **Age de départ et nombre de frères et sœurs**

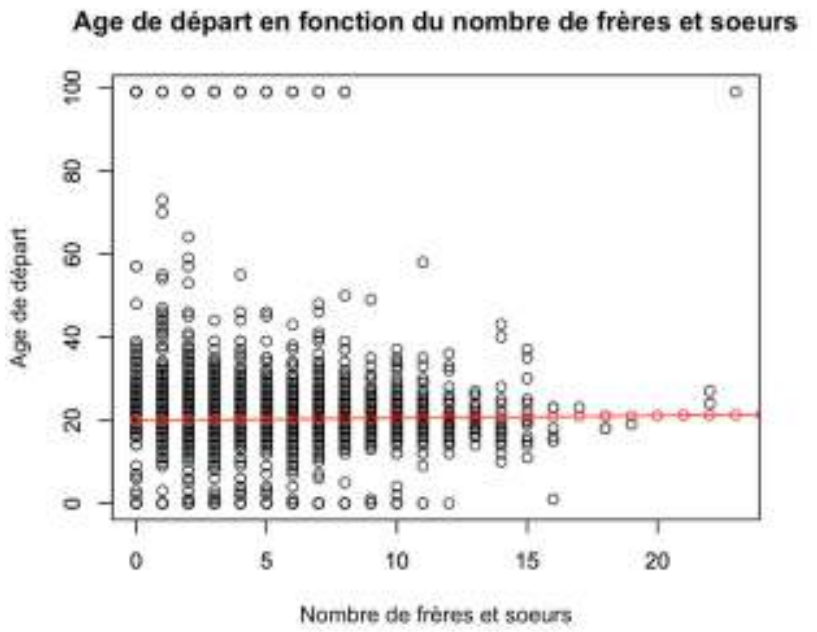
```
Call:
lm(formula = ASBQUITE.N.4.0 ~ ASBNBFS.N.4.0)

Residuals:
    Min      1Q  Median      3Q     Max
-20.632  -2.060  -0.083   2.940  79.055

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.94550    0.17826  111.89   <2e-16 ***
ASBNBFS.N.4.0  0.05718    0.04173    1.37    0.171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 8401 degrees of freedom
Multiple R-squared:  0.0002735, Adjusted R-squared:  0.0001045
F-statistic: 1.878 on 1 and 8401 DF,  p-value: 0.1706
```

Si nous conservons le seuil de 0,1, nous avons donc $p_{value} > 0{,}1$; ce qui rejette l'hypothèse d'une régression linéaire entre les deux variables. Cependant, nous pouvons baisser le seuil de tolérance en prenant par exemple $\alpha = 0{,}2$ et ainsi mettre en avant l'existence d'un lien linéaire entre ces deux variables. Cependant, $R^2 = 1 * 10^{-4}$ ce qui reste très faible et limite la pertinence de la régression linéaire.

Nous pouvons également tracer le graphe suivant afin de se rendre compte de la répartition de l'âge de départ en fonction du nombre de frères et sœurs.

Comme nous pouvons le voir sur le graphe, nous ne pouvons pas établir un lien direct entre le nombre de frères et sœurs et l'âge de départ. En effet, il n'y a pas le même nombre d'enquêté n'ayant aucun, un seul ou 5 frères et sœurs, ce qui limite la comparaison. En revanche, nous pouvons essayer d'observer que plus le nombre de frères et sœurs est élevé, plus le nombre de départs tôt et tard est limité. Les départs semblent plus concentrés autour de 20-22 ans ; et une forme d'entonnoir qui se resserre apparait, lorsque l'on tend vers un nombre de frères et sœurs élevé.

La courbe rouge représente la régression linéaire. Nous pouvons observer que celle-ci est quasiment une droite constante, ce qui explique le fait que l'on accepte H_0 et que l'on ait $a_5 = 0$ et donc : $Y = b_5 + \varepsilon$. Le fait que les points soient aussi éloignés de la courbe confirme également que R^2soit aussi faible. Cela enlève donc de l'intérêt au tracé de la régression linéaire.



- **Age de départ et élevé par ses deux parents biologiques et les considère comme ses parents**

Dans cette analyse, si nous abaissons α à 0,15 nous pouvons observer que les deux variables de la régression linéaire (A$BELEVQ1.C.1 et A$BPAREQ1.C.1) ont $p_{value} < 0{,}15$, ce qui nous permet de dire que la régression linéaire multiple est acceptée et on a ainsi :

- $Y = -1{,}9178 * X_1 - 0{,}9011 * X_3 + 21{,}8958 + \varepsilon$

Si nous conservons $\alpha = 0{,}1$, seule la première variable (X_1) a une p_{value} < 0,1 et ainsi on obtient la même équation que pour la régression linéaire primaire :

- $Y = -1{,}9278 * X_1 + 21{,}8958 + \varepsilon$

Cependant, les coefficients ne sont plus les mêmes, ce qui montre un impact du lien sentimental. En effet, nous avions $Y = -1{,}8732 * X_1 + 21{,}8211 + \varepsilon$.

```
Call:
lm(formula = A$BQUITE.N.4.0 ~ A$BELEVQ1.C.1 + A$BPAREQ1.C.1)

Residuals:
    Min      1Q  Median      3Q     Max
-21.896  -1.978   0.022   3.022  79.022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.8958     0.3640  60.159  < 2e-16 ***
A$BELEVQ1.C.1 -1.9178     0.3811  -5.033 4.94e-07 ***
A$BPAREQ1.C.1 -0.9011     0.5941  -1.517    0.129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.41 on 8400 degrees of freedom
Multiple R-squared:  0.003157,  Adjusted R-squared:  0.00292
F-statistic:  13.3 on 2 and 8400 DF,  p-value: 1.705e-06
```

- **Elevé par un couple que ne sont pas ses parents mais les considère comme ses parents**

```
Call:
lm(formula = A$BQUITE.N.4.0 ~ A$BELEVQ6.C.1 + A$BPAREQ6.C.1)

Residuals:
    Min      1Q  Median      3Q     Max
-21.416  -2.119  -0.119   2.881  80.146

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.1185     0.1165 172.752   <2e-16 ***
A$BELEVQ6.C.1  1.2978     0.6759   1.920   0.0549 .
A$BPAREQ6.C.1 -2.5620     1.0782  -2.376   0.0175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 8400 degrees of freedom
Multiple R-squared:  0.0007143, Adjusted R-squared:  0.0004764
F-statistic: 3.002 on 2 and 8400 DF,  p-value: 0.04973
```

Dans cette étude, nous pouvons observer que les deux P_{value}sont inférieures à 0,1, ce qui valide l'hypothèse d'une régression linéaire multiple. Ainsi on a l'équation suivante :

$$Y = 1{,}2978 * X_2 - 2{,}5620 * X_4 + 20{,}1185 + \varepsilon$$

Nous pouvons observer que, précédemment, la régression linéaire primaire ne permettait pas de mettre en avant un lien linéaire entre l'âge de départ du foyer et le fait d'être élevé par un couple ne comportant aucun des deux parents biologiques. Cependant, l'ajout de la considération de ce couple semble changer l'analyse et rend donc linéaire le lien entre ces variables. Cependant, $R^2 = 4{,}8 * 10^{-4}$reste très faible car les points sont très éloignés de la droite étant donné que nous avons une des variables discrètes.

- **Parents biologiques + considère comme ses parents + nombre de frères et sœurs**

```
Call:
lm(formula = A$BQUITE.N.4.0 ~ A$BELEVQ1.C.1 + A$BPAREQ1.C.1 +
    A$BNBFS.N.4.0)

Residuals:
    Min      1Q  Median      3Q     Max
-22.292  -1.964   0.141   3.088  79.194

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.71177    0.39196  55.393  < 2e-16 ***
A$BELEVQ1.C.1 -1.90564    0.38119  -4.999 5.88e-07 ***
A$BPAREQ1.C.1 -0.91206    0.59417  -1.535    0.125
A$BNBFS.N.4.0  0.05271    0.04168   1.265    0.206
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.41 on 8399 degrees of freedom
Multiple R-squared:  0.003347,  Adjusted R-squared:  0.002991
F-statistic: 9.402 on 3 and 8399 DF,  p-value: 3.359e-06
```

Dans ce cas, l'ajout du nombre de frères et sœurs ne peut être retenu car $p_{value} = 0{,}206 > 0{,}1$. De plus, cela ne permet pas non plus d'accepter la variable du lien sentimental. Comme pour la première régression linéaire multiple précédente (parents biologiques et lien sentimental), seule le fait d'avoir été élevé par ses parents biologiques a un impact sur l'âge de départ du domicile familial. On garde alors quasiment la même régression linéaire que celle de la régression primaire.

Pour conclure sur l'étude de l'âge de départ du domicile familiale lorsque l'enquêté a été élevé par ses deux parents biologiques en couple jusqu'à ses 18 ans, seule la variable X_1(élevé par ses parents biologiques) a un impact sur l'âge de départ. En effet, l'étude de l'ajout du lien sentimental et du nombre de frères et sœurs ne permet pas de mettre en avant un impact significatif de ces deux variables sur l'âge de départ.

- **Parents non biologiques + considère comme ses parents + nombre de frères et sœurs**

Dans cette régression linéaire, nous pouvons observer que l'ajout du nombre de frères et sœurs n'a pas d'impact sur les valeurs de la $\mathrm{p_{value}}$ des deux premières variables déjà étudiées ensemble (parents non biologiques et lien sentimental). Cela a également très peu d'impact sur les coefficients de la régression linéaire qui restent les mêmes jusqu'à la deuxième décimale. Ainsi, nous pouvons dire que l'ajout du nombre de frères et sœurs ne semble pas avoir d'impact supplémentaire sur l'âge de départ de l'enquêté si celui-ci a été élevé par un couple qui ne sont pas ses parents biologiques mais qu'il considère comme ses vrais parents.

Si nous passons à $\alpha = 0{,}2$, alors nous pouvons considérer une régression linéaire entre ces 3 variables, avec $Y = 1{,}29214X_2 - 2{,}56545X_4 + 0{,}05718X_5 + 19{,}93071 + \varepsilon$.

```
Call:
lm(formula = A$BQUITE.N.4.0 ~ A$BELEVQ6.C.1 + A$BPAREQ6.C.1 +
    ASBNBFS.N.4.0)

Residuals:
    Min      1Q  Median      3Q     Max
-21.509  -2.045   0.012   2.955  80.171

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.93071    0.17983 110.832   <2e-16 ***
ASBELEVQ6.C.1  1.29214    0.67586   1.912   0.0559 .
ASBPAREQ6.C.1 -2.56545    1.07811  -2.380   0.0174 *
ASBNBFS.N.4.0  0.05718    0.04172   1.371   0.1705
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 8399 degrees of freedom
Multiple R-squared:  0.0009378, Adjusted R-squared:  0.0005809
F-statistic: 2.628 on 3 and 8399 DF,  p-value: 0.04855
```

En conclusion sur l'étude de l'âge de départ du domicile familial des enquêtés qui n'ont pas été élevé par leurs parents biologiques nous pouvons dire que si la régression linéaire primaire n'a pas été retenue, l'ajout du lien sentimental rend la seconde régression linéaire tout à fait acceptable. En revanche, l'ajout du nombre de frères et sœurs semble moins pertinent.

IV- Conclusion

Tableau récapitulatif de l'existence ou non d'une régression linéaire primaire ou multiple :

	Parents biologiques	Parents non biologiques	Nombre de frères et sœurs
Age de départ	Totale	Aucune	Aucune
	Parents biologiques et considère comme ses parents	Parents non biologiques et considère comme ses parents	
Age de départ	Partielle	Totale	
	Parents biologiques + considère comme ses parents + nb de frères et sœurs	Parents non biologiques + considère comme ses parents + nb de frères et sœurs	
Age de départ	Partielle	Partielle	

Pour conclure notre travail, nous pouvons dire que si l'analyse descriptive mettait en avant un lien entre le fait d'avoir été élevé par un couple ne comportant aucun des deux parents biologiques et un âge de départ plus jeune du foyer, que ceux élevés par leurs deux parents biologiques ; les régressions linéaires primaires montraient le contraire.

Cependant, l'ajout du lien sentimental rend alors pertinent le lien linéaire entre l'âge de départ du foyer et le fait d'avoir été élevé par aucun des deux parents biologiques. Celui-ci n'a en revanche aucun impact sur les enquêté élevé par leurs parents biologiques.

Enfin, l'ajout du nombre de frères et sœurs reste peu pertinent même si la variable seule semble montrer une tendance dite « d'entonnoir ».

Ainsi, nous n'avons pas vraiment réussi à montrer ce que nous voulions, même si nous avons plus ou moins mis en avant certains liens, comme le fait que les enfants adoptés ou confiés ont tendance à partir plus jeune du domicile familial.

Il est nécessaire de prendre du recul sur notre travail mais également sur la pertinence de nos données. En effet, certaines populations sont sur-représentées, ce qui peut influencer les résultats. De plus, l'utilisation de variables qualitatives complique la lecture des régressions linéaires et leur pertinence.

V- Références bibliographiques

- Lien des données

https://www.insee.fr/fr/statistiques/2532244

- Informations sur l'enquête

http://www.progedo-adisp.fr/enquetes/XML/lil.php?lil=lil-0190

https://www.insee.fr/fr/metadonnees/source/serie/s1246/

- Questionnaire relatif à l'étude

http://www.progedo-adisp.fr/documents/lil-0190/lil-0190q.pdf

- Dictionnaire des codes

file:///C:/Users/morga/Downloads/hdv03fd_dictionnaire_des_codes.pdf