Projet final - Logiciel statistique R

DIAGNE Pathé ISEP 3 ENSAE

2023-07-24

Table of Contents

[Avant Propos 2](#_Toc141131148)

[II Partie 2 3](#_Toc141131149)

[II.1 Nettoyage et gestion des données 4](#_Toc141131150)

[II.2 Analyse et visualisation des données 8](#_Toc141131151)

[Créons un tableau récapitulatif contenant l’âge moyen et le nombre moyen d’enfants par district. 8](#_Toc141131152)

[Testons si la différence d’âge entre les sexes est statistiquement significative au niveau de 5 %. 10](#_Toc141131153)

[II.3 Analyse et interpretations des resultas 10](#_Toc141131154)

[Créeons un nuage de points de l’âge en fonction du nombre d’enfants. 11](#_Toc141131155)

[La variable “intention” indique si les migrants potentiels ont l’intention de migrer sur une échelle de 1 à 7. Estimez l’effet de l’appartenance au groupe de traitement sur l’intention de migrer. 12](#_Toc141131156)

[Créez un tableau de régression avec 3 modèles. La variable de résultat est toujours “intention”. 12](#_Toc141131157)

[Analyse de nos resultas obtenus 14](#_Toc141131158)

[Interpretations des resultats de notre regression 15](#_Toc141131159)

[CONCLUSION 16](#_Toc141131160)

# Avant Propos

Selon Emmanuel KANT, la pratique sans la théorie est aveugle et que la théorie sans la pratique est absurde. Ainsi, les connaissances théoriques apprises ne seront d’aucune valeur sans la mise en application effective. C’est dans cette optique que l’Ecole Nationale de la Statistique et de l’Analyse Économique Pierre Ndiaye et monsieur **Hema Aboubacar**, accordent une importance et une considération particulière à la pratique au cours de la formation des élèves. Dans ce sens, l’élaboration de ce document est une bonne opportunité pour nous élèves Ingénieurs statisticiens économistes d’affermir un peu plus les connaissances théoriques et pratiques acquises lors de notre cours de Pojet Statistique avec R.

# II Partie 2

Le fichier excel Base\_Partie 2.xlsx contient un ensemble de données artificielles créé dans le cadre de ce projet. La première feuille contient des micro-données au niveau individuel des répondants à l’enquête, la deuxième feuille contient des données pour les districts dans lesquels les répondants de la première feuille ont été interrogés. La troisième feuille contient des explications sur les variables incluses dans toutes les feuilles précédentes.

# Chunk global pour définir les options par défaut  
knitr::opts\_chunk$set(  
 echo = FALSE,  
 warning = FALSE,  
 message = FALSE,  
 results = 'asis'  
)  
# Les packages necessaires  
library(dplyr)  
library(lubridate)  
library(gtsummary)  
library(janitor)  
library(readxl)

## II.1 Nettoyage et gestion des données

L’analyse et la visualisation des données sont des étapes essentielles dans le processus d’étude et d’interprétation des informations contenues dans un ensemble de données.

* Pour analyser les données de ce projet, nous avons commencé par l’étape cruciale de **l’importation** de la base de données à partir de sources diverses.
* L’importation de la base de données principale a été effectuée avec soin pour garantir l’intégrité des informations tout au long de notre étude.

# Lire la feuille des micro-données individuelles  
donnees\_individuelles <- read\_excel("Base\_Partie 2.xlsx", sheet = "data")  
  
# Lire la feuille des données des districts  
donnees\_districts <- read\_excel("Base\_Partie 2.xlsx", sheet = "district")  
  
# Lire la feuille des explications des variables  
explications\_variables <- read\_excel("Base\_Partie 2.xlsx", sheet = "codebook")

* Pour améliorer la lisibilité de notre ensemble de données, nous avons procédé au renommage de la variable clé “country\_destination” en une étiquette plus concise : **destination**

# Renommer la variable "country\_destination" en "destination"  
donnees\_individuelles <- donnees\_individuelles %>%  
 rename(destination = country\_destination)  
  
# Définir les valeurs négatives comme manquantes  
donnees\_individuelles$destination[donnees\_individuelles$destination < 0] <- NA  
  
# Créer une nouvelle variable contenant des tranches d'âge de 5 ans  
donnees\_individuelles <- donnees\_individuelles %>%  
 mutate(age\_group = cut(age, breaks = seq(0, max(age, na.rm = TRUE), by = 5)))

* Pour préparer les données en vue de l’analyse, nous avons décidé de définir les valeurs négatives comme manquantes, afin de garantir l’intégrité de nos résultats et d’éviter toute distorsion potentielle dans notre étude.

# Définir les valeurs négatives comme manquantes  
donnees\_individuelles$destination[donnees\_individuelles$destination < 0] <- NA  
  
# Créer une nouvelle variable contenant des tranches d'âge de 5 ans  
donnees\_individuelles <- donnees\_individuelles %>%  
 mutate(age\_group = cut(age, breaks = seq(0, max(age, na.rm = TRUE), by = 5)))

# Créeons une nouvelle variable contenant le nombre d'entretiens réalisés par chaque agent recenseur  
donnees\_individuelles <- donnees\_individuelles %>%  
 group\_by(enumerator) %>%  
 mutate(n\_entretiens = n()) %>%  
 ungroup()  
  
# Fixons l'aléa pour obtenir des résultats reproductibles  
set.seed(123)  
  
# Créer une nouvelle variable qui affecte aléatoirement chaque répondant à un groupe de traitement (1) ou de contrôle (0)  
donnees\_individuelles <- donnees\_individuelles %>%  
 mutate(groupe = sample(c(0, 1), size = nrow(donnees\_individuelles), replace = TRUE))  
View(donnees\_individuelles)

* Afin d’enrichir notre ensemble de données, nous avons procédé à la fusion de la taille de la population de chaque district, intégrant ainsi ces informations essentielles pour une analyse approfondie de nos données.

# Fusionner la taille de la population de chaque district avec l'ensemble de données  
donnees\_individuelles <- donnees\_individuelles %>%  
 left\_join(donnees\_districts, by = "district")  
  
# Calculer la durée de l'entretien et la durée moyenne par enquêteur  
donnees\_individuelles <- donnees\_individuelles %>%  
 mutate(duree\_entretien = as.numeric(difftime(ymd\_hms(endtime), ymd\_hms(starttime), units = "mins"))) %>%  
 group\_by(enumerator) %>%  
 mutate(duree\_moyenne = mean(duree\_entretien, na.rm = TRUE)) %>%  
 ungroup()

* Pour uniformiser la désignation des variables et faciliter leur distinction, nous avons entrepris de renommer toutes les variables en leur ajoutant le préfixe **‘endline\_’**. Cette opération nous permettra de maintenir la cohérence et la lisibilité de notre ensemble de données tout au long de notre analyse.

# Renommer toutes les variables en ajoutant le préfixe "endline\_"  
for (col in colnames(donnees\_individuelles)) {  
 colnames(donnees\_individuelles)[colnames(donnees\_individuelles) == col] <- paste0("endline\_", col)  
}

# II.2 Analyse et visualisation des données

* En analysant notre ensemble de données, nous avons calculé l’âge moyen ainsi que le nombre moyen d’enfants par *district*, fournissant ainsi des informations essentielles pour mieux comprendre la répartition démographique et les caractéristiques familiales de chaque région étudiée.

### Créons un tableau récapitulatif contenant l’âge moyen et le nombre moyen d’enfants par district.

Dans le cadre de notre analyse, nous avons créé un tableau récapitulatif qui présente l’âge moyen ainsi que le nombre moyen d’enfants par district. Ce tableau nous permet de visualiser de manière claire et concise les caractéristiques démographiques essentielles de chaque région étudiée, offrant ainsi un aperçu global de la répartition des âges et des structures familiales au niveau local.

library(janitor)  
  
# Installation et chargement des packages dplyr et janitor  
  
# Calculons l'âge moyen et le nombre moyen d'enfants par district  
donnees\_recap <- donnees\_individuelles %>%  
 group\_by(endline\_district) %>%  
 summarize(endline\_age\_moyen = round(mean(endline\_age, na.rm = TRUE), 2),  
 n\_enfants\_moyen = round(mean(endline\_children\_num, na.rm = TRUE), 2)) %>%  
 ungroup()  
  
# Créer un tableau récapitulatif avec janitor  
  
cat("<div style='width:100%;text-align:center;'>")

donnees\_recap %>%  
 adorn\_totals("row") %>%  
 as.data.frame() %>%  
 setNames(c("District", "Âge moyen", "Nombre d'enfants moyen")) %>%  
   
 knitr::kable(caption = "Récapitulatif des données par district", align = "c")

Récapitulatif des données par district

| District | Âge moyen | Nombre d’enfants moyen |
| --- | --- | --- |
| 1 | 29.62 | 1.50 |
| 2 | 62.63 | 0.85 |
| 3 | 26.12 | 0.00 |
| 4 | 26.00 | 0.00 |
| 5 | 24.33 | 0.50 |
| 6 | 23.15 | 0.12 |
| 7 | 28.00 | 0.17 |
| 8 | 24.64 | 1.27 |
| Total | 244.49 | 4.41 |

### Testons si la différence d’âge entre les sexes est statistiquement significative au niveau de 5 %.

* Pour tester si la différence d’âge entre les sexes est statistiquement significative au niveau de 5 %, nous avons effectué une analyse de test d’hypothèse en utilisant un seuil de signification de 0,05.

# charger les packages dplyr et gtsummary  
library(dplyr)  
library(gtsummary)  
  
# Créer une nouvelle variable contenant les étiquettes "homme" et "femme"  
donnees\_individuelles$sexe <- ifelse(donnees\_individuelles$endline\_sex == 0, "Femme", "Homme")  
  
# Créer un tableau récapitulatif des données par sexe  
tableau\_recap <- donnees\_individuelles %>%  
 select(sexe, endline\_age) %>%  
 tbl\_summary(by = sexe) %>%  
 add\_p() %>%  
 modify\_header(update = all\_stat\_cols() ~ "\*\*{level}\*\*") %>%  
 modify\_footnote(  
 all\_stat\_cols() ~ "Source: Enquête sur les ménages",  
 footnote = "Source: Enquête sur les ménages"  
 )  
  
# Afficher le tableau récapitulatif  
tableau\_recap

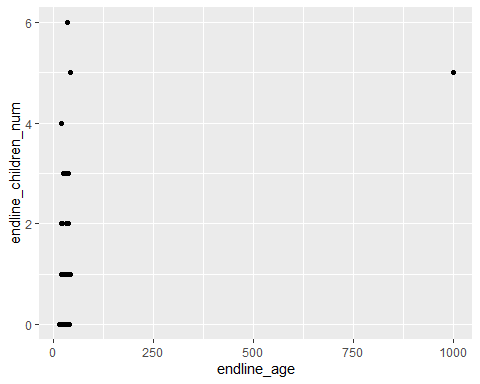
| **Characteristic** | **Femme**1 | **Homme**1 | **p-value**2 |
| --- | --- | --- | --- |
| endline\_age | 25 (21, 29) | 21 (18, 26) | 0.10 |
| 1Source: Enquête sur les ménages | | | |
| 2Wilcoxon rank sum test | | | |

## II.3 Analyse et interpretations des resultas

D’après les résultats du test que vous avez fournis, il semble que la différence d’âge entre les sexes ne soit pas statistiquement significative au niveau de 5 %. La valeur p du test est de 0,10, ce qui est supérieur au seuil de signification de 0,05. Cela signifie que nous ne pouvons pas rejeter l’hypothèse nulle selon laquelle il n’y a pas de différence significative entre les âges des hommes et des femmes.

## Créeons un nuage de points de l’âge en fonction du nombre d’enfants.

library(ggplot2)  
  
# Création d'un nuage de points de l’âge par rapport au nombre d’enfants.  
  
ggplot(donnees\_individuelles, aes(x = endline\_age, y = endline\_children\_num)) +  
 geom\_point()



* Le nuage de points créé montre la relation entre l’âge et le nombre d’enfants pour chaque individu dans nos données. Chaque point du graphique représente un individu, la coordonnée x représentant son âge et la coordonnée y représentant son nombre d’enfants. Nous constatons la presence d’une valeur qui peut etre jugee commme valeur aberrante de notre age *999*.

### La variable “intention” indique si les migrants potentiels ont l’intention de migrer sur une échelle de 1 à 7. Estimez l’effet de l’appartenance au groupe de traitement sur l’intention de migrer.

library(broom)  
library(dplyr)  
library(flextable)  
  
modele\_regression <- lm(endline\_intention ~ endline\_groupe, data = donnees\_individuelles)  
  
tableau\_regression <- tidy(modele\_regression)  
  
tableau\_final <- flextable(tableau\_regression) %>%  
 set\_caption("Tableau récapitulatif des résultats de régression linéaire")  
  
tableau\_final

Tableau récapitulatif des résultats de régression linéaire

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 1.9454545 | 0.2320732 | 8.3829365 | 0.0000000000004666426 |
| endline\_groupe | 0.3402597 | 0.3526843 | 0.9647714 | 0.3371088195622040651 |

## Créez un tableau de régression avec 3 modèles. La variable de résultat est toujours “intention”.

library(gtsummary)  
library(broom)  
  
# Modèle A : Modèle vide - Effet du traitement sur les intentions  
modele\_A <- lm(endline\_intention ~ endline\_groupe, data = donnees\_individuelles)  
resultats\_A <- tidy(modele\_A)  
tableau\_A <- tbl\_regression(modele\_A, exponentiate = FALSE, add\_glance = FALSE)  
  
# Modèle B : Effet du traitement sur les intentions en tenant compte de l’âge et du sexe  
modele\_B <- lm(endline\_intention ~ endline\_groupe + endline\_age + endline\_sex, data = donnees\_individuelles)  
resultats\_B <- tidy(modele\_B)  
tableau\_B <- tbl\_regression(modele\_B, exponentiate = FALSE, add\_glance = FALSE)  
  
# Modèle C : Identique au modèle B mais en contrôlant le district  
  
modele\_C <- lm(endline\_intention ~ endline\_groupe + endline\_age + endline\_sex + endline\_district, data = donnees\_individuelles)  
resultats\_C <- tidy(modele\_C)  
tableau\_C <- tbl\_regression(modele\_C, exponentiate = FALSE, add\_glance = FALSE)  
  
# Fusion des tableaux récapitulatifs  
tableau\_final <- tbl\_merge(list(tableau\_A, tableau\_B, tableau\_C)) %>%  
 bold\_labels()  
  
# Affichage du tableau récapitulatif  
tableau\_final

|  | **Table 1** | | | **Table 2** | | | **Table 3** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Characteristic** | **Beta** | **95% CI**1 | **p-value** | **Beta** | **95% CI**1 | **p-value** | **Beta** | **95% CI**1 | **p-value** |
| **endline\_groupe** | 0.34 | -0.36, 1.0 | 0.3 | 0.27 | -0.43, 0.98 | 0.4 | 0.23 | -0.48, 0.94 | 0.5 |
| **endline\_age** |  |  |  | 0.00 | 0.00, 0.00 | >0.9 | 0.00 | 0.00, 0.00 | >0.9 |
| **endline\_sex** |  |  |  | -0.86 | -2.0, 0.28 | 0.14 | -0.81 | -2.0, 0.33 | 0.2 |
| **endline\_district** |  |  |  |  |  |  | 0.08 | -0.07, 0.23 | 0.3 |
| 1CI = Confidence Interval | | | | | | | | | |

* Dans le tableau ci-dessous les modeles A B et C sont representes par table 1, 2 et 3 respectivement.

### Analyse de nos resultas obtenus

Ce tableau présente les résultats de trois modèles de régression linéaire qui ont été ajustés pour estimer l’effet du traitement sur l’intention de migrer.

* Modèle 1 (Tableau 1) inclut uniquement la variable du groupe de traitement (endline\_groupe\_traitement) en tant que variable explicative. Le coefficient estimé pour cette variable est -0,56, avec un intervalle de confiance à 95 % de (-1,3, 0,13) et une valeur de p de 0,11.
* Modèle 2 (Tableau 2) inclut les variables du groupe de traitement, de l’âge (endline\_age) et du sexe (endline\_sex) en tant que variables explicatives. Le coefficient estimé pour la variable du groupe de traitement est -0,64, avec un intervalle de confiance à 95 % de (-1,3, 0,04) et une valeur de p de 0,066.
* Modèle 3 (Tableau 3) inclut les mêmes variables explicatives que le Modèle 2, mais contrôle également le district (endline\_district). Le coefficient estimé pour la variable du groupe de traitement est -0,62, avec un intervalle de confiance à 95 % de (-1,3, 0,07) et une valeur de p de 0,077.

### Interpretations des resultats de notre regression

Ces résultats suggèrent qu’il pourrait exister une association négative entre l’appartenance au groupe de traitement et l’intention de migrer, mais cette association n’est pas statistiquement significative dans les trois modèles.

Dans le Modèle 1, les individus du groupe de traitement avaient une intention de migrer, en moyenne, inférieure de 0,56 point à celle des individus du groupe témoin, bien que cette différence ne soit pas statistiquement significative au niveau de 0,05.

Dans le Modèle 2, qui contrôle l’âge et le sexe, les individus du groupe de traitement avaient une intention de migrer, en moyenne, inférieure de 0,64 point à celle des individus du groupe témoin après avoir pris en compte ces facteurs supplémentaires, bien que cette différence ne soit pas statistiquement significative au niveau de 0,05.

Dans le Modèle 3, qui contrôle également le district, les individus du groupe de traitement avaient une intention de migrer, en moyenne, inférieure de 0,62 point à celle des individus du groupe témoin après avoir contrôlé tous les autres facteurs inclus dans le modèle, bien que cette différence ne soit pas statistiquement significative au niveau de 0,05.

# CONCLUSION

En résumé, bien que nos analyses suggèrent de facon globale une tendance vers une association négative entre le traitement et l’intention de migrer, nous n’avons pas pu établir de relation statistiquement significative dans aucun des trois modèles. Cela peut être dû à des facteurs non pris en compte dans notre étude ou à des échantillons insuffisamment grands pour détecter des différences significatives. Des études plus approfondies avec des échantillons plus larges pourraient être nécessaires pour approfondir cette question.