**Projet Fouilles de données**

**Description de la base de données**

SwissLabor est une base de données qui comprend 872 observations et 7 variables.

1. participation : c’est une variable qualitative, qui indique si l’individu participe ou non au marché du travail.
2. income : Logarithme du revenu hors travail ou du revenu
3. âge : Âge en décennies (années divisées par 10).
4. education : Années d'éducation
5. youngkids : Nombre de jeunes enfants de moins de 7 ans
6. oldkidsold : Nombre d'enfants plus âgés de plus de 7 ans
7. foreign : variable factorielle (yes, no).

*L'individu est-il un étranger (c'est-à-dire non suisse) ?*

**Source**  
Journal of Applied Econometrics Data Archive.<http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/>

**Références**  
Gerfin, M. (1996). Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation. *Journal of Applied Econometrics*, **11**, 321–339.

## Préparation des données

Après observations de la base nous voyons que celle-ci ne possède pas de valeurs NA. Nous n’avons pas d’identifiants uniques (numéro, nom de famille ...) pour s’assurer qu' il n’y ai pas de doublons. Nous partons donc de l'hypothèse selon laquelle il n’y a pas de doublons dans cette base.

**## ANALYSE VISUELLES**

--

# income

h\_income <-hist(data$income)

h\_income

p\_income <- plot(density(data$income))

p\_income

p1 <- ggplot(data) +

aes(x = "", y = income) +

geom\_boxplot(fill = "#9ECAE1") +

theme\_minimal()

p1

skim(data$income)

# Les revenus sont sous forme de logarithme pour diminuer e'echelle et resserer la distribution

# La moyenne est de 10.6 avec un ecart type de 0.412.

# On observe Un individu qui a un revenu bien inférieur aux autres :

# Il a un revenu de 7.19 alors que la moyenne est autour de 10.6

# On voit nettement qu il est isolé sur l'histogramme et le boxplot

# Hormis cet individu la distribution suit globalement une loi normale

# Il serait intéressant de se renseigner sur cet individu spécifiquement et

# d'essayer de réaliser une modélisation sans celui-ci car il peut avoir une influence notable.

# il faut essayer les deux

--

# "age"

h\_age <-hist(data$age)

h\_age

p\_age <- plot(density(data$age))

# p\_age

p2 <- ggplot(data) +

aes(x = "", y =age

) +

geom\_boxplot(fill = "#9ECAE1") +

theme\_minimal()

p2

skim(data$age)

# La distribution de l'age suit globalement une loi Normale

# Nous n'observons pas d'outlier

# les ages soint dispersé entre 2 decades et 6.2 decades , la moyenne est de 4 decade et l'ecart type de 1.06 decade

--

# "education"

h\_education <-hist(data$education)

h\_education

p\_education <- plot(density(data$education))

# p\_education

p3 <- ggplot(data) +

aes(x = "", y =education

) +

geom\_boxplot(fill = "#9ECAE1") +

theme\_minimal()

p3

skim(data$education)

# la moyenne est de 9.31 années avec un ecart type de 3.04 années

# la repartition est assez large

# on onserve deux individus a 19 et un a 21 années qui sont des outliers hauts concernant l'education

# IL existe egalement un individus avec la valeur a 1 année d'education

# il faudrait realiser un modelisation sans ces individus pour verifier qu'ils n'influencent pas trop le modele.

--

# "youngkids" (Les enfants de moins de 7 ans)

h\_youngkids <-hist(data$youngkids)

# h\_youngkids

p\_youngkids <- plot(density(data$youngkids))

p\_youngkids

p4 <- ggplot(data) +

aes(x = "", y =youngkids

) +

geom\_boxplot(fill = "#9ECAE1") +

theme\_minimal()

# p4

skim(data$youngkids)

# la classe majoritaire est la classe sans enfants de moins de 7 ans.

# C'est pourquoi on a une moyenne de 0.613 et un écart type assez élevé a 0.312

--

"oldkids"

h\_oldkids <-hist(data$oldkids)

# p\_oldkids <- plot(density(data$oldkids))

p5 <- ggplot(data) +

aes(x = "", y =youngkids

) +

geom\_boxplot(fill = "#9ECAE1") +

theme\_minimal()

skim(data$oldkids)

# La repartition des enfants de plus de 7 ans aupres des individus du data frame

# En moyenne les individus ont 0.983 enfants de plus de 7 ans soit un peu moins de 1.

# Cela semble cohérents avec le nombre d'enfants de moins de 7 ans analysé au dessus.

# "foreign"

p6<-ggplot(data, aes(x = foreign, fill = foreign)) + geom\_bar() +

scale\_fill\_brewer(type="qual", palette = 6)

p6

skim(data$foreign)

Dans la base de données, on voit 216 personnes non suisse et 656 personnes suisses soit environ un quart des individus de la base sont étrangers et les ¾ sont sont suisses.

-----

--Variable d'intérêt : participation , variable binaire (yes, no)

p7<-ggplot(data, aes(x = participation, fill = participation)) + geom\_bar() +

scale\_fill\_brewer(type="qual", palette = 6)

p7

skim(data$participation)

Sur les 876 individus on voit que 471 ne sont pas actifs, c’est à dire qu il ne participe pas

au marché du travail.Tandis que 401 personnes participent au marché du travail, soit un peu moins de la moitié.