

RAPPORT : Modèles Linéaires

Master Statistique pour l'Événement et la Prévision

Léa Pimpernelle

Introduction	2
Analyse de la consommation quotidienne de cigarettes dans les 31 pays de l'UE	2
Les revenus du premier quintile des fumeurs quotidiens expliqués par les autres quintiles	3
Conclusion	3
Annexe	4
Carte 1 : prix moyen d'un paquet de cigarette (en €) en Europe	4
Tableau 1 : prix moyen d'un paquet de cigarette (en €) et taxe (en %) dans l'UE	4
Tableau 2 : Pays ayant le plus de fumeurs de cigarettes quotidiens	5
Tableau 3 : Pays ayant le moins de fumeurs de cigarettes quotidiens	5
Histogramme 1 : distribution du premier quintile	5
Histogramme 2 : distribution du second quintile	5
Histogramme 3 : distribution du troisième quintile	6
Histogramme 4 : distribution du quatrième quintile	6
Histogramme 5 : distribution du cinquième quintile	7
Graphique 1 : Régression linéaire multiple	7
Résultat 1 : test de student sur la Régression linéaire multiple	8
Résultat 2 : Corrélation des variables explicatives	9
Résultat 3 : Colinéarité	10
Résultat 4 : homoscedasticité	10
Programmations effectuées sur SAS et R	11
SAS	11
RStudio	13
Sources	17

Introduction

Le tabagisme est un problème de santé publique majeur en Europe. Les taux de tabagisme varient entre les différents pays de l'Union Européenne.

Le prix d'un paquet de cigarettes varie considérablement d'un pays à l'autre en dans l'Union Européenne. En annexe ce trouve une carte ainsi qu'un tableau regroupant le prix moyen d'un paquet de cigarettes et la taxe associée.

Il y a de nombreuses causes qui peuvent contribuer à la consommation de tabac. Les principales causes incluent :

- La disponibilité: Le tabac est largement disponible dans la plupart des pays d'Europe, ce qui en fait facilement accessible pour les consommateurs.
- L'addiction: La nicotine, qui est contenue dans les produits du tabac, est une substance addictive qui peut rendre difficile pour les fumeurs de cesser de fumer.
- Les images de marque: Les campagnes publicitaires et les images de marque peuvent contribuer à la consommation de tabac en créant une image positive du tabagisme.
- Les influences culturelles et sociales: Le tabagisme peut être perçu comme une norme sociale ou culturelle dans certaines régions d'Europe, ce qui peut inciter les gens à fumer.

L'Union européenne a adopté de nombreuses réglementations pour réduire la consommation de tabac et ses effets néfastes sur la santé. Ces réglementations incluent des restrictions sur la publicité et la promotion du tabac, des exigences pour les avertissements sanitaires sur les emballages de cigarettes, et des taxes sur les produits du tabac.

En général, les études montrent qu'il existe un lien entre les revenus plus faibles et la consommation de tabac. En travaillant sur la base de données (fumeurs quotidiens de cigarettes de 31 pays de l'Union Européenne par quintile de revenu) de l'office statistique de l'Union européenne datant de 2019, on cherchera à répondre à la problématique : la consommation de cigarettes par des fumeurs quotidiens est-elle induite par leurs types de revenus classés en cinq groupes ?

Analyse de la consommation quotidienne de cigarettes dans les 31 pays de l'UE

Les pays de l'Union Européenne où le taux de tabagisme est le plus élevé sont généralement ceux d'Europe de l'Est. Selon les données utilisées, en 2019, les taux de tabagisme chez les adultes les plus élevés se trouvaient en Grèce, en Bulgarie, en Roumanie (35%), en

Allemagne, en Serbie, en Turquie et en Hongrie. Ces pays ont plus de 25% de fumeurs quotidiens de cigarettes dans au moins un des cinq quintiles de revenus.

Les pays de l'Union Européenne où le taux de tabagisme est le plus faible sont généralement ceux d'Europe du Nord et de l'Ouest. Selon les données utilisées, en 2019, les taux de tabagisme chez les adultes les plus faibles en Europe se trouvaient en Suède, en Islande, en Norvège, en Finlande et au Portugal. Ces pays n'ont pas plus de 15% de fumeurs quotidiens de cigarettes dans les cinq quintiles de revenus.

La répartition des différents pourcentages des fumeurs quotidiens de cigarettes dans les cinq quintiles est dissemblable. Le taux de tabagisme est plus élevés parmi les fumeurs ayant un faible revenu (premier quintile et second quintile). Pour les trois quintiles suivants, le taux de tabagisme tend à diminuer au fur et à mesure.

Les revenus du premier quintile des fumeurs quotidiens expliqués par les autres quintiles

Une régression multiple avec nos valeurs quantitatives est effectuée afin d'expliquer le taux de tabagisme des premiers quintiles des 31 pays l'UE. Les autres quintiles ne peuvent pas expliquer le taux de tabagisme pour les revenus appartenant au premier quintile hormis le second quintile.

Lorsque l'on regarde la corrélation des quintiles de revenu, le constat est simple, la corrélation entre un quintile est toujours plus forte pour les quintiles proches de celui-ci. Il reste important de ne pas confondre corrélation et causalité, il est possible que l'on observe une corrélation entre deux variables mais que l'une n'est pas la cause de l'autre.

La variance des erreurs d'un modèle de régression n'est pas constante, la dispersion des erreurs varie en fonction des valeurs des variables explicatives. L'hétéroscédasticité cause des problèmes lors de l'interprétation et de la validité des résultats d'un modèle de régression.

Conclusion

Les fumeurs quotidiens aux revenus les plus bas ont tendance à fumer d'avantage que les personnes aux revenus les plus élevés. Il existe des variations entre les pays de l'UE, et le lien entre le revenu et le tabagisme peut varier considérablement d'un pays à l'autre. L'étude statistique montre que la consommation de cigarettes par des fumeurs quotidiens n'est pas forcément induite par leurs types de revenus.

Annexe

CARTE 1 : PRIX MOYEN D'UN PAQUET DE CIGARETTE (EN €) EN EUROPE

Prix moyen d'un paquet de cigarettes en Europe (en euros)

2,77 15,4 €

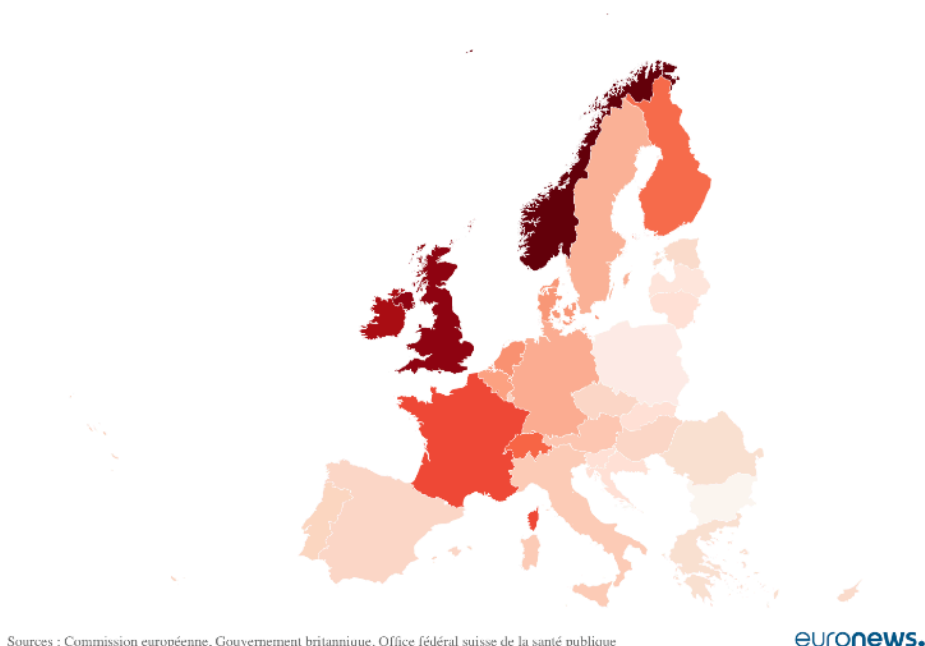


TABLEAU 1 :

PRIX MOYEN D'UN PAQUET DE CIGARETTE (EN €) ET TAXE (EN %) DANS L'UE

Prix moyen d'un paquet de cigarettes en Europe (en euros)								
Pays	Paquet	Taxe	Pays	Paquet	Taxe	Pays	Paquet	Taxe
Belgique	6,93 €	83,38 %	Croatie	4,02 €	80,52 %	Pologne	3,22 €	81,71 %
Bulgarie	2,77 €	81,96 %	Italie	5,20 €	77,83 %	Portugal	4,71 €	76,01 %
Tchéquie	4,54 €	80,06 %	Chypre	4,36 €	75,20 %	Roumanie	4,22 €	73,01 %
Danemark	7,25 €	92,77 %	Lettonie	3,79 €	84,10 %	Slovénie	3,98 €	81,99 %
Allemagne	6,33 €	70,17 %	Lituanie	3,95 €	79,97 %	Slovaquie	3,94 €	80,03 %
Estonie	4,47 €	87,52 %	Luxembourg	4,78 €	69,54 %	Finlande	8,90 €	90,88 %
Irlande	13,43 €	84,62 %	Hongrie	4,65 €	78,15 %	Suède	6,16 €	73,46 %
Grèce	4,18 €	84,79 %	Malte	5,75 €	/	Islande	10,08 €	/
Espagne	4,55 €	70,21 %	Pays-Bas	7,56 €	81,54 %	Norvège	15,40 €	/
France	10,19 €	84,15 %	Autriche	5,35 €	76,98 %	Serbie	3,24 €	/
						Turquie	1,51 €	/

TABLEAU 2 : PAYS AYANT LE PLUS DU FUMEURS DE CIGARETTES QUOTIDIEN

Les pays ayant le plus de fumeurs de cigarettes quotidien						
Obs.	Pays	premier_quintile	second_quintile	troisieme_quintile	quatrieme_quintile	cinquieme_quintile
2	Bulgarie	23	23.4	26.8	31.5	34.6
5	Allemagne	28.6	23	22.6	20.5	15.8
8	Grèce	31.5	25.1	22.4	20.8	21.5
17	Hongrie	29.5	19.7	17	16.5	13.1
30	Serbie	29.3	26.2	23.8	26.9	25.2
31	Turquie	24.7	25.8	26.2	29.1	29.7

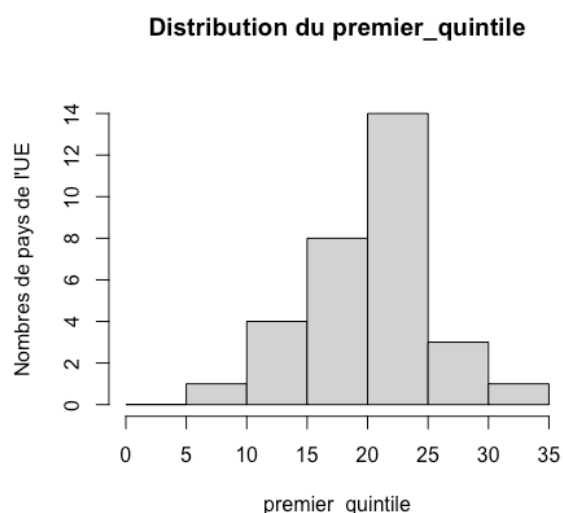
Ces pays n'ont pas plus de 15% de fumeurs quotidiens de cigarettes dans les cinq quintiles de revenus.

TABLEAU 3 : PAYS AYANT LE MOINS DU FUMEURS DE CIGARETTES QUOTIDIEN

Les pays ayant le moins de fumeurs de cigarettes quotidien						
Obs.	Pays	premier_quintile	second_quintile	troisieme_quintile	quatrieme_quintile	cinquieme_quintile
22	Portugal	11.2	10.4	12.5	13	10.4
26	Finlande	11.2	8.9	9.9	10.2	10.1
27	Suède	11.5	7	6.5	5.6	3.4
28	Islande	6.8	7.6	12.4	6.8	4.5
29	Norvège	13.8	12.2	11	8.3	6

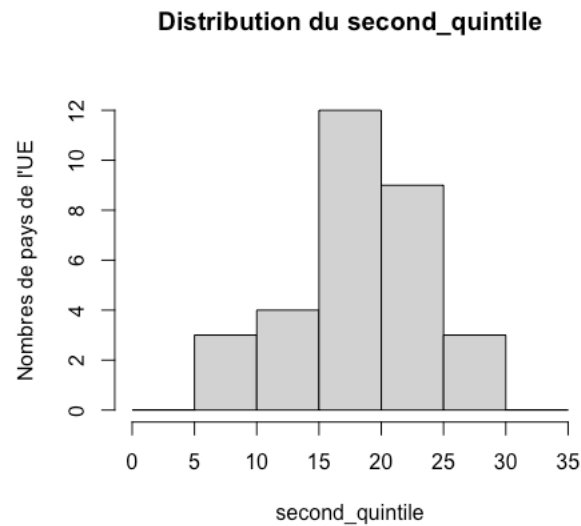
Ces pays ont plus de 25% de fumeurs quotidiens de cigarettes dans au moins une des cinq quintiles de revenus.

HISTOGRAMME 1 : DISTRIBUTION DU PREMIER QUINTILE



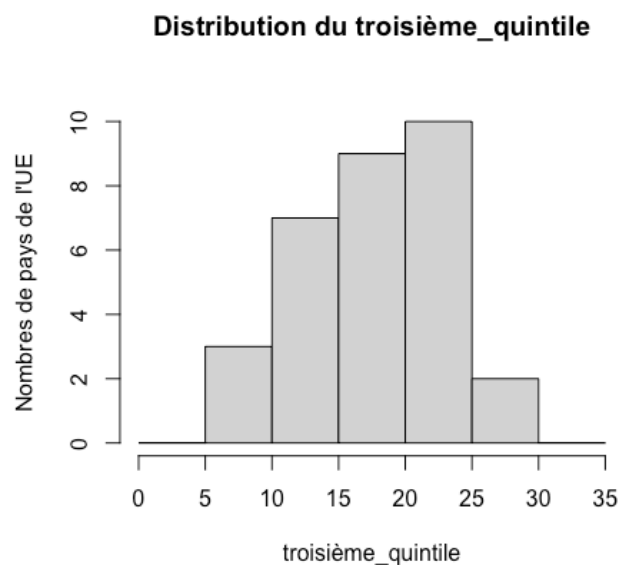
L'histogramme montre que les fumeurs quotidiens de cigarettes appartenant au groupe du premier quintile de revenu se situent entre 20% et 25% pour 14 pays de l'UE.

HISTOGRAMME 2 : DISTRIBUTION DU SECOND QUINTILE



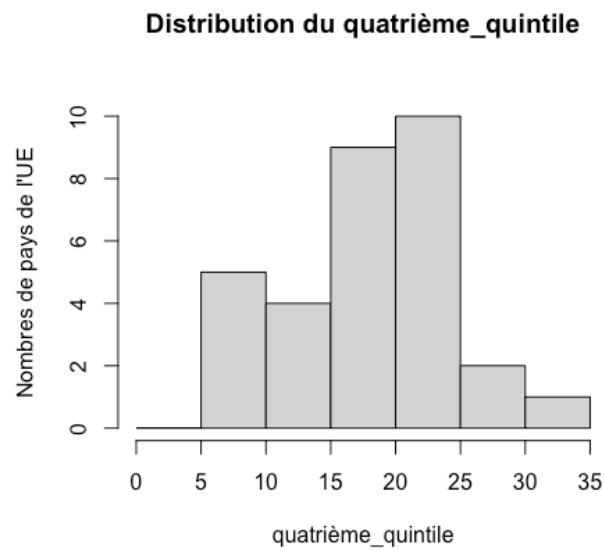
L'histogramme montre que les fumeurs quotidiens de cigarettes appartenant au groupe du second quintile de revenu se situent entre 15% et 20% pour 12 pays de l'UE.

HISTOGRAMME 3 : DISTRIBUTION DU TROISIÈME QUINTILE



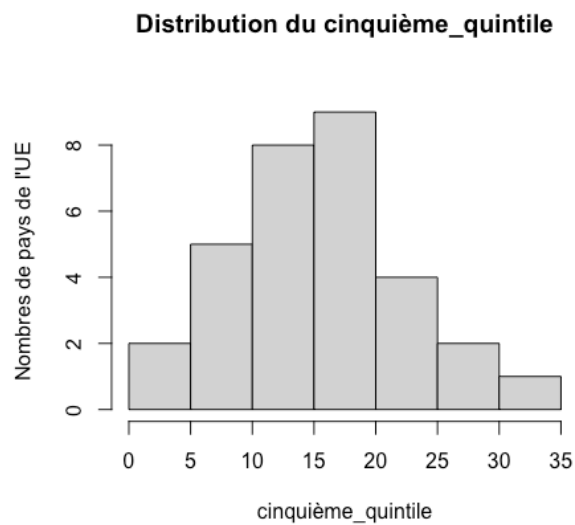
L'histogramme montre que les fumeurs quotidiens de cigarettes appartenant au groupe du troisième quintile de revenu se situent entre 20% et 25% pour 10 pays de l'UE.

HISTOGRAMME 4 : DISTRIBUTION DU QUATRIÈME QUINTILE



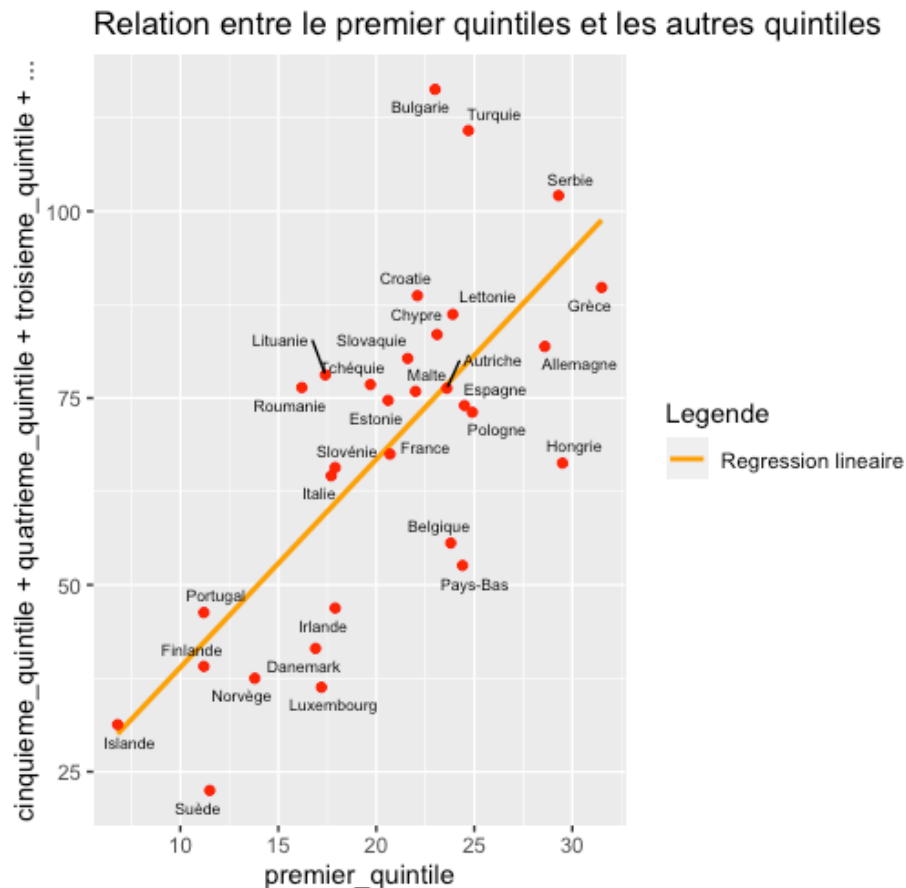
L'histogramme montre que les fumeurs quotidiens de cigarettes appartenant au groupe du second quintile de revenu se situent entre 20% et 25% pour 10 pays de l'UE.

HISTOGRAMME 5 : DISTRIBUTION DU CINQUIÈME QUINTILE



L'histogramme montre que les fumeurs quotidiens de cigarettes appartenant au groupe du second quintile de revenu se situent entre 15% et 20% pour 9 pays de l'UE.

GRAPHIQUE 1 : RÉGRESSION LINÉAIRE MULTIPLE



Sur ce graphique, on peut voir que la Bulgarie a le taux de fumeurs quotidiens de cigarettes le plus élevé pour tous quintiles confondu. La Grèce a le taux le plus élevé de fumeurs quotidiens de cigarettes et pour l'Islande inversement.

Call:

```
lm(formula = don$premier_quintile ~ don$cinquieme_quintile +
  don$quatrieme_quintile + don$troisieme_quintile + don$second_quintile)
```

Coefficients:

(Intercept)	don\$cinquieme_quintile	don\$quatrieme_quintile	don\$troisieme_quintile
2.09708	-0.09120	-0.13656	0.09387
don\$second_quintile			
1.12617			

La formule mathématique de la régression multiple se formule comme suit : valeur de la variable dépendante = intercept + coefficient de régression β_1 x valeur de la variable indépendante1 + coefficient de régression β_2 x valeur de la variable indépendante2 + ... + coefficient de régression β_N x valeur de la variable indépendanteN

```

Call:
lm(formula = don$premier_quintile ~ don$cinquieme_quintile +
    don$quatrieme_quintile + don$troisieme_quintile + don$second_quintile)

Residuals:
    Min       1Q   Median       3Q      Max
-4.5762 -1.8896 -0.3873  1.5983  7.0694

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.09708    2.08966   1.004   0.325
don$cinquieme_quintile -0.09120    0.24558  -0.371   0.713
don$quatrieme_quintile -0.13656    0.38035  -0.359   0.722
don$troisieme_quintile  0.09387    0.28354   0.331   0.743
don$second_quintile    1.12617    0.18718   6.017 2.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.933 on 26 degrees of freedom
Multiple R-squared:  0.7817,    Adjusted R-squared:  0.7481
F-statistic: 23.28 on 4 and 26 DF,  p-value: 2.849e-08

```

La p-value est largement inférieur à 5%. Le second quintile est significatif, l'échantillon provient de population ayant des moyennes similaire. Le r^2 est utilisé pour évaluer la force et la qualité d'une relation de corrélation linéaire entre deux variables, ici il n'est pas suffisamment correcte.

RÉSULTAT 2 : CORRÉLATION DES VARIABLES EXPLICATIVES

Coefficients de corrélation de Pearson, N = 31 Proba > r sous H0: Rho=0					
	premier_quintile	second_quintile	troisieme_quintile	quatrieme_quintile	cinquieme_quintile
premier_quintile premier_quintile	1.00000	0.87458 <.0001	0.68154 <.0001	0.62677 0.0002	0.54452 0.0015
second_quintile second_quintile	0.87458 <.0001	1.00000	0.82690 <.0001	0.79931 <.0001	0.72137 <.0001
troisieme_quintile troisieme_quintile	0.68154 <.0001	0.82690 <.0001	1.00000	0.92860 <.0001	0.86027 <.0001
quatrieme_quintile quatrieme_quintile	0.62677 0.0002	0.79931 <.0001	0.92860 <.0001	1.00000	0.94990 <.0001
cinquieme_quintile cinquieme_quintile	0.54452 0.0015	0.72137 <.0001	0.86027 <.0001	0.94990 <.0001	1.00000

Une corrélation proche de 1 est une corrélation parfaite. La corrélation entre un quintile est toujours plus forte pour les quintiles proche de celui-ci.

RÉSULTAT 3 : COLINÉARITÉ

```

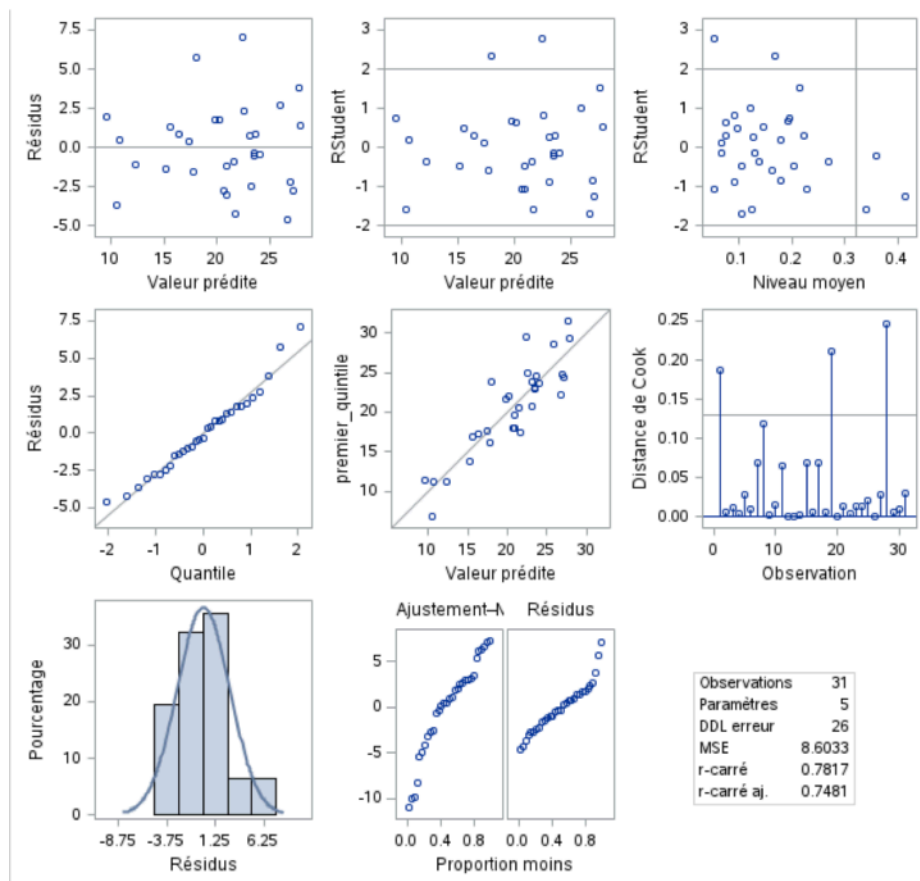
> #coefficients estimés :
> reg$coefficients
      (Intercept)    xsecond_quintile xtroisieme_quintile xquatrieme_quintile xcinquieme_quintile
      2.09708005      1.12617301      0.09387456      -0.1365946      -0.09120077
> # A comparer avec
> cor(Y,x)           # on constate une inversion de signe pour les quintile 4 et 5.
      second_quintile troisieme_quintile quatrieme_quintile cinquieme_quintile
[1,]      0.8745796      0.6815443      0.6267712      0.5445177
> vif(lm(Y~x1+x2+x3+x4)) #on considere que des VIF > 4 ou 5 sont problematiques.
      x1      x2      x3      x4
3.303713 8.614742 20.902920 10.835655

```

On constate une inversion de signe pour les quintile 4 et 5, Il y a un problème de colinéarité entre ces deux variables et le premier quintile.

Les valeurs supérieures à 4 sont problématiques, elles augmentent les erreurs d'estimation des autres variables indépendantes dans le modèle. Seule le second quantile est en dessous de 4.

RÉSULTAT 4 : HOMOSCÉDASTICITÉ



Il y a hétéroscédasticité, la variance des erreurs d'un modèle de régression n'est pas constante, la dispersion des erreurs varie en fonction des valeurs des variables explicatives. Cela cause des problèmes lors de l'interprétation et de la validité des résultats d'un modèle de régression.

Il n'y a pas de valeurs aberrantes par la distance de Cook. (graphique au milieu à droite)

Les valeurs se trouvent entre -2 et 2, seulement 2 au dessus, il y a normalité. (graphique en haut au milieu)

Programmations effectués sur SAS et R

Code	SAS	R
Structure de la Base de Donnée	X	X
Histogrammes	X	X
Conditions	X (where)	X (subset)
Régression Linéaire Multiple	X	X
Student	X	X
Fisher	X	X
Colinéarité	X (VIF)	X
Multicolinéarité		X
Homoscédasticité	X	X
Distance de Cook	X	
AIC/BIC	X	

SAS

```
libname projet '/home/u62350341' ; /*bibliothèque*/
```

```
FILENAME tabac '/home/u62350341/projet/bdd.xlsx'; /*modification du nom du chemin*/
```

```
/*Importation de la base de données*/
```

```
PROC IMPORT DATAFILE= tabac
```

```
    DBMS=XLSX
```

```
    OUT=projet.tabac;
```

```
    GETNAMES=YES;
```

```
RUN;
```

```
/*Affichage de la base de données*/
```

```
title "Fumeur quotidien de cigarettes par quantile de revenu dans les pays d'Europe";
```

```
PROC print DATA=projet.tabac;
```

```
RUN;
```

```
/*Structure de la base de données*/
```

```
title "Structure de la base de données";
```

```
PROC contents DATA=projet.tabac;
```

```
RUN;
```

```
/*Graphique : histogrammes des 5 quintiles*/
```

```
TITLE "Histogramme du premier quintile de revenu dans 31 pays de l'UE";
```

```
PROC UNIVARIATE DATA = projet.tabac NOPRINT;  
HISTOGRAM premier_quintile ; RUN;
```

```
TITLE "Histogramme du second quintile de revenu dans 31 pays de l'UE";  
PROC UNIVARIATE DATA = projet.tabac NOPRINT;  
HISTOGRAM second_quintile ; RUN; TITLE 'Summary of Weight Variable (in pounds)';
```

```
TITLE "Histogramme du troisième quintile de revenu dans 31 pays de l'UE";  
PROC UNIVARIATE DATA = projet.tabac NOPRINT;  
HISTOGRAM troisieme_quintile ; RUN;
```

```
TITLE "Histogramme du quatrième quintile de revenu dans 31 pays de l'UE";  
PROC UNIVARIATE DATA = projet.tabac NOPRINT;  
HISTOGRAM quatrieme_quintile ; RUN;
```

```
TITLE "Histogramme du cinquième quintile de revenu dans 31 pays de l'UE";  
PROC UNIVARIATE DATA = projet.tabac NOPRINT;  
HISTOGRAM cinquieme_quintile ; RUN;
```

```
/*Les pays qui fument le plus et le moins*/
```

```
proc print DATA= projet.tabac;  
    var pays premier_quintile second_quintile troisieme_quintile quatrieme_quintile  
    cinquieme_quintile;  
    where premier_quintile>25 or second_quintile>25 or troisieme_quintile>25 or  
    quatrieme_quintile>25 or cinquieme_quintile>25;  
run;
```

```
proc print DATA= projet.tabac;  
    var pays premier_quintile second_quintile troisieme_quintile quatrieme_quintile  
    cinquieme_quintile;  
    where premier_quintile<15 and second_quintile<15 and troisieme_quintile<15 and  
    quatrieme_quintile<15 and cinquieme_quintile<15;  
run;
```

```
/*Régression Linéaire multiple*/
```

```
proc REG data=projet.tabac;  
    model premier_quintile=second_quintile troisieme_quintile quatrieme_quintile  
    cinquieme_quintile / vif;  
run;
```

```
proc glm data=projet.tabac;
model premier_quintile=second_quintile troisieme_quintile quatrieme_quintile
cinquieme_quintile / ss3 solution tolerance;
run;

/*Corrélation de Pearson*/
proc CORR data=projet.tabac;
    title 'Corrélation de Pearson';
    var premier_quintile second_quintile troisieme_quintile quatrieme_quintile
cinquieme_quintile;
run;

/* Homoscédasticité */
proc reg data=projet.tabac ;
    title 'Homoscédasticité';
    premier_quintile : model premier_quintile=second_quintile troisieme_quintile
quatrieme_quintile cinquieme_quintile / SPEC;
    run;

    /* QQ Plot */
    plot R.* NQQ.;
run;
```

RSTUDIO

```
#-----#
# Projet : Consommation de cigarette quotidien des fumeurs de 31 pays de l'UE
# Date : 12/01/2023
# Auteur : Léa PIMPERNELLE
#-----#

require(openxlsx)
require("ggrepel")
require(scales)
require(leaps) # pour le calcul de la recherche exhaustive de choix de mod?les
require(readr)
require(car) #pour le calcul des VIF

library(openxlsx)
library("ggrepel")
library(scales)
library(leaps)
library(readr)
```

```
library(car)
```

```
# chemin vers le répertoire courant où est placé le fichier de données (à adapter) :  
setwd("/Users/lea/Desktop/M1/s1/SEP731-ModLin/projet")
```

```
# Importation de la base de données  
don <- openxlsx::read.xlsx(xlsxFile = "bdd.xlsx")
```

```
# Affichage de la base de données ----  
don # Affichage
```

```
# Structure de la base de données -----
```

```
str(don)      # pour vérifier la nature des variables importées  
names(don)    # liste des noms de variables (ou colnames(don))
```

```
## Graphique : histogrammes des 5 quintiles -----
```

```
h1<-hist(don$premier_quintile,  
        breaks = seq(0,35,5),  
        main = 'Distribution du premier_quintile',  
        xlab='premier_quintile',  
        ylab="Nombres de pays de l'UE")
```

```
h2<-hist(don$second_quintile,  
        breaks = seq(0,35,5),  
        main = 'Distribution du second_quintile',  
        xlab='second_quintile',  
        ylab="Nombres de pays de l'UE")
```

```
h3<-hist(don$troisieme_quintile,  
        breaks = seq(0,35,5),  
        main = 'Distribution du troisième_quintile',  
        xlab='troisième_quintile',  
        ylab="Nombres de pays de l'UE")
```

```
h4<-hist(don$quatrieme_quintile,  
        breaks = seq(0,35,5),  
        main = 'Distribution du quatrième_quintile',  
        xlab='quatrième_quintile',  
        ylab="Nombres de pays de l'UE")
```

```

h5<-hist(don$cinquieme_quintile,
        breaks = seq(0,35,5),
        main = 'Distribution du cinquième_quintile',
        xlab='cinquième_quintile',
        ylab="Nombres de pays de l'UE")

## Les pays qui fument le plus et le moins ----
plus<-subset(don,don$premier_quintile>25 | don$second_quintile>25 |
don$troisieme_quintile>25 | don$quatrieme_quintile>25 | don$cinquieme_quintile>25 )
plus

moins<-subset(don,don$premier_quintile<15 & don$second_quintile<15 &
don$troisieme_quintile<15 & don$quatrieme_quintile<15 & don$cinquieme_quintile<15 )
moins

### Régression linéaire multiple ----
ggplot(don,aes(premier_quintile,cinquieme_quintile+quatrieme_quintile+troisieme_quintile
+second_quintile,label=Pays))+
  ggtitle("Relation entre le premier quintiles et les autres quintiles")+
  geom_smooth(method = 'lm',se=FALSE, aes(color='orange'))+
  scale_color_identity(name = "Legende",
                      breaks = c("orange"),
                      labels = c("Regression lineaire"),
                      guide = "legend")+
  scale_y_continuous(labels = comma)+
  scale_x_continuous(labels = comma)+
  geom_point(col='red')+
  geom_text_repel(max.overlaps = 30,size=2.3)

# Ajustement de la base de données
rownames(don)<-don[,1] # La premiere colonne contient les noms des pays.
don<-don[,c(-1)]      # on retire la colonne 1 du contenu des données

# Régression
reg <-
lm(don$premier_quintile~don$cinquieme_quintile+don$quatrieme_quintile+don$troisieme
_quintile+don$second_quintile)
confint(reg)
s_reg <- summary(reg)
s_reg

```



```
# test anova
a_reg <- anova(reg)
a_reg

### Colinéarité ----
x <- as.matrix(don[,c(-1)]) # on enlève la variable à expliquer (= on enlève le premier
quintile)
Y <- as.matrix(don[,1])      # variable endogène (le premier quintile)

# Correlation des variables explicatives autre que la constante
cor(x)      # beaucoup de corrélation (>0.8)

# déterminant petit ?
det(cor(x))  # confirme la multicolinéarité (=0.003927748)

# comparaison des estimateurs des beta avec les corrélations des régresseurs avec le premier
quintile
reg <- lm(Y ~ x)
# Résumé de la régression :
s_reg <- summary(reg)
# coefficients estimés :
reg$coefficients

# À comparer avec
cor(Y,x)      # on constate une inversion de signe pour les quintiles 4 et 5.
              # Il y a un problème de colinéarité entre ces deux variables et le reste des variables.

# Significativité globale et pour individuel
anova(reg) # F de Fisher est bien significatif.
s_reg[["r.squared"]] # R2

# Quelles valeurs de la variance des coef ? 2nde colonne de std.error dans summary
etbeta_sig <- summary(reg)$coef[,2]
etbeta <- summary(reg)$coef[,1]
etbeta/etbeta_sig # variance trop grande par rapport aux valeurs estimées : multicolinéarité ?
# pourrait expliquer le pb de significativité

# On stocke le R2
R2 <- summary(reg)$r.squared # ou $adj.r.squared ou [[9]] si on veut l'ajuster
R2
```

#Regle de Klein

C<- cor(Y,x)

C^2 # proches de R2 pour le second quintile

C^2 > R2 #coef plus gd que R2 ?

#VIF

x1<-x[,1] # second quintile

x2<-x[,2] # troisième quintile

x3<-x[,3] # quatrième quintile

x4<-x[,4] # cinquième quintile

vif(lm(Y~x1+x2+x3+x4)) #on considere que des VIF > 4 ou 5 sont problematiques.

seulement le second quintile n'est pas problématique

#####@

#fournir un AIC : cadre de prévision

AIC(reg) #

#fourbir BIC :d'un modèle explicative

BIC(reg) #

#recherche exhaustive : importer la library leaps pour ce qui suit

reg.exhaustive<-regsubsets(Y~x,data=don, method="exhaustive")

#avec differents criteres

plot(reg.exhaustive, scale="adjr2") # le plus grand

plot(reg.exhaustive, scale="bic") # expliqué

plot(reg.exhaustive, scale="Cp") # le plus petit possible

Sources

Base de données :

https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_SK3I_custom_4077910/default/table?lang=fr

Prix des paquets de cigarettes :

<https://fr.euronews.com/2022/09/26/quel-pays-a-les-cigarettes-les-plus-cheres-deurope-tour-dhorizon-de-la-taxation-sur-le-tab>

<https://www.combien-coute.net/cigarette/>