Faculté de Sciences Économiques de Rennes 1
Master mention Mathématiques appliquées, statistique

Rapport du projet d'économétrie

année universitaire 2021-2022

ANALYSE DES CAMBRIOLAGES EN FRANCE

Par
NGUYEN Manh Hung
DHENNIN Nolwenn
PHAM Thi Minh Ngoc

Enseignant:
Véronique Thelen

Table des matières

Introduction

Le cadre de notre étude porte sur le taux de criminalité par département en France en 2016, et tout particulièrement sur les cambriolages. Le taux de criminalité en termes de cambriolages par département en France est la proportion de cambriolages constatés par les services de police et de gendarmerie par rapport à la population totale du département français concerné, au cours d'une année civile.
Les cambriolages recensées pour notre analyse sont les cambriolages de locaux d'habitations principales, de résidences secondaires, de locaux industriels ainsi que les cambriolages d'autres lieux non spécifiés. Nous avons décidé de les regrouper afin d'avoir une analyse globale des cambriolages par département en France.
Afin de pouvoir expliquer le taux de cambriolages en France, nous voulons porter notre attention sur l'existence de facteurs qui contribuent à l'évolution de ce phénomène tel que le poids du taux de chômage sur le taux de cambriolages par exemple. C'est pourquoi, il est intéressant d'analyser le comportement des cambriolages de chaque département de la France.
La France, malgré sa puissance et son niveau de vie moyen élevé, reste un pays où nous pouvons observer de nombreux crimes. Ainsi, ce pays va être en outre très intéressant à étudier pour relever la relation entre les inégalités et les cambriolages.
Dans notre étude, nous utilisons un jeu de données provenant de l'INSEE, l'Institut National de la Statistique et des Études Économiques. C'est pourquoi, nous pouvons assurer une certaine réalité des données analysées et de la quantité de personnes interrogées pour refléter la réalité des faits.
L'objectif de ce projet est d'estimer un modèle économétrique sur les cambriolages en France métropolitaine basé sur des données de 2016.
Concernant les départements français étudiés, nous avons pris la décision de ne garder que les départements de France métropolitaine et donc d'exclure les départements de France d'outre-mer.
Nous pouvons alors nous questionner ainsi : **Quels sont les facteurs qui déterminent la criminalité en France ?**
Nous allons tout d'abord commencer par présenter les données collectées, ensuite nous décrirons les statistiques descriptives de celles-ci et enfin, nous modéliserons les données de manière à observer les différentes associations, relations et contraintes relatives aux données. Pour terminer, nous conluerons afin de résumer les résultats obtenus de notre étude.

I. Présentation des données

1) Choix des variables

Pour effectuer notre étude sur les cambriolages et les différents facteurs pouvant être déterminants, nous avons choisi d'utiliser le logiciel RStudio. À partir d'un jeu de données de l'INSEE, nous avons choisi d'étudier 6 documents (transformés en table via RStudio) recensant les variables :

- *diplome* : les niveaux de diplômes par département en 2016 en France

  Le niveau d'éducation est souvent mentionné comme ayant une importance forte sur la criminalité. Cependant, il n'est pas si facile de déterminer leur étroit lien. Cela est peut-être dû au fait que nous imaginons que plus le niveau d'éducation d'une personne est élevé, moins celle-ci aura l'occasion de commettre des crimes tels que les cambriolages. Ainsi, une corrélation négative pourrait être attendue entre les deux variables. Cependant, on peut aussi observer que plus une personne est scolarisée, plus elle peut être influencée à commettre des crimes via la rencontre de nouvelles personnes ou encore l'effet de groupe. C'est pourquoi, la corrélation entre le niveau d'éducation sur le taux de cambriolages peut aussi être positive.
- *pauvrete* : les niveaux de vie des personnes pauvres par département en 2016 en France

  Le niveau de pauvreté apparaît comme une variable permettant d'expliquer le taux de cambriolages en France. En effet, il se peut que plus on est pauvre, plus nous avons besoin de ressources et donc plus nous sommes prêts à cambrioler potentiellement.
- *revenu* : la distribution des revenus disponibles par unité de consommation et composition du revenu disponible par départements en France en 2016.

  Le revenu d'une personne détermine son niveau et sa qualité de vie. C'est pourquoi, nous pouvons estimer que plus nous avons un revenu élevé, moins nous sommes portés à cambrioler.
- *tauxChomage* : le taux de chômage (en %) par département depuis 1982 jusqu'à 2019 en France

  Le taux de chômage peut avoir un effet participant à l'augmentation du nombre de cambriolages. En effet, on peut observer que plus une personne n'est pas sur le marché du travail, plus la probabilité de prendre part à un crime tel qu' un cambriolage est grande. Nous pouvons alors estimer qu'il y a une relation positive entre le taux de chômage et le taux de cambriolages.
- *delits_fr_2016* : le nombre des différents délits (cambriolages, vols, homicides, prises d'otages, etc) commis par département en France en 2016.
- *pop* : recensant les populations légales des départements, régions, etc en vigueur au 1er janvier 2016 en France.

  Pour pouvoir comparer chaque département, il est obligatoire de les mettre sous la même échelle. C'est pourquoi nous devons mettre en rapport les données obtenues du nombre de cambriolages par département suivant la population totale de celui-ci.

2) Regroupement des données intéressantes à l'étude

a) Choix des individus

Nos individus étudiés pour ce projet sont les départements de la France métropolitaine c'est-à-dire les départements de 1 à 95 en gardant les départements de la Corse : 2A et 2B. Ainsi, nous avons 96 individus au total (car le numéro de département 20 n'existe pas).

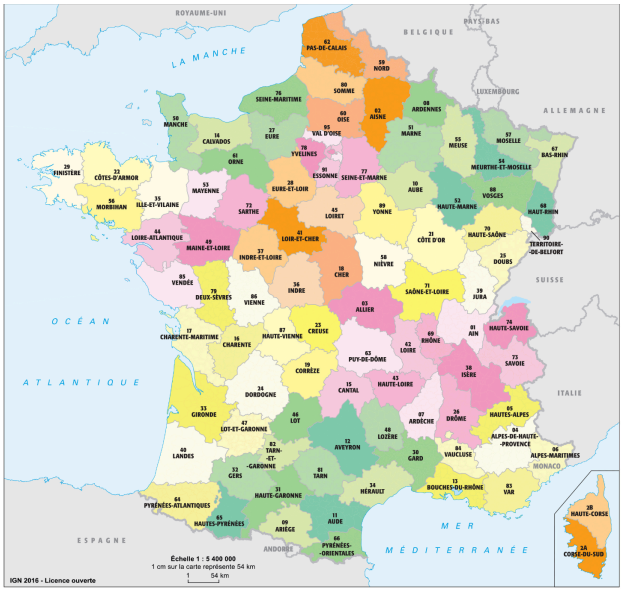

Figure I.1 - Carte des départements en France

b) Filtrage des variables

Pour pouvoir faire une étude intéressante de notre modèle, il est nécessaire de filtrer et résumer les variables préalablement décrites afin de ne garder que les informations importantes regroupant toute la population des départements. Nous allons donc récupérer les données nous intéressant pour notre étude en modifiant les variables précédemment importées ou en en créant d'autres.

Pour la variable diplôme, nous ne gardons que la part des personnes dont le diplôme le plus élevé est le bepc ou le brevet, dans la population non scolarisée de 15 ans ou plus en 2016, pour les départements de France métropolitaine. Il est intéressant de n'étudier que cette catégorie afin de voir si par exemple, dans les départements où le niveau d'étude maximal est le brevet ou le bepc, il y a plus de cambriolages ou non en comparant avec les autres départements. Dans notre modèle, cette variable se nommera **'NbBrevetBepc'**.

Depuis la variable pauvreté, on ne garde que le taux de pauvreté au seuil de 60% (%) tout simplement pour comparer un seul seuil de pauvreté pour tous les départements, cela simplifie l'analyse de ces données. Dans notre modèle, cette variable se nommera '**pauvrete**'.

Depuis la variable pop, on ne récupère que la population totale de chaque département afin de pouvoir affecter un effet taille à notre variable endogène ultérieurement.

Depuis la variable revenu, on ne garde que la médiane du revenu afin d'étudier la distribution moyenne des revenus disponibles par unité de consommation et composition du revenu disponible. Dans notre modèle, cette variable se nommera '**revenu_median**'.

Depuis la variable tauxChomage, on garde le taux de chômage moyen pour l'année 2016. Le taux de chômage est normalement étudié en semestre, mais ici, nous avons choisi d'étudier la moyenne de celui-ci sur l'année de 2016 complète. Dans notre modèle, cette variable se nommera '**tauxChomagemoyen**'.

À partir de la variable delits_fr_2016, on récupère la somme totale des différents types de cambriolages (cambriolages de locaux d'habitations principales + de résidences secondaires + de locaux industriels + d'autres lieux non spécifiés). Cette variable s'appellera nbr_cambriolage mais n'aura pas d'impact dans notre modèle.

c) Création de nouvelles variables

- **Y_cambriolagesur10000** : Cette variable va représenter notre variable endogène, le Y de notre équation, c'est-à-dire la variable réponse. Elle représente le nombre de cambriolages par département pour 10 000 habitants.

- **Grandmetropole** :

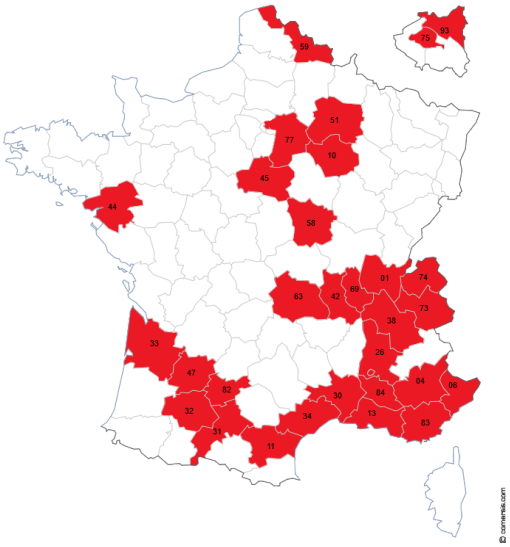

Figure I.2 - Les 30 départements les plus cambriolés en 2016

Cette nouvelle variable représente la variable indicatrice du modèle. Elle est égale à 1 si le département de l'individu appartient aux 30 départements les plus cambriolés pour 10 000 habitants en 2016 et 0 sinon.
Comme ces départements représentent là où de nombreuses métropoles se trouvent, nous avons décidé d'appeler notre variable indicatrice "Grandmetropole".
Sur cette carte, représentant les différents départements de la variable indicatrice, nous pouvons observer que les départements forment comme des groupes pour la majorité. En effet, on remarque qu'il y a beaucoup de cambriolages pour 10 000 habitants soit au Sud, Sud-Est, Sud-Ouest de la France ou encore sur la région autour de Paris, mais, il y a deux départements qui restent seuls : les départements du 44 (Loire-Atlantique) et du 59 (Nord).

- **Rev2** : Cette variable représente le revenu au carré, qui va nous aider à améliorer notre modèle, c'est pourquoi, nous créons une variable le représentant.

3) Équation du modèle

Notre modèle final que nous allons étudier est :
Y_cambriolagesur10000 = B0 + B1pauvrete + B2NbBrevetBepc + B3revenu_median + B4tauxChomagemoyen + B5Grandmetropole + Beta6Rev2
où les variables pauvrete, NbBrevetBepc, revenu_median, tauxChomagemoyen, Rev2 sont les variables exogènes du modèle, c'est-à-dire les variables explicatives.

II. Statistiques descriptives

1) Statistiques univariées

```
-- Variable type: numeric ------------------------------------------------------------------------------------------------------------------------------
# A tibble: 9 x 11
  skim_variable        n_missing complete_rate          mean           sd          p0         p25         p50         p75        p100 hist
* <chr>                    <int>         <dbl>         <dbl>        <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl> <chr>
1 pauvrete                     0             1       14.6          3.11         9.2        12.6        14.4        15.8        28.6  
2 NbBrevetBepc                 0             1        5.83         0.885        3.6         5.3         5.7         6.5         8.4  
3 revenu_median                0             1    20485.       1590.       16996.      19575.      20113       21024.      26808  
4 tauxchomagemoyen             0             1        9.75         1.78         6.25        8.39        9.59       10.6        15.4  
5 Population totale            0             1   685778.      519047.       80141      309930.     551103      866211     2639070    
6 Nbrcambriolage               0             1     3939.        3897.         207        1343.       2652.       4524       18795    
7 Y_cambriolagesur10000        0             1       52.3         15.2         24.9        42.3        48.7        61.2       101.  
8 Grandmetropole               0             1        0.312        0.466        0           0           0           1           1    
9 Rev2                         0             1 422151412.    69890150.    288872514.  383168974.  404532862.  442025413.  718668864  
```

- Commençons par observer la part des personnes dont le diplôme le plus élevé est le bepc ou le brevet, dans la population non scolarisée de 15 ans ou plus en 2016, en fonction de si le département appartient aux départements les plus cambriolés (variable indicatrice) ou non.

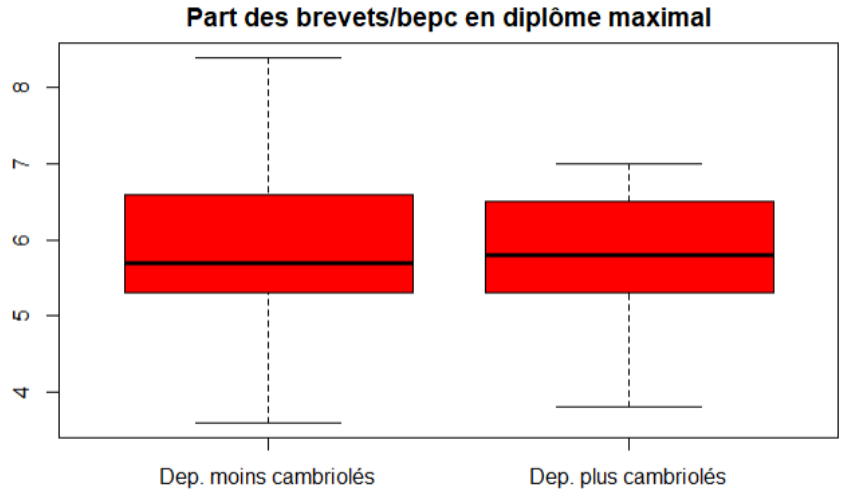

Figure II.1 - Nombre de brevets/bepc en 2016 selon 2 catégories de départements

On remarque que les départements les 'moins' cambriolés ont une moyenne de 5.7 en ce qui concerne la part des personnes dont le diplôme le plus élevé est le bepc ou le brevet, dans la population non scolarisée de 15 ans ou plus, ce qui est légèrement inférieur à la moyenne des départements les plus cambriolés qui elle est à 5.8. Il n'y a pas de grande différence entre les moyennes, et non plus en termes de variance des deux groupes de données.
Nous pouvons interpréter cela comme ceci : il n'y a pas plus de cambriolages dans les départements où la part des personnes dont le diplôme le plus élevé est le bepc ou le brevet est moins grande. C'est-à-dire que là où il y a le plus de cambriolages, ne se trouve pas forcément des personnes dont le niveau d'études est grand, et donc ont plus de richesse. Ainsi, ici, on ne peut pas conclure avec un simple graphique que le niveau d'étude a un réel impact sur le fait que le département soit cambriolé plus ou non.

- Maintenant, étudions le revenu médian des départements les ‘moins’ cambriolés en le comparant à ceux les ‘plus’ cambriolés. Le revenu médian est une variable quantitative.

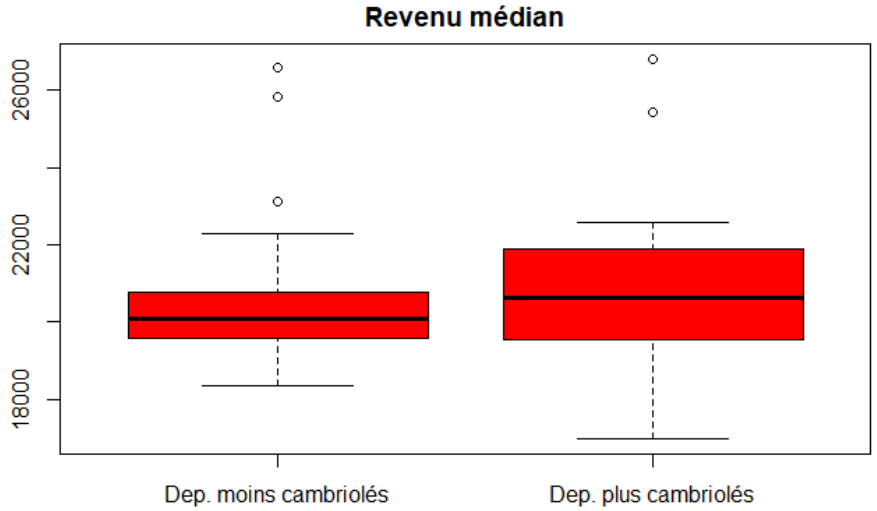

Figure II.1 - Présentation du revenu médian en 2016 selon 2 catégories de départements

Nous pouvons déjà observer que les revenus médians dans ces 2 types de départements suivent une distribution symétrique.
Nous remarquons que les départements les moins cambriolés ont une moyenne des revenus (disponibles par unité de consommation) à environ 20 000€ par an, ce qui est inférieur aux départements les plus cambriolés, qui ont eux une moyenne des revenus égale à environ 20 500€ par an. Il y a 500€ de différence en revenu sur une année ce qui représente environ 42€ par mois entre les deux différents types de départements. Cette différence peut changer beaucoup pour certains foyers, mais nous ne pouvons pas faire de réelles conclusions quant aux fait que les départements les plus cambriolés sont largement plus aisés/riches que les autres départements en moyenne.
Cependant, nous pouvons observer qu’il y a une variance largement supérieure des revenus pour les départements les plus cambriolés, presque deux fois supérieure à celle des autres départements. Ce que nous pouvons en conclure, c’est qu’il n’y a pas de réelle ‘classe sociale’ associée aux départements les plus cambriolés, les revenus sont assez disparates entre les différents départements, mais 50% d’entre eux ont un revenu entre environ 19 000€ et 21 900€ par an alors que pour les départements les moins cambriolés, cela se situe entre 19 000€ et 20 500€. Ainsi, on peut remarquer qu’en termes de variance des données, les départements les plus cambriolés en ont une supérieure aux départements les moins cambriolés.

Ces analyses univariées nous permettent d’avoir une première approche quant à l’étude de nos variables par rapport à notre variable indicatrice.

2) Statistiques descriptives bivariées

Pour faire des statistiques descriptives bivariées, nous cherchons à décrire l’association entre plusieurs variables quantitatives, y compris en regardant les corrélations et en analysant un nuage de points.

- Étudions la corrélation de chaque variable entre elles :

On remarque que la variable réponse de notre équation, Y_cambriolagessur10000, est légèrement positivement corrélée avec le revenu médian à 0.15 ainsi que le taux de pauvreté à 0.20 et le taux de chômage moyen à 0.33. C’est-à-dire que nous pouvons dire qu' il existe une relation de dépendance entre ces variables explicatives et notre variable réponse.

De plus, il y a une forte corrélation positive entre le niveau de pauvreté et le taux de chômage moyen à 0.76, ainsi qu'une corrélation négative entre le revenu médian et le taux de pauvreté à -0.59 ainsi que le nombre de brevets/bepc à -0.60, ce qui semble assez logique.
Pour les autres corrélations, nous pouvons regarder le graphique suivant :

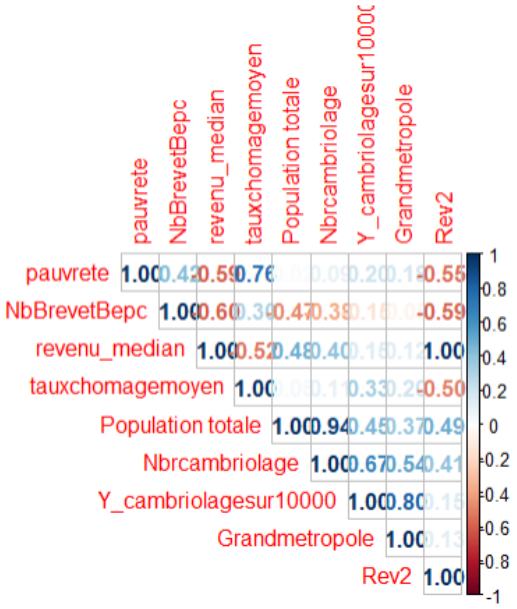

Figure II.2 - Visualisation graphique de la matrice de corrélation

- À présent, étudions notre variable réponse Y, le taux de cambriolages pour 10000 habitants, en fonction du taux de chômage moyen.

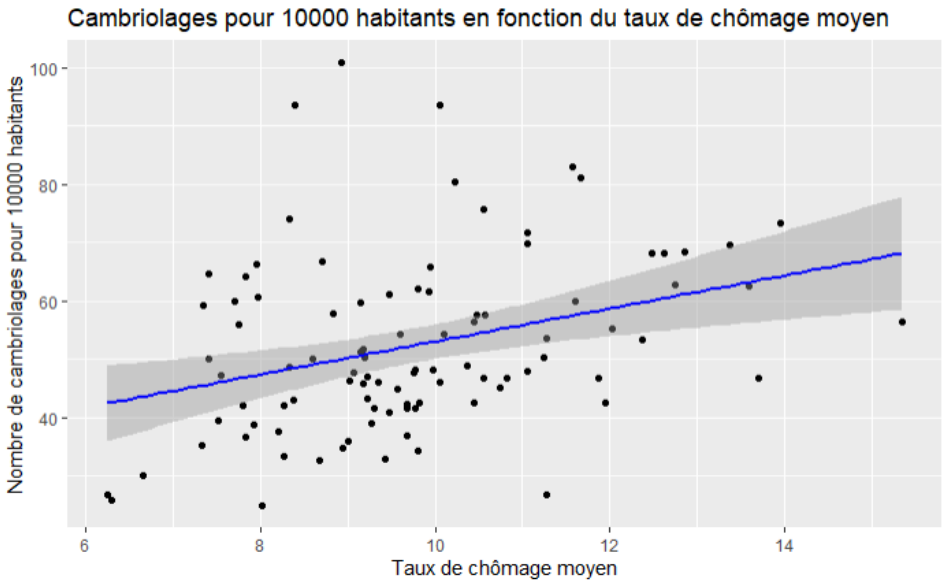

Figure II.3 - Nuage de points du nombre de cambriolages sur 10 000 habitants en fonction de taux de chômage moyen

Nous remarquons directement qu'il existe une relation linéaire positive entre les deux variables. Après avoir analysé la relation positive entre Y, notre variable réponse, et le taux de chômage moyen, nous pouvons conclure de l'implication des variables sur le modèle : un modèle linéaire est adapté à l'étude de celui-ci.

III. Modélisation économétrie

1) Estimation du modèle par la méthode des MCO

a) Analyse de l'équation en modèle linéaire niveau-niveau

On rappelle l'équation du modèle :
Y_cambriolagesur10000 = B0 + B1pauvrete + B2NbBrevetBepc + B3revenu_median + B4tauxChomagemoyen + B5Grandmetropole + Beta6Rev2

Commentons les résultats du modèle de la sortie de R :

```
Call:
lm(formula = Y_cambriolagesur10000 ~ pauvrete + NbBrevetBepc +
    revenu_median + tauxchomagemoyen + Rev2 + Grandmetropole,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.060  -5.632  -0.281   3.768  27.779

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.740e+02  1.322e+02  -2.072   0.0412 *
pauvrete         5.673e-01  6.098e-01   0.930   0.3547
NbBrevetBepc    -1.733e+00  1.200e+00  -1.444   0.1522
revenu_median    2.539e-02  1.126e-02   2.254   0.0266 *
tauxchomagemoyen 2.512e+00  7.715e-01   3.256   0.0016 **
Rev2            -5.297e-07  2.472e-07  -2.143   0.0349 *
Grandmetropole   2.279e+01  1.969e+00  11.578   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.201 on 89 degrees of freedom
Multiple R-squared:  0.727,     Adjusted R-squared:  0.7086
F-statistic: 39.51 on 6 and 89 DF,  p-value: < 2.2e-16
```

```
Resultats d'estimation (modèle niveau-niveau)
================================================================================
                L'influence des variables sur les licences modèle niveau-niveau
                --------------------------------------------------------------
                                    Y_cambriolagesur10000
--------------------------------------------------------------------------------
pauvrete                                      0.567
                                             (0.610)

NbBrevetBepc                                 -1.733
                                             (1.200)

revenu_median                                0.025**
                                             (0.011)

tauxchomagemoyen                            2.512***
                                             (0.771)

Rev2                                        -0.00000**
                                            (0.00000)

Grandmetropole                              22.793***
                                             (1.969)

Constant                                    -273.972**
                                            (132.244)

--------------------------------------------------------------------------------
Observations                                    96
R2                                           0.727
Adjusted R2                                  0.709
================================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Figure III.1 - Résumé du modèle régression linéaire niveau - niveau

- **Qualité d'ajustement du modèle** : on observe le coefficient de détermination ajusté nommé 'Adjusted R-squared' est à 0.7086. Cela veut dire que 70,86% de la variance de notre variable réponse Y, c'est-à-dire du nombre de cambriolages pour 10 000 habitants, est expliquée par le modèle c'est-à-dire par les variables explicatives. Cela est suffisant pour pouvoir mener à bien notre étude.
- **Validité globale du modèle** : on analyse le test de validité global (i.e le test de Fisher) qui a pour hypothèse nulle H0: "Tous les coefficients sont nuls sauf la constante". On a que 'F-statistic' est à 39.51 et que la 'p-value' est inférieure à 2.2e-16 c'est-à-dire à 5% . Ainsi, on peut rejeter H0 : le modèle est globalement satisfaisant.
- **Variables significatives** : on observe le test de Student avec l'hypothèse nulle est H0 : "Beta i = 0" où i=1,..,6 afin de savoir si une variable est significative ou non pour le modèle. En effet, si la p-value de celle-ci dans ce test est inférieur à 5%, alors nous pouvons rejeter H0 et donc, affirmer que cette variable est significative. Ainsi, d'après la sortie de R ci-dessus, on peut conclure que le revenu médian, le taux chômage moyen, le revenu médian au carré, la variable grande métropole représentant notre variable indicative, sont les variables significatives de notre modèle et ont donc un effet significatif sur le nombre de cambriolages pour 10000 habitants. Ainsi, on peut analyser que :
    - Si le revenu médian augmente de 1€, alors le nombre de cambriolages pour 10000 habitants variera de 0.02539 cambriolages.
    - Si le taux de chômage moyen augmente de 1%, alors le nombre de cambriolages pour 10000 habitants augmentera de 2.512 cambriolages.
    - Si le carré du revenu médian augmente de 1€, alors le nombre de cambriolages pour 10000 habitants variera de 0.02539 - 0.0000005297 cambriolages.

b) Comparaison avec le modèle linéaire log-log

Afin d'observer si nous avons bien choisi le modèle pour étudier notre équation (niveau-niveau précédemment), comparons le avec modèle log-log. On a :

```
Call:
lm(formula = log(Y_cambriolagesur10000) ~ log(pauvrete) + log(NbBrevetBepc) +
    log(revenu_median) + log(tauxchomagemoyen) + log(Rev2) +
    Grandmetropole, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45929 -0.09448 -0.00889  0.09979  0.37102

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -6.29392    3.83297  -1.642   0.1041
log(pauvrete)        -0.05712    0.13891  -0.411   0.6819
log(NbBrevetBepc)    -0.14821    0.13733  -1.079   0.2834
log(revenu_median)    0.89760    0.35507   2.528   0.0132 *
log(tauxchomagemoyen) 0.70141    0.14206   4.937 3.62e-06 ***
log(Rev2)                NA         NA      NA       NA
Grandmetropole        0.40745    0.03753  10.856  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1597 on 90 degrees of freedom
Multiple R-squared:  0.7091,    Adjusted R-squared:  0.693
F-statistic: 43.88 on 5 and 90 DF,  p-value: < 2.2e-16
```

```
Resultats d'estimation (modèle log-log)
================================================================================
                        L'influence des variables sur les licences modèle log-log
                        ---------------------------------------------------------
                                    log(Y_cambriolagesur10000)
--------------------------------------------------------------------------------
log(pauvrete)                                -0.057
                                             (0.139)

log(NbBrevetBepc)                            -0.148
                                             (0.137)

log(revenu_median)                           0.898**
                                             (0.355)

log(tauxchomagemoyen)                        0.701***
                                             (0.142)

log(Rev2)


Grandmetropole                               0.407***
                                             (0.038)

Constant                                     -6.294
                                             (3.833)

--------------------------------------------------------------------------------
Observations                                   96
R2                                           0.709
Adjusted R2                                  0.693
================================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Figure III.2 - Résumé du modèle régression linéaire log - log

On a que NA signifie que la variable log(Rev2) est un cas particulier de colinéarité. (corrélation = 1 avec log(revenu_median)).
Regardons la qualité d'ajustement du modèle : on observe que le coefficient de détermination ajusté nommé 'Adjusted R-squared' est à 0.693, c'est-à-dire à 69.3 %, ce qui est inférieur au coefficient de détermination ajusté du modèle niveau-niveau, qui lui est à 70.86%. Ainsi, nous avons fait le bon choix quant au choix du modèle niveau-niveau à la place du modèle log-log, car il explique plus de variance.

2) Test de spécification de Ramsey

Nous réalisons le test de Ramsey, qui a comme fonction de voir si le modèle subit l'omission d'une ou plusieurs variables pertinentes en introduisant une ou plusieurs variables fictives. Si on analyse alors que, la ou les variables fictives sont significatives, alors des variables susceptibles d'influencer les variations de la variable dépendante seront introduites.
Ce test va estimer, avec le modèle des MCO, la série des valeurs prédites par le modèle (yp), puis va réestimer avec les MCO le modèle augmenté des variables yp au carré et yp au cube et enfin, va tester l'égalité à 0 des coefficients de ces 2 variables.
L'hypothèse nulle est alors H0: "Deux coefficients des variables yp au carré et yp au cube sont nuls, le modèle est correctement spécifié."

```
> modele_nvnvt <- lm(Y_cambriolagesur10000 ~ pauvrete +
NbBrevetBepc + revenu_median + tauxchomagemoyen + Rev2 +
Grandmetropole + yp2 + yp3, data=data )
> summary(modele_nvnvt )

Call:
lm(formula = Y_cambriolagesur10000 ~ pauvrete + NbBrevetBepc +
    revenu_median + tauxchomagemoyen + Rev2 + Grandmetropole +
    yp2 + yp3, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1666  -4.9258  -0.5084   3.5222  27.7793

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.062e+03  1.681e+03  -1.821   0.0720 .
pauvrete         6.134e+00  3.405e+00   1.802   0.0751 .
NbBrevetBepc    -1.801e+01  9.932e+00  -1.814   0.0732 .
revenu_median    2.686e-01  1.468e-01   1.830   0.0707 .
tauxchomagemoyen 2.605e+01  1.420e+01   1.835   0.0700 .
Rev2            -5.606e-06  3.064e-06  -1.830   0.0707 .
Grandmetropole   2.489e+02  1.357e+02   1.834   0.0701 .
yp2             -1.780e-01  1.091e-01  -1.632   0.1064
yp3              1.045e-03  6.568e-04   1.591   0.1152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.161 on 87 degrees of freedom
Multiple R-squared:  0.7358,    Adjusted R-squared:  0.7115
F-statistic: 30.28 on 8 and 87 DF,  p-value: < 2.2e-16
```

```
> anova(modele_nvnv,modele_nvnvt )

Analysis of Variance Table

Model 1: Y_cambriolagesur10000 ~ pauvrete + NbBrevetBepc + revenu_median +
    tauxchomagemoyen + Rev2 + Grandmetropole
Model 2: Y_cambriolagesur10000 ~ pauvrete + NbBrevetBepc + revenu_median +
    tauxchomagemoyen + Rev2 + Grandmetropole + yp2 + yp3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     89 5986.0
2     87 5793.8  2    192.23 1.4433 0.2418
```

Figure III.3 - Résumé du modèle augmenté

Conclusion du test:

```
        RESET test

data:  modele_nvnv
RESET = 1.4433, df1 = 2, df2 = 87, p-value = 0.2418


        RESET test

data:  modele_loglog
RESET = 3.5246, df1 = 2, df2 = 87, p-value = 0.03374
```

Figure III.8 - Résultats du test de Ramsey

Pour le modèle niveau-niveau, (que nous avons choisi comme le meilleur modèle), on a : p-value > 5%, donc nous acceptons H0, le modèle est correctement spécifié.
Par exemple, si nous faisons le test Ramsey sur le modèle log-log, p-value < 5%, donc nous rejetons H0, le modèle log-log est mal spécifié.

3) Test de Chow

Nous faisons des sous-groupes d'observations pour créer 2 échantillons: un groupe avec les départements les plus cambriolés, ici nous l'appelons Grande Métropole et un groupe avec ceux qui ne le sont pas, qu'on appellera Non Grande Métropole. Puis, nous allons tester si les coefficients de régression sont identiques ou différents sur ces 2 échantillons. Le test de Chow est un test statistique et économétrique qui est mis en place afin de déterminer si les coefficients de deux séries linéaires des deux échantillons sont égaux. Les coefficients sont établis par régression linéaire. On a :

Les départements où il y a plus de cambriolages -> Grande Métropole

```
Call:
lm(formula = Y_cambriolagesur10000 ~ pauvrete + NbBrevetBepc +
    revenu_median + tauxchomagemoyen + Rev2 + Grandmetropole,
    data = fichier1)

Residuals:
    Min      1Q  Median      3Q     Max
-14.105  -7.139  -1.607   3.180  20.731

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -5.984e+02  3.428e+02  -1.746   0.0937 .
pauvrete          1.974e+00  1.631e+00   1.210   0.2380
NbBrevetBepc     -4.871e+00  3.569e+00  -1.365   0.1850
revenu_median     5.876e-02  2.888e-02   2.034   0.0531 .
tauxchomagemoyen  7.280e-01  1.976e+00   0.368   0.7158
Rev2             -1.291e-06  6.281e-07  -2.056   0.0508 .
Grandmetropole           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 on 24 degrees of freedom
Multiple R-squared:  0.2104,    Adjusted R-squared:  0.04589
F-statistic: 1.279 on 5 and 24 DF,  p-value: 0.3054
```

Les départements où il y a moins de cambriolages -> Non Grande Métropole

```
Call:
lm(formula = Y_cambriolagesur10000 ~ pauvrete + NbBrevetBepc +
    revenu_median + tauxchomagemoyen + Rev2 + Grandmetropole,
    data = fichier2)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4613  -3.9688   0.3339   4.0144  11.3886

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.153e+02  1.371e+02  -0.841 0.403930
pauvrete          3.339e-01  6.105e-01   0.547 0.586471
NbBrevetBepc     -1.378e+00  1.229e+00  -1.121 0.266770
revenu_median     1.039e-02  1.172e-02   0.887 0.378629
tauxchomagemoyen  3.105e+00  7.825e-01   3.968 0.000196 ***
Rev2             -1.882e-07  2.583e-07  -0.729 0.468890
Grandmetropole           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.738 on 60 degrees of freedom
Multiple R-squared:  0.3821,    Adjusted R-squared:  0.3306
F-statistic: 7.421 on 5 and 60 DF,  p-value: 1.79e-05
```

Figure III.4 - Résumé des modèles de régression des sous-échantillons.

(NA signifie que les variables sont linéairement dépendantes, i.e. l'une des autres variables peut être exprimée comme une combinaison linéaire de ces variables, cela a du sens puisque Grandmetropole ne contient que la valeur 1 ou 0).

Le résultat du test de Chow est un test Fisher $F = \dfrac{[SCR-(SCR_1+SCR_2)]/\mathrm{ddl}_n}{(SCR_1+SCR_2)/\mathrm{ddl}_d}$

	Tous les départements	Grande métropole	Non grande métropole
SCR	5986.005	2744.196	2724.014
Observations	96	30	66
Ddl	89	24	60

FChow	1.59082
p-value	0.17155

Figure III.5 - Résultat du test de Chow

Dans le contexte du test de Chow, l'hypothèse nulle est qu'il n'y a pas de changement structurel, i.e. les coefficients sont égaux pour les deux groupes de données.
Donc, ici on ne rejette pas l'hypothèse nulle (p-value = 0.17155 > 5%). On accepte H0, ainsi il n'y a pas de différence structurelle entre les grandes métropoles et les non grandes métropoles, c'est-à-dire entre les départements les plus cambriolés et ceux moins cambriolés.

4) Analyse des résidus

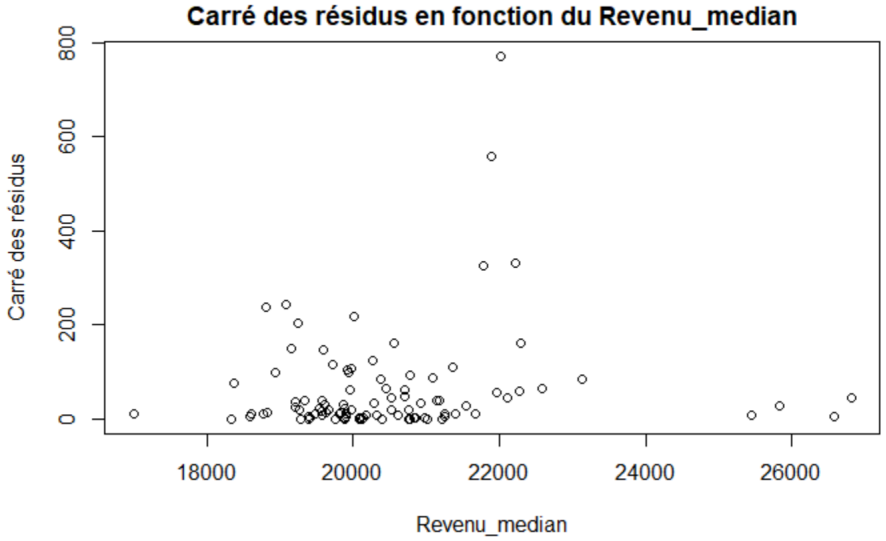

Figure III.6 - Carré des résidus en fonction du Revenu_median

On représente le carré des résidus en fonction de la variable revenu_median. On ne remarque pas de tendance spécifique. Il y a quelques valeurs extrêmes mais dans l'ensemble les valeurs sont regroupées dans un même intervalle (entre 0 et 200). Le nuage de points n'a pas de forme particulière donc il n'y a sans doute pas de problème d'hétéroscédasticité, nous allons le vérifier ultérieurement.

5) Validation des hypothèses des MCO

a) Hétéroscédasticité

Regardons tout d'abord si nous n'avons pas un problème d'hétéroscédasticité des résidus à l'aide d'un **test de White** sur notre modèle niveau-niveau. L'hypothèse nulle de ce test est qu'il y a homoscédasticité des résidus. On a :

```
        studentized Breusch-Pagan test

data:  modele_nvnv
BP = 4.179, df = 2, p-value = 0.1237
```

Figure III.7 - Résultat du test de White

Ici, on remarque que le test est égal à 4.179 pour une p-value égale à 0.1237 ce qui est supérieur à 5%. Ainsi, on ne rejette pas l'hypothèse nulle, donc, il y a homoscédasticité des résidus.

Vérifions cela à l'aide d'un **test de Goldfeld-Quandt** sur notre modèle niveau-niveau. L'hypothèse nulle est la même que le test de White : il y a homoscédasticité des résidus. On a :

```
        Goldfeld-Quandt test

data:  modele_nvnv
GQ = 1.6612, df1 = 38, df2 = 38, p-value = 0.06105
alternative hypothesis: variance increases from segment 1 to 2
```

Figure III.8 - Résultat du test de Goldfeld-Quandt

Le test de Golden-Quandt est égal à 1.6612 et nous avons une p-value de 0.06105, qui est supérieur à 5 %. Ainsi, nous pouvons accepter l'hypothèse H0 : il y a homoscédasticité des résidus. L'hypothèse des MCO est vérifiée, il n'y a pas besoin de faire une correction du modèle.

b) Autocorrélation

Regardons s'il y a un problème d'autocorrélation des résidus en effectuant un **test de Durbin-Watson** sur notre modèle niveau-niveau. L'hypothèse nulle est qu'il y a une non-autocorrélation des résidus.

```
        Durbin-Watson test

data:  modele_nvnv
DW = 1.9553, p-value = 0.3205
alternative hypothesis: true autocorrelation is greater than 0
```

Figure III.9 - Résultat du test de Durbin-Watson

Le test de Durbin-Watson est égal à 1.9553, DW=1,9553.
À l'aide d'une table de Durbin-Watson, on trouve que pour n=96, le nombre d'individus, et pour k=6, le nombre de variable explicatives de l'équation du modèle, que les bornes sont : [1.53,1.80] à un niveau de 5%. Ainsi, d'après la table :

DW	$[0\ ;\ d_1]$	$[d_1\ ;\ d_2]$	$[d_2\ ,\ 4-d_2]$	$[4-d_2\ ,\ 4-d_1]$	$[4-d_1\ ;\ 4]$
Analyse	$\rho > 0$ Auto-corrélation positive	Indéterminée	Hypothèse nulle valide	Indéterminée	$\rho < 0$ Auto-corrélation négative

Figure III.10 - Table de Durbin-Watson

En posant d1=1.53 et d2=1.80, on remarque que comme 4 - d1=2.47 et 4 - d2=2.2, on a :
d2 < DW < 4 - d2 : ainsi, l'hypothèse nulle est valide et donc les résidus ne sont pas corrélés. L'hypothèse des MCO est bien vérifiée, il n'y pas besoin de faire une correction du modèle.

Conclusion

Cette étude a pour objectif l'analyse des déterminants des cambriolages en France. Les jeux de données proviennent de l'INSEE, l'Institut National de la Statistique et des Études Économiques.

Pour atteindre l'objectif fixé, une analyse univariée puis bivariée est effectuée sur les données comportant les nombre de cambriolages sur 10 000 habitants de chaque département, les niveaux de diplômes, le taux de pauvreté au seuil de 60%, la distribution moyenne des revenus, le taux de chômage moyen, et le facteur de Grande métropole.

Un modèle régression linéaire niveau - niveau a été estimé, et a été montré comme le meilleur modèle avec une bonne spécification.

Pour conclure, grâce à cette étude, nous pouvons dire que notre modèle met en avant que le nombre de cambriolages pour 10 000 habitants par département en France métropolitaine est impacté par des variables extérieures.

Tout d'abord, ce taux de cambriolages est impacté par la moyenne des revenus disponibles par unité de consommation et composition du revenu disponible des personnes. En effet, le revenu d'une personne est un critère déterminant quant à sa qualité de vie et son pouvoir d'achat. Cependant, parfois, le revenu acquis d'une personne est en dessous de ses attentes, et donc peut affecter ses envies et ses ambitions, ou bien encore, peut ne pas être suffisant pour subvenir à ses besoins fondamentaux. Ces exemples parmi d'autres peuvent être des motivations quant au fait de commettre un crime tel qu'un cambriolage. C'est pourquoi, le revenu moyen d'une personne impact directement le nombre de cambriolages commis.

Ensuite, nous avons aussi constaté que le taux de chômage moyen impacte le taux de cambriolages. En effet, nous avons pu alors estimer qu'il y a une relation positive entre le taux de chômage et le taux de cambriolages, plus des personnes ne sont pas sur le marché du travail, plus la probabilité qu'un cambriolage se déroule augmente. Cela peut être en lien avec le temps passé sans réinsertion professionnelle d'une personne, les missions proposées, la pénibilité des travaux et le salaire associé, il y a tant de possibilités qui peuvent participer à pousser une personne à commettre un crime tel qu'un cambriolage.

De plus, nous avons analysé que le niveau de pauvreté, au seuil de 60%, n'était pas significatif quant au taux de cambriolages commis, ce qui nous a étonné. Nous nous attendions à trouver une forte significativité entre le nombre de cambriolages et le taux de pauvreté, car, moins une personne a de ressources, plus elle peut être plus apte à cambrioler. Mais ici, nous ne pouvons pas conclure sur le fait que la pauvreté a un effet quelconque sur le taux de cambriolages. Peut-être que pour un autre seuil de pauvreté, un autre niveau de significativité aurait pu être présent.

De même pour le niveau d'éducation, il n'est pas significatif sur notre modèle. En effet, ici notre variable étudiée était la part des personnes dont le diplôme le plus élevé est le bepc ou le brevet, dans la population non scolarisée de 15 ans ou plus, et elle n'a pas d'impact sur le taux de cambriolages. Encore une fois, c'est étonnant, cependant, nous l'avions observé dans l'analyse de la statistique descriptive sur notre variable indicatrice. Car nous aurions pu nous attendre à trouver que suivant le nombre de personnes ayant un certain niveau d'étude, cela aurait pu avoir une implication sur le nombre de cambriolages effectués. Peut-être qu'avec un autre type de diplôme, tel que le niveau baccalauréat, nous aurions pu trouver que cela a un effet sur le taux de cambriolages.