

05 JANVIER 2023


DETERMINANTS DE L'INSECURITE EN FRANCE

PROJET ECONOMETRIE ET MODELISATION ECONOMIQUE

REALISE PAR

YANN KIBAMBA & NABIL LOUDAOUI

MASTER 1 Mathématiques Appliquées, Statistique



SOMMAIRE

Introduction

1. Problématique & Modèle

1.1 Les variables

1.1.1 Présentation des variables

1.1.2 Variable indicatrice

1.1.3 Equation à estimer

1.2 Statistique descriptive univarié

1.2.1 Résumé des données

1.2.2 Représentations graphiques

1.2.3 Algorithme progressif-rétrogressif

1.3 Statistique descriptive bivarié

1.3.1 Distribution bivarié

1.3.2 Corrélation

1.3.3 Analyse & choix des variables du modèle

2. Validation & Estimation du modèle

2.1 Choix meilleur modèle

2.2 Analyse des résidus

2.3 Tests

2.3.1 Test de spécification (Test de Ramsey)

2.3.2 Test d'autocorrélation

2.3.3 Test Hétéroscédasticité (White/Goldfeld Quandt)

2.3.4 Changement de structure

Conclusion

Bibliographie

Annexe

Introduction

« Au cours des 10 dernières années de 2008 à 2018, les vols ont considérablement augmenté. En 2008, c'est 386 000 vols et tentatives de vols contre la résidence principale qui ont été enregistré en France. Dix ans plus tard, en 2018, le nombre de vols a augmenté de 171 000. C'est une augmentation de plus de 44%, des cambriolages en France ! » (www.portes-et-serrures.fr). Véritable fléau de nos sociétés actuelles, les cambriolages sont en constante croissance ces dernières années en France. Bijoux, instruments monétaires (argent, carte bancaire, chéquier etc...) et matériels multimédias constituent les trois premières sources d'objets de convoitises des ravisseurs qu'ils dérobent dans les foyers français. Les cambriolages font partie intégrante des infractions liés à l'insécurité dont sont victimes individus et ménages français. L'insécurité compte d'autres facteurs comme les violences physiques volontaires auquel nous nous intéresserons également.

D'après l'enquête CVS (Cadre de Vie et Sécurité) réalisé chaque année par l'Institut National de la Statistique et des Etudes Economiques (INSEE) et en collaboration avec l'ONDPR (Observatoire Nationale de la Délinquance et des Réponses Pénales) et SSMSI (Service Statistique Ministériel de la Sécurité Intérieur), « En 2018, près de 221 000 personnes de quinze ans ou plus ont été enregistrées comme victimes de coups et blessures volontaires par la police et la gendarmerie en France métropolitaine, ce qui représente 4 victimes pour 1 000 habitants. ». Tout comme le nombre de cambriolage, le nombre de coups et blessures volontaires est également en hausse sur le territoire tant intrafamiliale qu'extrafamiliale.

L'insécurité peut se définir comme un environnement physique ou social favorisant les atteintes aux personnes et aux biens (Larousse). Dans ce contexte, il est intéressant de comprendre les déterminants des actes d'insécurité en France car ces derniers peuvent impacter la qualité de vie urbaine, l'environnement physique ou bien social liés aux personnes et aux biens.

L'objet du présent rapport est de s'interroger sur les principaux facteurs contribuant aux faits d'insécurité sur le territoire métropolitain, plus précisément les actes de cambriolages et actes de coups et blessures volontaires pour l'année civile 2018. En d'autres termes, quels sont les déterminants responsables des actes d'insécurité au travers des départements français ?

A la suite d'une collecte de variables diverses et variées pour chacun des départements français métropolitain, nous pourrons décrire les actes dérogeant au cadre de santé et sécurité des citoyens où nous procéderons comme ci-après :

Le choix et l'analyse des variables collectées pour la modélisation économique du phénomène constitueront le premier volet de l'étude. Dans un second temps, nous choisirons le modèle le plus adéquat pour mettre en exergue la relation entre les variables sélectionnées et l'insécurité. Nous veillerons à ce que notre modèle respecte les propriétés fondamentales d'un modèle de régression linéaire tel que l'homoscédasticité.

Pour nos tâches, nous nous appuieront sur le logiciel Rstudio. Le terme insécurité pour cette étude désignera uniquement les cambriolages et les violences physiques volontaires.

1. Problématique & Modèle

1.1 Les variables

Afin d'apporter une réponse économique à la problématique, nous avons recueilli un ensemble de 13 variables pour 96 individus correspondant aux départements français métropolitains dans le but de créer notre base de données. Rattaché à chaque circonscription du territoire pour l'année d'étude 2018, les données sont individuelles et non temporelles. A noter qu'aucune variable ne compte de valeurs manquantes.

1.1.1 Présentation des variables

Ci-dessous, le descriptif et l'origine des variables :

Code variable	Définition de la variable	Unité de la variable	Source de l'information
Delit (variable endogène)	Moyenne entre le taux de cambriolages pour 1000 logements et le taux de coups et blessures volontaires pour 1000 personnes (par département)	En (%)	www.insee.fr
Population	Nombre d'habitants recensés dans un département en 2018	Nombre d'habitant en milliers	https://public.opendatasoft.com/
Densite_pop	Nombre d'habitants au km ² dans le département	Habitants/km ²	www.observatoire-des-territoires.gouv.fr
Revenu_median	Revenu médian de la population active du département	En euros	www.insee.fr
X15.29ans	Part des habitants ayant entre 15 et 29 ans dans le département	En (%)	www.observatoire-des-territoires.gouv.fr
X60.74ans	Part des habitants ayant entre 60 et 74 ans dans le département	En (%)	www.observatoire-des-territoires.gouv.fr
Taux_pauvrete	Part des ménages dans le département vivant en dessous du seuil de pauvreté. <i>(Ce seuil est fixé à 60% du niveau de vie médian en France.)</i>	En (%)	www.insee.fr
Urbanisation	Taux d'urbanisation, la part de la population vivant en ville par rapport à la population totale.	En (%)	www.insee.fr
Scolarisation	Taux de scolarisation, la part des peu ou pas diplômé parmi les 16-24 ans qui ne sont pas en études en 2017-2018.	En (%)	www.insee.fr
Taux_chomage	Pourcentage des personnes de la population active qui sont au chômage.	En (%)	www.observatoire-des-territoires.gouv.fr
Immigration	Correspond à un taux d'immigration, la part de la population immigrée selon les pays de naissance autre que la France (UE et Reste du monde).	En (%)	www.insee.fr
Police	Taux de policiers et gendarmes pour 10 000 habitants dans les départements.	Nombre de policiers + gendarmes /10 000 habitants	www.insee.fr
Politique	Orientation politique départementale	/	www.interieur.gouv.fr

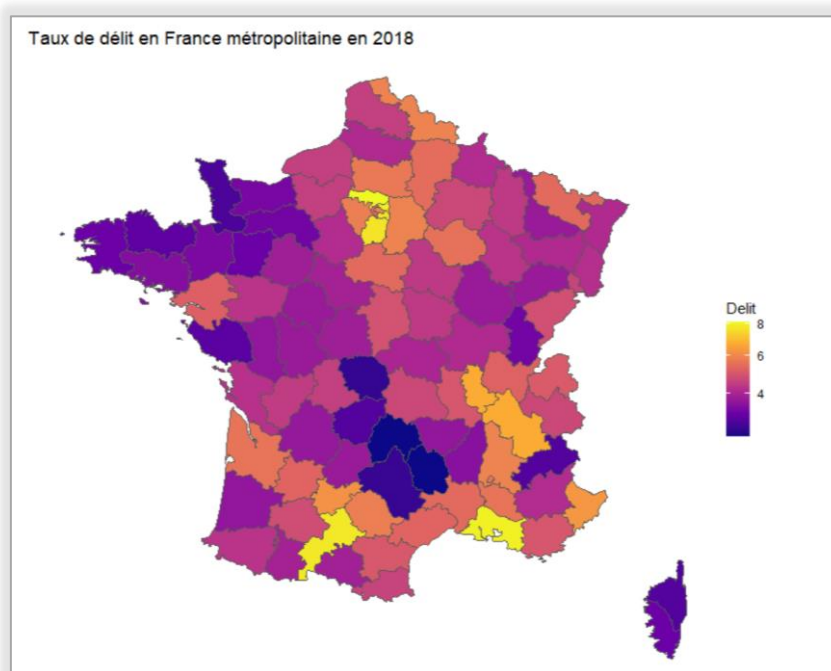
1.1.2 Changement de structure

La variable Politique est codée avec les chaînes de caractères « gauche » et « droite » pour chaque département dans la base de données en fonction de la tendance de la politique. Cette variable aura un rôle de variable indicatrice. Cette variable sera par la suite labélisée en facteur 0 et 1. La variable indicatrice Politique permettra d'observer s'il existe un changement de structure par rapport à notre échantillon. C'est-à-dire qu'on pourra évaluer l'impact de la variable sur les deux groupes construit à partir de cette dernière. En d'autres termes, existe-t-il une différence structurelle entre les départements de « gauche » et les départements de « droite » ?

La région Ile-de-France et ses départements ne ressemble à aucun autre territoire français. C'est pourquoi l'étude d'un potentiel changement de structure y sera également abordée.

1.1.3 Equation à estimer

Notre équation à estimer pour le moment regroupe toutes les variables exogènes en fonction de l'endogène (Delit). Nous analyserons par la suite les différentes variables explicatives afin d'en déterminer les plus pertinentes à conserver afin d'expliquer le taux de délit dans les départements (voir représentation ci-dessous).



1.2 Statistique descriptive univarié

1.2.1 Résumé des données

A l'aide de la fonction « summary » du logiciel Rstudio, il est possible d'obtenir un résumé statistique de la distribution des variables quantitatives d'un jeu de données. Les résultats obtenus de la fonction sur notre dataset sont présentés ci-dessous :

Delit	Population	Densite_pop	
Min. :2.200	Min. : 75.46	Min. : 14.81	
1st Qu.:3.700	1st Qu.: 299.22	1st Qu.: 50.04	
Median :4.400	Median : 538.86	Median : 83.02	
Mean :4.562	Mean : 677.27	Mean : 565.85	
3rd Qu.:5.325	3rd Qu.: 850.00	3rd Qu.: 162.40	
Max. :8.100	Max. :2613.87	Max. :20641.38	
Revenu_median	X15.29ans	X60.74ans	
Min. :17740	Min. :12.30	Min. :11.50	
1st Qu.:20415	1st Qu.:14.30	1st Qu.:16.57	
Median :21010	Median :15.60	Median :18.10	
Mean :21395	Mean :16.14	Mean :18.09	
3rd Qu.:21908	3rd Qu.:18.00	3rd Qu.:20.10	
Max. :28270	Max. :23.70	Max. :23.30	
Taux_pauvrete	Urbanisation	Scolarisation	
Min. : 8.90	Min. : 21.40	Min. :14.00	
1st Qu.:12.38	1st Qu.: 56.00	1st Qu.:18.68	
Median :14.40	Median : 66.75	Median :21.25	
Mean :14.53	Mean : 68.10	Mean :21.20	
3rd Qu.:15.72	3rd Qu.: 81.05	3rd Qu.:23.10	
Max. :28.40	Max. :100.00	Max. :30.60	
Taux_chomage	Immigration	Police	Politique
Min. : 8.30	Min. : 2.500	Min. : 16.00	droite:65
1st Qu.:11.28	1st Qu.: 5.050	1st Qu.: 25.00	gauche:31
Median :12.60	Median : 6.950	Median : 30.00	
Mean :12.73	Mean : 7.933	Mean : 32.23	
3rd Qu.:13.53	3rd Qu.: 9.625	3rd Qu.: 36.25	
Max. :18.90	Max. :30.700	Max. :117.00	

Nous observons dans un premier temps que la distribution des variables ne contient pas de valeurs négatives et ne semble pas contenir de valeurs aberrantes. Dans un second temps, on peut constater que la disparité connue entre les départements français se répercute sur la distribution des variables.

1.2.2 Représentations graphiques

Ci-dessous, les représentations graphiques des distributions des variables permettant d'observer la distribution et les potentiels individus atypiques pour chaque variable.

La variable endogène :

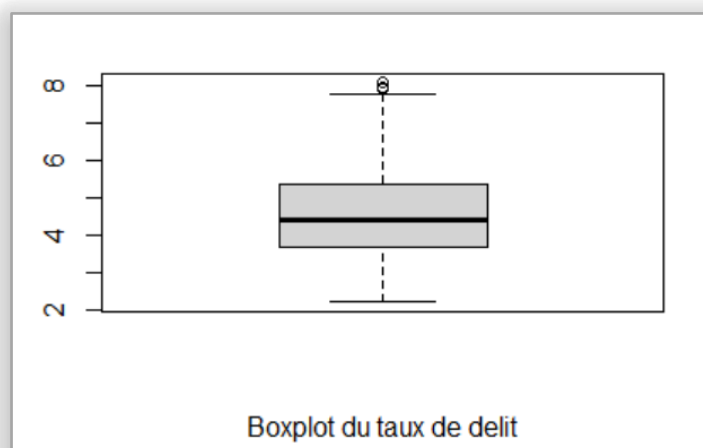


Figure n°1 – Distribution du taux de délit dans les départements

Le Boxplot met en évidence 4 valeurs "atypiques" qui sont très légèrement supérieures au seuil du $\min(\max, Q3+1.5*(Q3-Q1))$. Ces départements sont les Bouches-du-Rhône (13), la Haute-Garonne (31), la Seine-Saint-Denis (93) et le Val-de-Marne (94). Ces départements connaissent les taux de délits les plus importants par rapport à notre jeu de données.

Les variables exogènes : Population, Immigration, Revenu_median, Police et Taux de chômage.

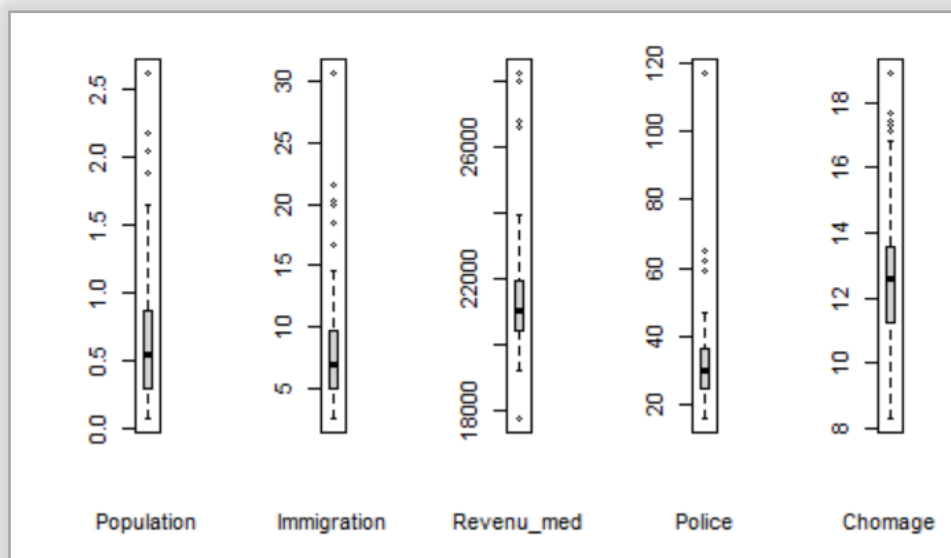


Figure n°2 – Distributions de variables exogènes (1)

Ces boîtes à moustaches mettent en évidence la pluralité d'individus « atypiques » aussi bien au-dessus du maximum déterminé par $\min(\max, Q3+1.5*(Q3-Q1))$ que du minimum déterminé par $\max(\min, Q1-1.5*(Q3-Q1))$. Par conséquent, on observe une importante variance entre les individus faisant écho au contraste d'un département à un autre.

Les variables exogènes : Urbanisation, Scolarisation, Densite_pop et Taux_pauvrete.

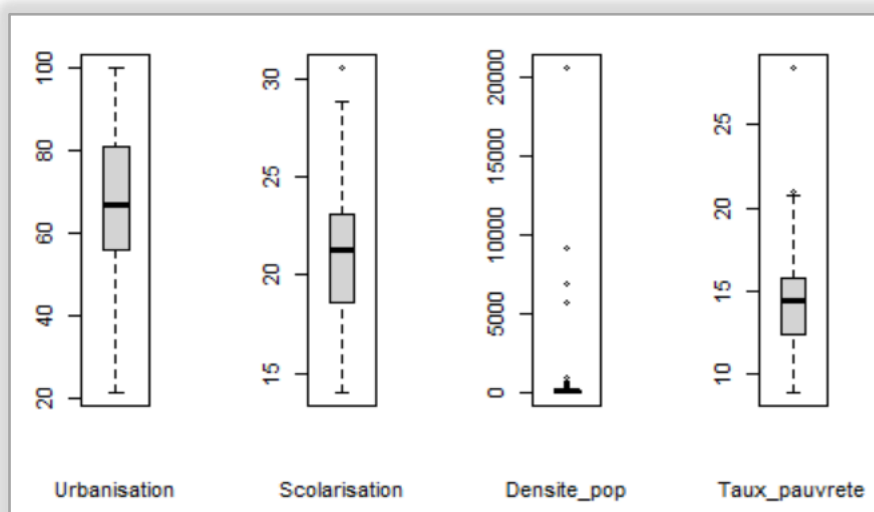


Figure n°3 – Distributions de variables exogènes (2)

La boîte à moustache de la densité de la population met en évidence la densité dominante du département de Paris et des départements de la première couronne parisienne. Environ 20 000 habitants/km² à Paris! Les trois autres boxplots comptent peu, voire pas de valeurs atypiques.

Au regard du nombre important de valeurs atypiques au sein de la distribution des variables, il sera probablement intéressant d'incorporer des logarithmes dans l'équation afin de réduire les valeurs atypiques et stabiliser les régresseurs.

1.2.3 Algorithme progressif-rétrogressif

Etant donné le nombre important de variables, il est nécessaire de porter un regard sur celles pouvant nous permettre d'expliquer notre variable Y (le taux de délit). Afin de nous guider dans le choix des variables explicatives, nous appliquerons les méthodes « Backward », « Forward » et « Stepwise » à l'aide du logiciel Rstudio. Ces méthodes ont pour but d'affiner le modèle. Leur processus est défini ci-dessous :

- **Backward** : Débute à partir du modèle complet et enlève une à une les variables les moins significatives pour le test de Student (pour $\alpha = 5\%$) jusqu'à obtenir un modèle uniquement de variables significatives.
- **Forward** : Débute d'un modèle contenant uniquement une constante et on ajoute pas à pas la variable explicative la plus significative permettant de minimiser l'AIC.
- **Stepwise** : Dérivé du Backward et Forward, cette méthode consiste à ajouter et à supprimer de manière itérative les variables dans le modèle prédictif afin de trouver le sous-ensemble de variable résultant en le modèle le plus performant. (Minimisant l'AIC)

Après application, les algorithmes progressif-rétrogressif renvoient les modèles suivants :

Backward : Immigration, Population, Revenu_median, Scolarisation et Taux_pauvrete.

Forward : Densite_pop, Immigration, Population, Revenu_median, Scolarisation, Taux_chomage et Taux_pauvrete.

Stepwise : Immigration, Population, Scolarisation et Taux_pauvrete

Ces résultats seront précieux en vue du choix des variables qui composeront le modèle prédictif final.

1.3 Statistique descriptive bivarié

1.3.1 Distribution bivarié

La médiane de la variable 60-74ans est nettement plus élevée que celle de la variable 15-29ans. On observe que 50% des départements français ont une part des 60-74 ans inférieure à 18% et l'autre moitié y est supérieure. Tandis que la médiane dans les départements des 15-29 ans est d'environ 16%. On peut en déduire que les départements comptent plus d'individus dans la tranche d'âge 60-74ans que dans la tranche 15-29ans. Il y a également plus variabilité concernant la variable 60-74ans par rapport à la variable 15-29ans.

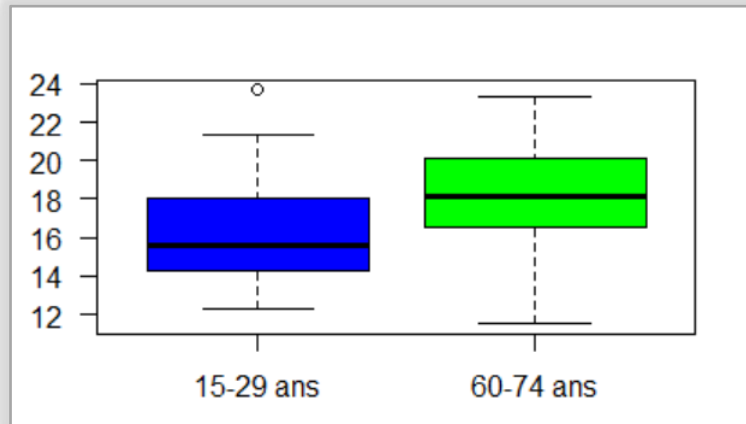


Figure n°4 – Distribution de l'âge dans les départements

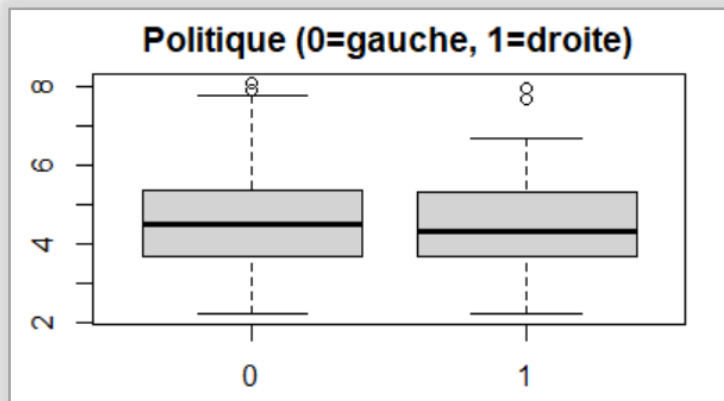


Figure n°5 – Distribution de la variable Politique

Ci-dessus, on observe la distribution du taux de délit dans les départements selon leur parti politique via des boxplots. Le jeu de données comptabilise 31 départements de gauche et 65 de droite. La médiane du taux de délit dans les départements de gauche est légèrement plus élevée que dans celle des départements de droite. Les deux boxplots disposent de deux individus atypiques. On note par ailleurs, que les départements de gauche ont une variance plus importante. Néanmoins, les deux distributions ne semblent pas éloignées.

1.3.2 Corrélation

La corrélation se traduit comme la manière dont deux variables ou plus vont évoluer ensemble. Le coefficient de corrélation oscille entre les -1 et 1. Une corrélation positive indique que les variables vont dans le même sens, c'est-à-dire croissent ou décroissent en même temps. Plus la relation est forte entre des variables corrélées positivement et plus le coefficient de corrélation tendra vers 1. A l'inverse, une corrélation négative signifie que les variables vont dans des sens opposés ; c'est-à-dire qu'à mesure qu'une décroît, l'autre croît et inversement. Plus la relation est forte entre des variables corrélées négativement et plus le coefficient de corrélation tendra vers -1. En conséquence, une valeur nulle signifie l'absence d'effet entre les variables.

Ainsi à travers une matrice de corrélation nous observerons les corrélations entre chaque paire de variables. Dans la constitution du modèle, il est important que plusieurs variables exogènes n'est pas de corrélation forte entre elles.

Précédemment, nous avons appliqué les algorithmes progressif-rétrogressif afin d'avoir un aperçu sur les meilleurs modèles potentiels, ceux minimisant l'AIC (ou BIC) et contenant uniquement des variables significatives. Ces derniers ne tiennent pas compte des potentielles corrélations entre paire de variables explicatives. C'est pourquoi cette matrice nous permettra d'ajuster notre décision concernant le choix des variables pour le modèle et anticiper les problèmes de colinéarité.

Les résultats de la matrice de corrélation obtenus sont présentés ci-dessous :

	Delit	Population	Densite_pop	Revenu_median	X15.29ans	X60.74ans	Taux_pauvrete	Urbanisation	Scolarisation	Taux_chomage	Immigration	Police
Delit	1	0.648	0.291	0.193	0.527	-0.631	0.285	0.681	0.362	0.310	0.715	0.100
Population	0.648	1	0.452	0.388	0.724	-0.709	0.056	0.788	0.035	0.127	0.533	0.199
Densite_pop	0.291	0.452	1	0.485	0.460	-0.373	0.136	0.365	-0.159	-0.032	0.574	0.649
Revenu_median	0.193	0.388	0.485	1	0.351	-0.419	-0.523	0.409	-0.439	-0.515	0.381	0.285
X15.29ans	0.527	0.724	0.460	0.351	1	-0.879	0.095	0.635	0.013	0.072	0.519	0.208
X60.74ans	-0.631	-0.709	-0.373	-0.419	-0.879	1	-0.010	-0.665	-0.016	0.041	-0.642	-0.047
Taux_pauvrete	0.285	0.056	0.136	-0.523	0.095	-0.010	1	0.114	0.731	0.749	0.351	0.122
Urbanisation	0.681	0.788	0.365	0.409	0.635	-0.665	0.114	1	0.106	0.203	0.663	0.129
Scolarisation	0.362	0.035	-0.159	-0.439	0.013	-0.016	0.731	0.106	1	0.776	0.185	-0.089
Taux_chomage	0.310	0.127	-0.032	-0.515	0.072	0.041	0.749	0.203	0.776	1	0.117	0.016
Immigration	0.715	0.533	0.574	0.381	0.519	-0.642	0.351	0.663	0.185	0.117	1	0.285
Police	0.100	0.199	0.649	0.285	0.208	-0.047	0.122	0.129	-0.089	0.016	0.285	1

Au vue du nombre de variable, il sera plus facile de visualiser les corrélations à l'aide d'un corrélogramme.

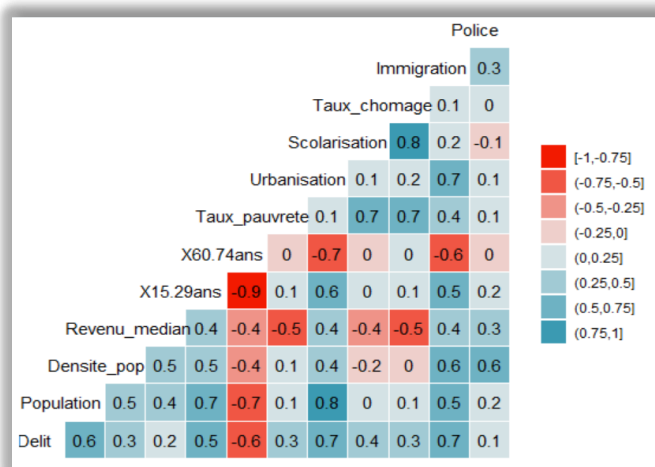


Figure n°6 – Corrélogramme

Mettant en évidence les corrélations plus ou moins fortes en fonction des couleurs, on constate que certaines variables explicatives collectées pour l'étude sont au final très peu corrélées voire corrélées négativement avec le taux de délit. C'est le cas par exemple de la variable Police ou encore X60.74ans. Cette faible corrélation avec l'endogène explique pourquoi ces variables n'ont pas été retenues par les algorithmes de sélection de modèles.

On remarque tout de même que le taux d'urbanisation est corrélé positivement à 0.7 avec le taux de délit. En analysant les corrélations entre les variables explicatives, on s'aperçoit que le taux d'urbanisation est fortement corrélé avec des variables proposées par les algorithmes de sélection, tel que le taux d'immigration à 0.7 et la variable population à 0.8. C'est pourquoi cette variable ne pourra être incluse dans un même temps dans le modèle malgré sa forte liaison avec l'endogène.

Au regard de la matrice de corrélation et des algorithmes de sélection nous conserverons les variables Immigration, Population, Scolarisation et Taux_pauvrete pour notre modèle. Ces variables sont toutes significatives dans le modèle ($\text{Délit} \sim \text{Population} + \text{Immigration} + \text{Scolarisation} + \text{Taux_pauvrete}$).

1.3.3 Analyse et interprétation des variables du modèle

Afin d'optimiser la spécification économétrique de notre modèle, nous appliquerons sur l'un des côtés ou sur les deux côtés de notre équation une transformation logarithmique. Nous observerons ainsi pour chacune des variables du modèle en fonction de l'endogène à l'aide de nuage de points les modèles niveau-niveau, niveau-log, log-log et log-niveau. Les droites de régression sont ajoutées à chaque nuage de point.

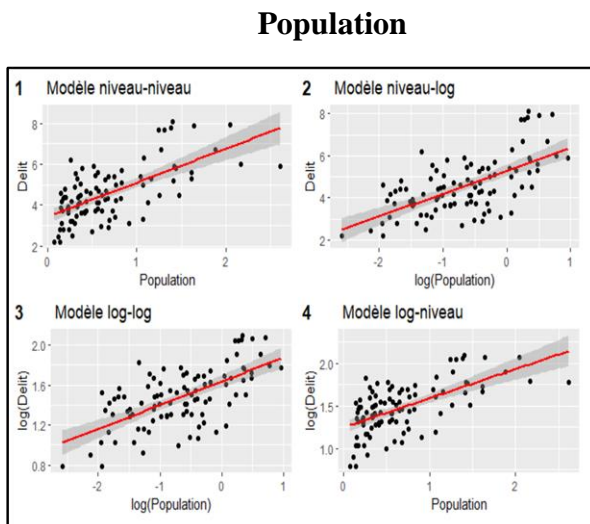


Figure n°7 – Nuage de points des différents modèles de la variable Population

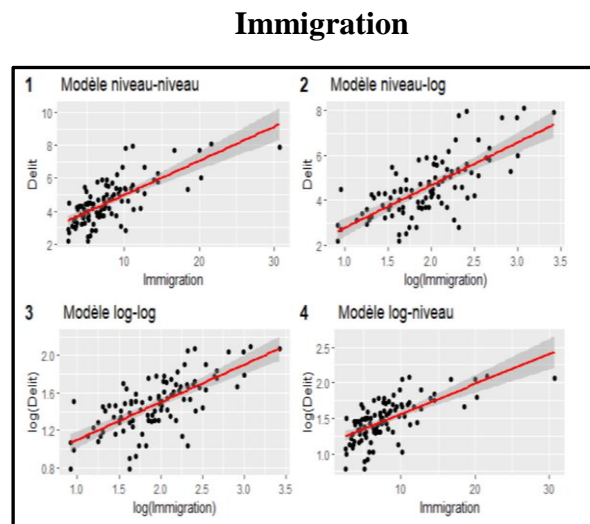


Figure n°8 – Nuage de points des différents modèles de la variable Immigration

Scolarisation

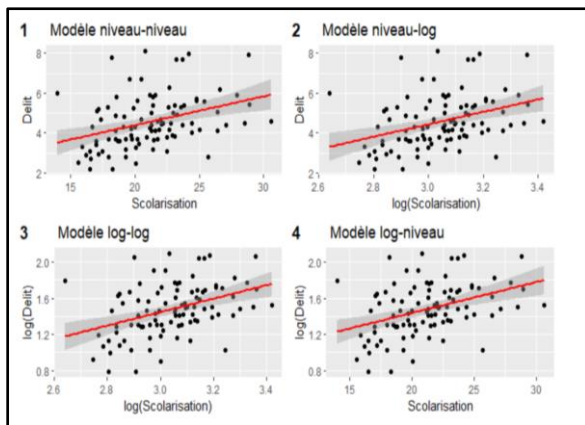


Figure n°9 – Nuage de points des différents modèles de la variable Scolarisation

Taux de pauvreté

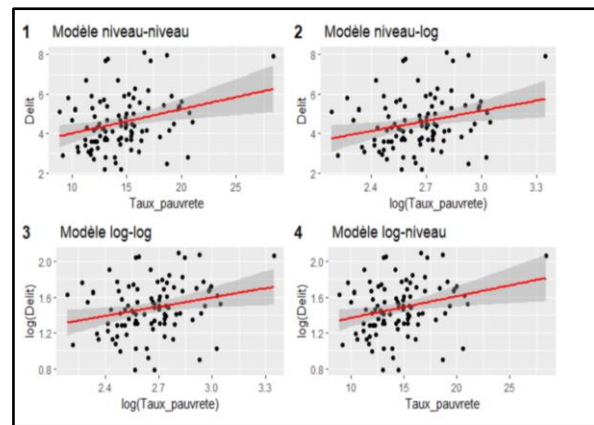


Figure n°10 – Nuage de points des différents modèles de la variable Taux_pauvrete

Pour tous les graphiques avec ou sans application du logarithme, il existe une relation linéaire entre la variable exogène et l'endogène. La variable endogène augmente en fonction de la variable exogène.

Dans l'ensemble, la relation linéaire semble être bien mieux expliquée par les modèles log-log et niveau-log. On observe un intervalle de confiance plus constant des points autour de la droite pour ces modèles. Le modèle niveau-log et le modèle log-log apportent une amélioration.

Par la suite, nous effectuerons 3 modélisations économétriques. Une modélisation niveau-niveau (modèle initiale), log-log et niveau-log. Pour chaque modèle, on évaluera la qualité de l'ajustement et on déterminera les variables significatives. Enfin, on testera si le modèle est globalement satisfaisant à l'aide de la commande "summary". Ainsi, on validera le meilleur modèle.

2 Validation & Estimation du modèle

Voici les 3 équations étudiés :

- Modèle niveau-niveau : $\text{Delit} \sim \log(\text{Population}) + \log(\text{Immigration}) + \text{Scolarisation}$
- Modèle niveau-log : $\text{Delit} \sim \log(\text{Population}) + \log(\text{Immigration}) + \log(\text{Scolarisation})$
- Modèle log-log : $\log(\text{Delit}) \sim \log(\text{Population}) + \log(\text{Immigration}) + \log(\text{Scolarisation})$

On considéra donc le modèle niveau-niveau avec l'application d'un logarithme sur les variables Population et Immigration afin de réduire l'effet de leurs observations atypiques. La variable taux_pauvrete avec ou sans l'application du logarithme a été retiré des modèles fautes de significativité après transformation du modèle. L'absence de significativité signifie que son apport d'information au modèle est négligeable.

2.1 Choix meilleur modèle

Modèle niveau-niveau

Modèle niveau-niveau (Estimation)	
Dependent variable:	
Delit	
log(Population)	0.739*** (0.106)
log(Immigration)	1.298*** (0.168)
Scolarisation	0.095*** (0.023)
Constant	-9.684*** (1.400)

Observations	96
R2	0.709
Adjusted R2	0.700
Residual Std. Error	0.730 (df = 92)
F Statistic	74.776*** (df = 3; 92)
=====	
Note:	*p<0.1; **p<0.05; ***p<0.01

Qualité d'ajustement : La valeur du R2 ajusté est de 0.7, c'est à dire que 70% de la variance de la variable Delit est expliqué par le modèle, autrement dit par les variables log(Population), log(Immigration) et Scolarisation. Le modèle est donc relativement en adéquation avec la variable endogène.

Modèle globalement satisfaisant : On suppose l'hypothèse nulle H0 de la statistique de Fisher tel que: "Tous les coefficients sont nuls sauf la constante".

La p-value est strictement significative au regard des 3 étoiles, donc bien inférieur au seuil de 5% alors le modèle est globalement satisfaisant.

Significativité des variables : On test individuellement sur chaque variable explicative la significativité avec comme hypothèse nulle H0 : "coefficient de la variable est égale à 0".

Les variables log(Population), log(Immigration) et Scolarisation ont une p-value inférieure à 5% (3 étoile). Ces variables ont un effet significatif sur le phénomène qu'on explique Y (Delit).

Interprétation du modèle :

- Si la variable log(Population) augmente d'une unité alors la variable Y (Delit) varie de 0.73857 habitants en milliers dans le département.
- Si la variable log(Immigration) augmente d'une unité alors la variable Y (Delit) varie de 1.29824 %.
- Pour toutes variations (positive ou négativement) d'un point de pourcentage de la variable Scolarisation, la variable Delit va varier respectivement (positivement ou négativement) de 0.09526%.

Modèle log-log

Modèle log-log (Estimation)	
Dependent variable:	
log(Delit)	
log(Population)	0.169*** (0.023)
log(Immigration)	0.257*** (0.036)
log(Scolarisation)	0.545*** (0.107)
Constant	-2.897*** (0.439)

Observations	96
R2	0.714
Adjusted R2	0.705
Residual Std. Error	0.157 (df = 92)
F Statistic	76.500*** (df = 3; 92)

Note:	*p<0.1; **p<0.05; ***p<0.01

Qualité d'ajustement : Nous avons un R2 ajusté de 0.705, c'est à dire que 70,5% de la variance de la variable log(Delit) est expliqué par le modèle, autrement dit par les variables Log (Population), log(Immigration) et log(Scolarisation). Le modèle est donc relativement en adéquation avec la variable Y.

Modèle globalement satisfaisant : On suppose l'hypothèse nulle H0 de la statistique de Fisher tel que: "Tous les coefficients sont nuls sauf la constante".

La p-value est strictement significative au regard des 3 étoiles, donc bien inférieur au seuil de 5% alors le modèle est globalement satisfaisant.

Significativité des variables : On test individuellement sur chaque variable explicative la significativité avec comme hypothèse nulle H0 : "coefficient de la variable est égale à 0".

Les variables log(Population), log(Immigration) et log(Scolarisation) ont une p-value inférieure au seuil de 5% donc ces variables ont un effet significatif sur le phénomène qu'on explique Y (log(Delit)).

Interprétation des variables :

- Si la variable log(Population) augmente de 1% alors log(Delit) augmente de 0.16%.
- Si la variable log(Immigration) augmente de 1% alors log(Delit) augmente de 0.27%.
- Si la variable log(Population) augmente de 1% alors log(Delit) augmente de 0.54%.

Modèle niveau-log

Modèle niveau-log (Estimation)	
Dependent variable:	
Delit	
log(Population)	0.751*** (0.106)
log(Immigration)	1.281*** (0.167)
log(Scolarisation)	2.122*** (0.493)
Constant	-14.253*** (2.022)
Observations	96
R2	0.713
Adjusted R2	0.704
Residual Std. Error	0.725 (df = 92)
F Statistic	76.303*** (df = 3; 92)
Note: *p<0.1; **p<0.05; ***p<0.01	

Qualité d'ajustement : Nous avons un R2 ajusté de 0.704, c'est à dire que 70,4% de la variance de la variable Delit est expliqué par le modèle, autrement dit par les variables log(Population), log(Immigration) et log(Scolarisation). Le modèle est donc relativement en adéquation avec la variable Y.

Modèle globalement satisfaisant : On suppose l'hypothèse nulle H0 tel que: "Tous les coefficients sont nuls sauf la constante".

La p-value est strictement significative au regard des 3 étoiles, donc bien inférieur au seuil de 5% alors le modèle est globalement satisfaisant.

Significativité des variables : On test individuellement sur chaque variable explicative la significativité avec comme hypothèse nulle H0 : "coefficient de la variable est égale à 0".

Les variables log(Population), log(Immigration) et log(Scolarisation) ont une p-value inférieure au seuil de 5% donc ces variables ont un effet significatif sur le phénomène qu'on explique Y (Delit).

Interprétation des variables :

- Si la variable log(Population) augmente de 1% alors le taux de delit augmente de (0.7457/100) %. (Soit 0.007457%).
- Si la variable log(Immigration) augmente de 1% alors le taux de delit augmente de (1.2902/100) %. (Soit 0.012902%).
- Si la variable log(Scolarisation) augmente de 1% alors le taux de delit augmente de (2.2172/100) %. (Soit 0.022172%).

Malgré que les qualités (d'ajustement R2 ajusté) soient assez proches, nous conservons le modèle log-log qui a le R2 ajusté le plus élevé et le critère d'information d'Akaike le plus faible. Le modèle log-log est donc le meilleur modèle.

2.2 Analyse des résidus

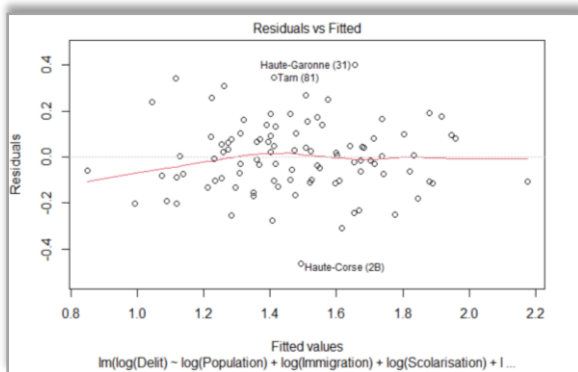


Figure n°11 – Residuals vs Fitted

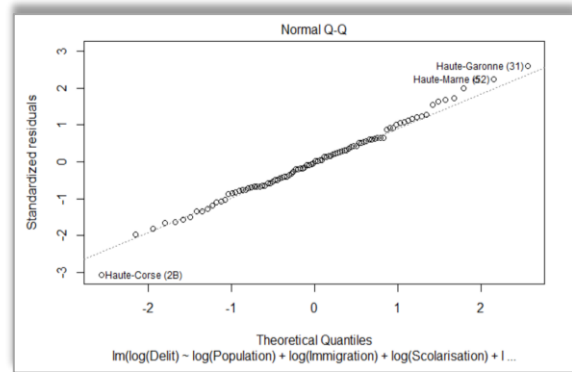


Figure n°12 – Normal Q-Q

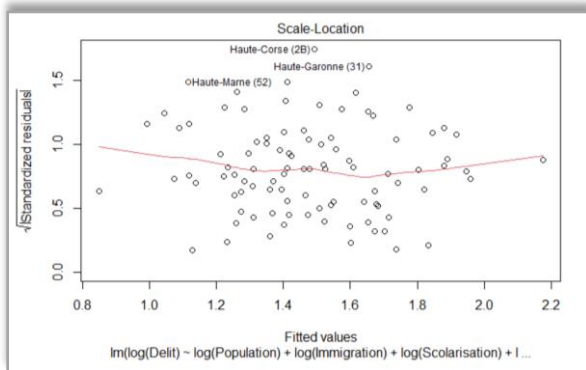


Figure n°13 – Scale Location

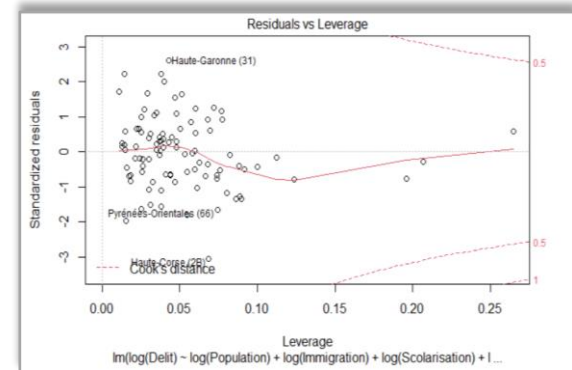


Figure n°14 – Residuals vs Leverage

- La figure n°11 (Residuals vs Fitted) nous indique que la linéarité est bonne car la ligne rouge est proche de la droite d'abscisse égale à 0. La relation linéaire est bien expliquée par le modèle. Plus l'on se déplace vers la droite de l'axe des abscisses, l'hétéroscédasticité se réduit et laisse place à une homogénéité presque parfaite des résidus. On note néanmoins, que les départements de la Haute-Corse (2B), Haute-Garonne (31) et du Tarn (81) ont des valeurs résiduelles importantes et peuvent être considéré comme individu atypique.
- La figure n°12 (Normal Q-Q) permet de mettre en relation la distribution de notre échantillon avec une distribution correspondant aux mêmes points obéissant à une distribution normale. Le comportement des points de notre échantillon nous indique une symétrie car les points situés au centre de la distribution sont très proche de la droite théorique. Nous retiendrons que la majorité des points sont bien alignés sur la droite. La Haute-Corse (2B) et la Haute-Garonne (31) se démarque une nouvelle fois.
- La figure n°13 (Scale-Location) montre que les résidus apparaissent répartis de manière aléatoire. La ligne est « horizontale ».
- La figure n°14 (Residuals vs Leverage) met en évidence l'effet levier de chaque individu sur l'axe des abscisses et son résidu standardisé sur l'axe des ordonnées. Les lignes rouges délimitent les distances de Cook.

Une observation hors des limites indiquerait que l'observation est influente et sans sa présence les coefficients du modèle changeraient considérablement. On peut noter que le département Haute-Garonne est proche de la frontière de la distance de Cook sans l'avoir franchie. En somme, notre modèle de régression ne compte pas de points influents.

Au final, nous sommes proches d'une homogénéité parmi nos observations. On ne relève pas de problème d'hétéroscédasticité. Malgré que La Haute-Corse (2B) et la Haute-Garonne (31) semblent se détacher des autres départements, aucun des deux n'est au-delà des distances de Cook. Dans cette étude nous cherchons à avoir une représentation globale des départements métropolitains raisons pour laquelle aucune observation ne se verrait exclue dans le cas où cette dernière ne comporterait pas d'erreur.

2.3 Tests

2.3.1 Test de spécification (Test de Ramsey)

L'idée du test de Ramsey est de savoir si le modèle a pris en compte toutes les variables pertinentes afin d'expliquer la variable endogène. Le test réside en la significativité d'une variable fictive ajoutée dans le modèle. Si elle n'est pas significative alors on considère notre modèle comme complet. Dans le cas contraire, des variables pouvant influencer l'endogène seront à introduire.

```
RESET test
data: modele
RESET = 0.53037, df1 = 2, df2 = 90, p-value = 0.5902
```

Le test de spécification a pour hypothèse H_0 : modèle bien spécifié. On a ici une $p\text{-value}=0.5902>0.05$. On accepte donc l'hypothèse H_0 . Le modèle est bien spécifié.

2.3.2 Test d'autocorrélation

A l'aide de la fonction *dwtest*, on teste H_0 : "Il n'y a pas d'autocorrélation" contre H_1 : "Il y a de l'autocorrélation".

```
Durbin-Watson test
data: eq_niveau_log2
DW = 1.6326, p-value = 0.03109
alternative hypothesis: true autocorrelation is greater than 0
```

La $p\text{-value}$ est égale à 0.03109 (<5%) donc on rejette H_0 au niveau 5%. Il y a présence d'autocorrélation. Afin d'y pallier nous procéderons à la méthode de Cochrane Orcutt.

Méthode de Cochrane Orcutt

La procédure de Cochrane Orcutt est utilisée en économie pour ajuster un modèle linéaire pour la corrélation sérielle dans le terme d'erreur. (Procédure R `cochrane.orcutt()`)

Ci-dessous, les résultats après application de la fonction :

```
Durbin-Watson test  
data: orc  
DW = 1.9558, p-value = 0.449  
alternative hypothesis: true autocorrelation is greater than 0
```

Non rejet de H_0 à l'aide du test de Durbin Watson. Par conséquent l'autocorrélation a bien été supprimée sur le modèle corrigé. Si l'on compare le modèle initial et le modèle corrigé, on devrait remarquer que les coefficients, la valeur des statistiques et la qualité d'ajustement ont changé.

2.3.3 Test Hétéroscédasticité (White/Goldfeld Quandt)

Au moyen des tests statistiques de White et de GoldFeld Quandt nous vérifions s'il y a des problèmes d'hétéroscédasticité.

Test de White

On teste H_0 :" Il n'y pas de problème d'hétéroscédasticité" contre H_1 :" Il y a un problème d'hétéroscédasticité". A l'aide de la fonction `bptest` du logiciel Rstudio, nous pourrions estimer les résidus MCO (Moindres Carrés Ordinaire) au carré en fonction des variables explicatives, de leur carré et de leurs produits croisés.

Les résultats obtenus sont présentés ci-dessous :

```
studentized Breusch-Pagan test  
data: eq_niveau_log2  
BP = 10.664, df = 7, p-value = 0.154
```

La p-value est égale à 0.154, donc supérieure à 5%, alors on ne rejette pas H_0 . Au niveau 5%, il n'y a pas de problème d'hétéroscédasticité. La variance des erreurs du modèle est donc constante, conduisant à l'homoscédasticité.

Test de Goldfeld Quandt

Nous supposons que la variance est fonction de la variable $\log(\text{Population})$. On applique le test de Goldfeld – Quandt à notre modèle. Ce test consiste à ordonner les observations, puis à les séparer en deux échantillons et par la suite retirer un pourcentage des observations centrales. Dans notre cas, ce pourcentage est fixé à 1/6 (environ 16%). A l'aide de la commande `gqtest`, on teste H_0 :" Il n'y pas de problème d'hétéroscédasticité" contre H_1 :" Il y a pas de problème d'hétéroscédasticité".

Les résultats obtenus sont présentés ci-dessous :

```
Goldfeld-Quandt test
data:  eq_niveau_log2
GQ = 0.57892, df1 = 37, df2 = 37, p-value = 0.9496
alternative hypothesis: variance increases from segment 1 to 2
```

On constate que la p-value (égale à 0.9496) est supérieure à 5%, donc on ne rejette pas H_0 . Au niveau 5%, il n'y pas de problème d'hétéroscédasticité.

Ces tests confirment nos précédents dire concernant la non hétéroscédasticité dans le modèle observé lors de l'analyse des résidus.

2.3.4 Changement de structure

Variable Politique

On souhaite tester l'hypothèse H_0 = "Il n'y a pas de différence structurelle entre les départements de Droite et ceux de Gauche" contre l'hypothèse H_1 = "Il y a une différence structurelle entre les départements de Droite et ceux de Gauche". On effectue le test de Chow afin d'y répondre. Ci-dessous le résultat de sa p-value :

```
pvalue = pf(FChow,ddl_n,ddl_d,lower.tail=FALSE)
pvalue

[1] 0.05876403
```

On remarque que la p-value est égale à 0.05876 (>0.05), donc on ne rejette pas H_0 . Au seuil de 5%, il n'y a pas de différence structurelle entre les départements de droite et ceux de gauche.

Variable : ile_de_France

Afin de d'estimer un potentiel changement structure entre les départements de la région parisienne et le reste des départements métropolitain, nous avons procédé par effet croisé au vu des effectifs respectifs des deux échantillons. 8 (île-de-France) contre 88 (hors île-de-France).

Ci-dessous les résultats obtenus :

Nous avons une p-value de 0.5106 (>0.05), donc nous ne rejetons pas l'hypothèse nulle. On conclut qu'il n'y pas de changement de structure entre les départements d'Ile-de-France et les départements de province.

```
Analysis of Variance Table

Model 1: log(Delict) ~ log(Population) + log(Immigration) + log(Scolarisation) +
(log(Population) + log(Immigration) + log(Scolarisation))^2 +
(log(Population) + log(Immigration) + log(Scolarisation))^3
Model 2: log(Delict) ~ log(Population) + log(Immigration) + log(Scolarisation) +
(log(Population) + log(Immigration) + log(Scolarisation))^2 +
(log(Population) + log(Immigration) + log(Scolarisation))^3 +
idf_Pop + idf_Scolar + idf_Immigra + (idf_Pop + idf_Scolar +
idf_Immigra)^2 + (idf_Pop + idf_Scolar + idf_Immigra)^3
Res.Df    RSS Df Sum of Sq    F Pr(>F)
1      88 2.1022
2      81 1.9505  7    0.15174 0.9002 0.5106
```

Conclusion

Notre étude a débuté avec une dizaine de variables explicatives afin d'illustrer notre problématique. Dans le processus de réponse à la problématique, nous avons tout d'abord analysé les distributions des variables exogènes une à une ou deux par deux au moyen essentiellement de boxplots. Variabilité et points atypiques ont pu être observés et identifiés.

Progressivement, le nombre de variables exogènes s'est vu diminuer dû à la non significativité de certaines variables avec l'endogène. Matrice de corrélation et algorithme de sélection de modèle nous ont permis de choisir l'ensemble des variables formant le modèle le plus adéquat.

Dans un souci d'amélioration du modèle en vue d'obtenir une estimation future meilleur, nous avons appliqué plusieurs transformations logarithmiques sur nos variables dans l'équation du modèle. C'est à l'aide de visualisation par des nuages de points et de mesures statistiques telles que le coefficient de détermination ajusté (R^2 ajusté) et le critère d'information d'Akaike que le modèle log-log nous a paru être le meilleur.

Par la suite, il était nécessaire de s'assurer de la validation des différentes hypothèses des MCO (homoscédasticité, non autocorrélation et espérance des aléas nulle) à travers de tests statistiques. Test de Ramsey, de White et de GoldFeld-Quandt ont respectivement confirmé la spécification du modèle ainsi que sa non hétéroscédasticité. En revanche, le test de Durbin Watson a révélé des problèmes d'autocorrélation qui ont été corrigé par la procédure de Cochrane-Orcutt.

En se servant du test de Chow et de son test équivalent nous avons étudié les potentiels changements structurels entre les départements. L'orientation politique du département et l'appartenance d'un département à la région île de France étaient les deux variables indicatrices. En fin de compte aucune des deux variables indicatrices n'apportent de changement de structure.

Quand est-il des déterminants de l'insécurité en France ?

Certaines variables n'ont pas été retenues pour le modèle prédictif final, mais peuvent toutefois apporter un éclaircissement en réponse à la problématique posée. L'importante corrélation négative entre le taux de délit et la part des 60-74 ans peut laisser penser qu'un département dont la part des 60-74ans est importante a une probabilité plus faible d'observer des cas de délit en 2018.

Si l'on se réfère à deux groupes constitués respectivement de 3 variables, c'est-à-dire le premier groupe (Population, Immigration et Urbanisation) et le second (taux de pauvreté, taux de chômage et la scolarisation), on observe une forte corrélation entre chacune des 3 variables. Les variables du premier groupe mettraient en évidence directement ou indirectement l'accroissement du nombre d'habitants et celles du second groupe les inégalités en termes de niveau d'éducation et de vitalité économique. Ces deux classes semblent être les principales sources du taux de délit en 2018.

Tandis que la population française continuerait de croître jusqu'à l'horizon 2044 en atteignant un pic démographique de 69,3M d'habitants (contre 66,9M d'habitants en 2018) et que les aires urbaines compteront de plus en plus d'habitants selon les projections de l'INSEE, devons-nous anticiper une augmentation considérable et continue des actes d'insécurité sur le territoire français ?

Néanmoins la corrélation n'implique pas forcément la causalité, c'est en là que réside la limite de l'étude.

Bibliographie

- [Chapitre 2 Faire des cartes avec R | Tutoriel : visualisation avec R \(lrouviere.github.io\)](#)
- [Délinquance : les cambriolages et les violences, physiques et sexuelles, en forte hausse en France \(lemonde.fr\)](#)
- [Violence, agressions, cambriolages... Une étude dresse le portrait de la délinquance en 2018 \(lefigaro.fr\)](#)
- [Rapport d'enquête « Cadre de vie et sécurité » 2018 / L'enquête Cadre de vie et sécurité \(CVS\) / Interstats - Ministère de l'Intérieur \(interieur.gouv.fr\)](#)
- [Les chiffres du cambriolage en France en 2021 \(portes-et-serrures.fr\)](#)

Annexe

- AIC: Le critère d'information d'Akaike est une mesure de la qualité d'un modèle statistique proposée par Hirotugu Akaike en 1973.
- Autocorrélation : Corrélation existant entre les valeurs successives d'une grande variable dans le temps.
- BIC : Statistique de mesure de la qualité d'ajustement d'un modèle tenant compte du nombre de paramètres et du nombre de données.
- Boxplot : Graphique représentant la distribution d'une série statistique avec ses quantiles.
- Colinéarité : La colinéarité fait référence la corrélation statistiquement significative existant entre deux variables.
- Effet levier : Fait référence à la mesure dans laquelle les coefficients du modèle de régression changeraient si une observation particulière était supprimée de l'ensemble de données.
- Hétéroscédasticité : Variance des résidus des variables examinées différentes.
- Homoscédasticité : Variance des erreurs stochastiques de la régression similaire pour chaque observation i .
- Nuage de point : Représentation graphique dans un repère du plan d'une série statistique de deux variables X et Y . Chaque individu i est représenté par un point dont les coordonnées sont les valeurs respectives des variables X et Y prises par l'individu i .
- Test de Chow : Détermine si les coefficients de deux séries linéaires sont égaux.
- Test de Durbin-Watson : Teste l'autocorrélation des résidus dans un modèle de régression linéaire
- Test de Goldfeld Quandt : Teste si les perturbations sont hétéroscédastiques ou homoscédastiques.
- Test de spécification: Déterminer si le modèle théorique est parfaitement identique pour deux zones ou s'il existe des spécificités propres à chaque zone.
- Test de White : Teste si la variance des erreurs d'un modèle de régression est constante.