Rapport d'étude sur les disparités de développement économique entre les départements français

Analyse des facteurs d'attractivité

Vanechop Justine
Moaye Johanne Astrid Eliaka

SOMMAIRE

Table des tableaux et figures

Introduction

Le développement économique des départements français est influencé par plusieurs facteurs socio-économiques. La création d'entreprises, en particulier, est un indicateur clé pour mesurer l'attractivité d'un territoire. Cependant, des données récentes de l'INSEE révèlent un ralentissement inquiétant, avec moins de 90 000 créations d'entreprises en France en septembre 2024, un seuil inédit depuis plus d'un an selon le journal *Les Echos*. Ce constat souligne la nécessité de mieux comprendre les déterminants locaux de cette dynamique et leurs interactions avec des facteurs structurels.

Les disparités entre départements sont également documentées dans des travaux antérieurs, notamment dans l'étude de Guesnier (2008) intitulée "Dynamique des territoires et création d'entreprises : une analyse des départements français". Ces recherches montrent que des variables telles que la densité de population, le niveau de formation, ou encore les revenus peuvent expliquer l'attractivité des territoires.

Dans cette perspective, ce rapport explore **les disparités de développement économique entre les départements français à travers une analyse des facteurs d'attractivité.** L'objectif est de fournir des recommandations concrètes pour éclairer les décideurs publics sur les politiques à prioriser afin de réduire les inégalités territoriales et promouvoir l'innovation.

Pour répondre à cette problématique, une méthode économétrique basée sur la **régression linéaire multiple** (MCO) a été choisie. Cette approche est particulièrement adaptée à l'analyse des données socio-économiques, car elle permet de modéliser la relation entre une variable dépendante, ici la proportion de création d'entreprises, et un ensemble de variables explicatives.

I. Présentation et justification des données - Création de la base de données

La base de données et traitement

Les variables ont été sélectionnées pour représenter des dimensions variées de l'attractivité économique : démographie, qualification de la main-d'œuvre, richesse locale, santé du marché du travail, et présence de pôles économiques structurants.

La variable principale, la **proportion de création d'entreprises en 2022 (PCENT)**, correspond à la part des nouvelles entreprises créées dans un département par rapport au nombre total d'entreprises existantes déjà en 2021. Elle est l'indicateur direct du dynamisme entrepreneurial et de l'attractivité économique d'un territoire. Elle reflète la capacité d'un département à offrir un environnement favorable à la création d'entreprises, influencé par des facteurs tels que les revenus, les infrastructures et le marché de l'emploi.

Le nombre total d'entreprises en 2021 (nbENT) mesure le nombre d'entreprises actives dans un département. Cette variable illustre la solidité du tissu économique local. Un département comptant de nombreuses entreprises actives bénéficie d'un écosystème favorable à l'entrepreneuriat, grâce à des opportunités de collaboration et à une dynamique économique positive. Cette variable est donc attendue comme ayant un effet positif sur la création d'entreprises.

La population municipale totale en 2021 (POP) est une mesure de la taille du marché local. Elle reflète à la fois la disponibilité de main-d'œuvre et le potentiel de consommation. Les départements densément peuplés offrent davantage d'opportunités économiques, bien que des effets négatifs puissent apparaître dans les zones surpeuplées, en raison d'une concurrence accrue et de coûts plus élevés.

La proportion de diplômés d'un Bac+3 et plus en 2021 (DIPL) correspond au pourcentage d'habitants d'un département ayant obtenu un diplôme de niveau Bac+3 ou supérieur. Le niveau d'éducation est un facteur clé pour attirer les entreprises innovantes et à forte valeur ajoutée. Une proportion élevée de diplômés favorise l'innovation et l'entrepreneuriat, et devrait donc avoir un impact positif sur la proportion de création d'entreprises. Néanmoins, un département qui concentre beaucoup d'entreprises aura tendance à attirer les personnes qualifiées.

Le revenu médian annuel en 2021 (REV) traduit le niveau de vie et le pouvoir d'achat des habitants d'un département. Ce facteur est un indicateur de la solvabilité du marché local. Un revenu médian élevé est attendu comme un facteur d'attractivité économique, car il signale un marché dynamique et propice à l'entrepreneuriat.

Le nombre de grandes entreprises en 2021 (gndENT), défini comme le nombre d'entreprises employant plus de 250 salariés, témoigne de la structuration économique d'un département. Les grandes entreprises ont un effet d'entraînement sur l'économie locale, en créant des opportunités pour

les petites et moyennes entreprises, notamment dans la sous-traitance. Une forte présence de grandes entreprises est donc considérée comme un atout pour stimuler la création d'entreprises.

Le taux de chômage en 2021 (txCHOM) quant à lui reflète l'état du marché du travail. Un taux de chômage élevé peut être un frein à l'attractivité économique, car il traduit un environnement économique difficile. Cependant, dans certains cas, des politiques publiques ciblées (comme les subventions ou les exonérations fiscales) peuvent atténuer cet effet négatif.

La variable indicatrice METRO a été introduite pour identifier les départements disposant ou non d'une métropole. Cette variable binaire joue un rôle important dans l'analyse des disparités territoriales, car la présence d'une métropole peut influencer directement l'attractivité économique.Les métropoles concentrent des infrastructures, des services et des talents, ce qui en fait des pôles économiques dynamiques.

Code	Définition	Unité	Source	Traitement
pcENT	Proportion de création entreprises par rapport aux entreprises existantes	En pourcentage	INSEE	Division du nombre d'entreprises créées en 2022 par le nombre d'entreprises en 2021
nbENT	Nombre d'entreprise en 2021	En milliers	INEE	Division du nombre d'entreprises en 2021 par 1 000
POP	Population municipale	En milliers	INSEE	Division de la population par 1 000
DIPL	Part de diplômés d'un Bac+3 et plus	En pourcentage	INSEE	Somme de la part des diplômés d'un Bac+3/Bac+4 avec ceux qui ont un Bac+5 et plus.
REV	Revenu annuel médian	En milliers d'euros	INSEE	Division du revenu médian par 1000
gndENT	Nombre de grandes entreprises	En unité	Data.gouv	Division du nombre de grandes entreprises par 1000
txCHOM	Taux de chômage	En pourcentage	INSEE	
METRO	Présence d'une métropole	Indicatrice	Collectivite locales.gouv	1 : Département abritant un métropole 0 : Autres départements

Individus

Certaines variables sont exprimées en milliers, afin d'améliorer la lisibilité et faciliter l'interprétation. Cette transformation permet de mieux appréhender l'ordre de grandeur des chiffres et de rendre les résultats plus intuitifs.

Après avoir effectué les traitements spécifiques nécessaires, nous avons procédé au nettoyage des données. En identifiant les individus présentant des valeurs manquantes, nous avons constaté que celles-ci concernaient les départements d'Outre-mer. Notre étude concerne les départements français en métropole, afin de garantir la cohérence et la comparabilité des observations. Nous avons donc exclu les départements d'Outre-mer et de la Corse. À l'issue de cette étape, nos données sont prêtes pour l'analyse.

Variables

Concernant la sélection des variables, aucune variable n'a été éliminée pour l'instant. En effet, toutes les variables disposaient d'un nombre suffisant de données pour être incluses dans l'étude. Ce choix repose également sur leur pertinence théorique dans l'explication des disparités d'attractivité et de développement économique des départements français. En conservant l'ensemble des variables, nous assurons une analyse complète et évitons de biaiser les résultats en omettant des facteurs potentiellement significatifs.

Ainsi, notre échantillon final comprend **94 départements** métropolitains et **1 variable dépendante** et **7 variables explicatives**, dont une variable indicatrice.

Méthodologie

Par la suite, une analyse descriptive univariée et bivariée, une analyse de corrélation et des tests de robustesse seront réalisés. L'analyse univariée permet de décrire les caractéristiques individuelles des variables, tandis que l'analyse bivariée explore les relations entre la proportion de création d'entreprises et les variables explicatives. Un test de corrélation est effectué pour évaluer les relations linéaires entre les variables. Des tests de robustesse sont ensuite appliqués, incluant la vérification de l'homoscédasticité, la normalité des résidus, l'identification des points aberrants et de la présence d'un changement structurel dans le modèle.

Régression linéaire multiple

Le modèle étudié :

$$pcENT_i = \beta_0 + \beta_1 nbENT_i + \beta_2 POP_i + \beta_3 DIPL_i + \beta_4 REV_i + \beta_5 gndENT_i + \beta_6 txCHOM_i + \beta_7 METRO_i + \varepsilon_i$$

II. Analyse descriptive

2.1. Statistiques descriptive univariées

L'analyse descriptive univariée consiste à examiner chaque variable individuellement pour en décrire les caractéristiques essentielles. Cette analyse aide à identifier la distribution des variables, la présence de valeurs extrêmes et à détecter d'éventuelles anomalies dans les données. Elle est essentielle pour comprendre la structure des données avant d'effectuer des analyses plus complexes.

Ces statistiques soulignent les disparités marquées entre les départements français, qui sont au cœur de l'analyse économétrique.

	pcENT <dbl>	nbENT <dbl>	POP <dbl>	DIPL <dbl>	REV <dbl>	gndENT <dbl>	txCHOM <dbl>
min	12.40110	5.81400	76.5190	9.900000	19.020000	0.138000	4.300000
max	33.89703	455.11600	2611.2930	56.700000	29.730000	9.999000	12.300000
median	18.05859	35.84150	548.1920	16.450000	22.285000	0.810000	7.350000
mean	18.35400	54.25237	693.1661	17.768085	22.712340	1.165574	7.565957
std.dev	3.36438	59.13980	524.9164	6.856238	1.733914	1.455903	1.480958

Tableau 1 : Statistiques descriptives

Une première exploration des données, avec un tableau de statistiques descriptives résumant les caractéristiques des variables utilisées dans l'analyse. Ces statistiques comprennent les valeurs moyennes, médianes, écarts-types, minimums et maximums pour chacune des 6 variables qualitatives.

Voici plusieurs observations intéressantes à partir de ces données départementales françaises :

Disparités dans la création d'entreprises (pcENT) : Le taux varie de 12,4% à 33,9% et la moyenne est de 18,35%, ce qui suggère que certains départements sont des "outliers" avec des taux très élevés de création.

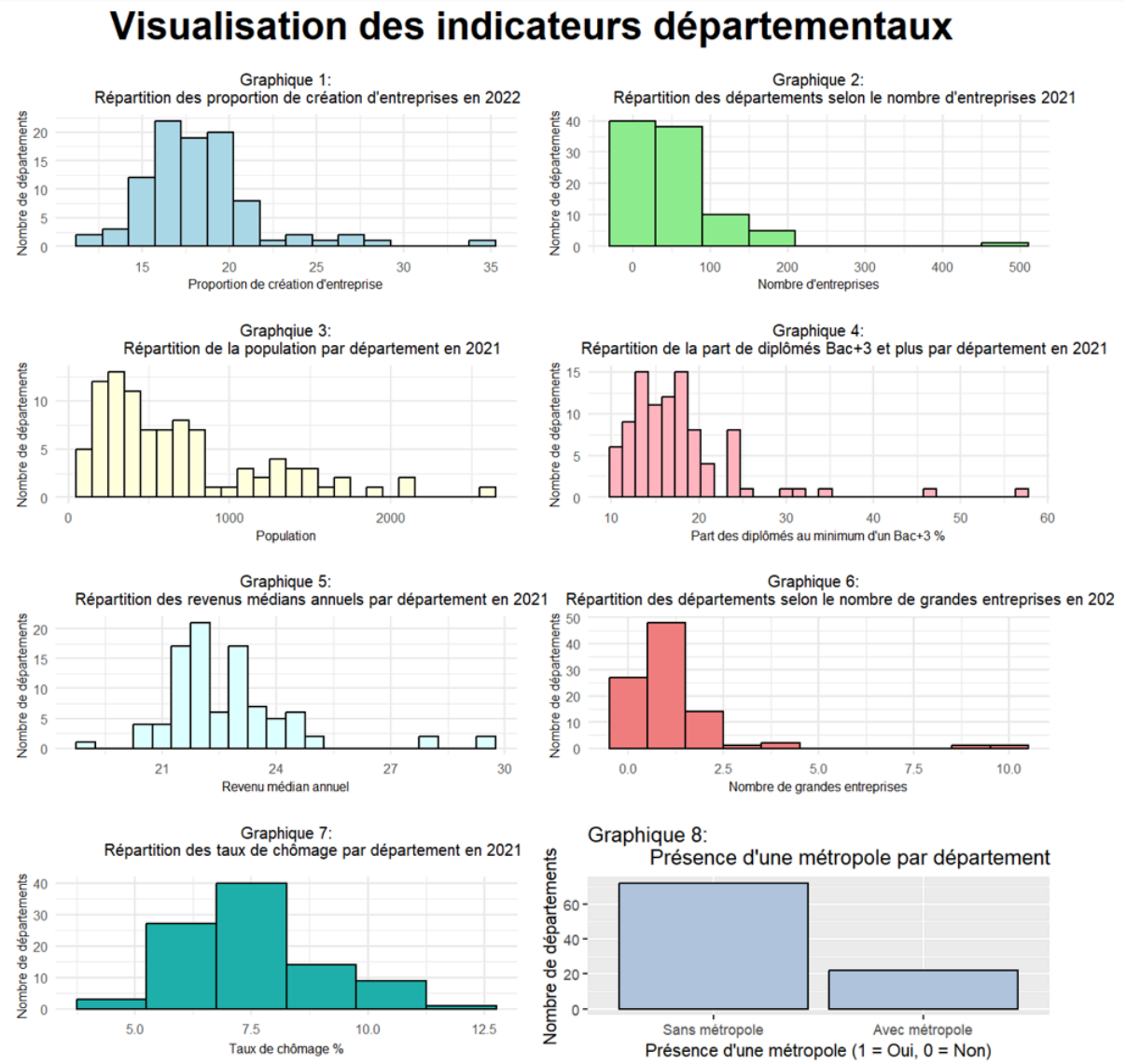
Concentration des entreprises (nbENT) : Une très grande disparité avec un minimum de 5,8k d'entreprises et un maximum de 455k. Et l'écart-type élevé (59,14) indique une forte concentration des entreprises dans certains départements, surement des zones urbaines

Population (POP) : Varie énormément la plus faible concentration est de 76,5k habitants contre le département ayant la plus grande densité qui est de 2611,29k habitants. L'écart-type très important (524,96) suggère une forte hétérogénéité démographique entre les départements.

Niveau d'éducation (DIPL) : Une moyenne de 17,76 % de diplômés (Bac+3/Bac+4) suggère que dans la plupart des départements, la population est relativement qualifiée. La part de diplômés d'un Bac+3 varie entre 9,9% et 56,7%, certains départements concentrent plus de personnes qualifiées que d'autres. Cela peut s'expliquer par la présence de campus universitaires.

Revenus (REV) et **Chômage (txCHOM)** : Les écarts-types sont relativement faibles (1,73 et 1,48 respectivement) suggèrent une distribution plus homogène des revenus entre départements, ainsi que le chômage qui indique que la majorité des départements ont des taux de chômage assez proches de la moyenne nationale que d'autres indicateurs.

Ces statistiques révèlent des inégalités territoriales importantes, particulièrement en termes de population et d'activité économique, mais aussi une certaine homogénéité sur d'autres aspects comme les revenus.



Graphique 1: Distribution des variables

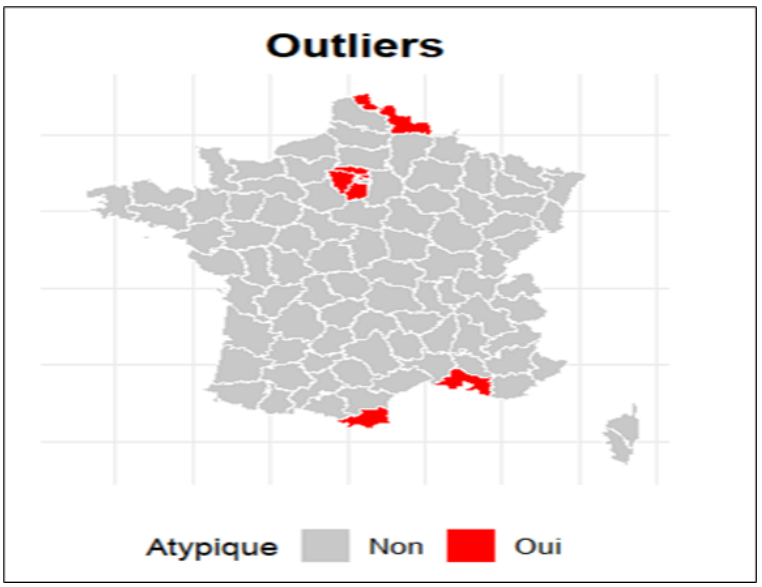
La visualisation de la répartition de chaque variable permet d'appuyer notre analyse du tableau des statistiques descriptives mais surtout d'identifier la présence d'individus atypiques, ainsi que la forme de la distribution.

La proportion de création d'entreprises (pcENT), le revenu médian annuel (REV) et le taux de chômage (txCHOM) suivent une distribution approximativement symétrique (ou approximativement normale). En effet, ces distributions sont centrées autour d'une valeur moyenne avec une répartition relativement équilibrée des deux côtés.

En revanche, le nombre d'entreprises (nbENT), la population (POP), la part des diplômés Bac+3 et plus (DIPL) et le nombre de grandes entreprises (gndENT) ont une distribution asymétrique à droite.

Ces variables ont donc une forte concentration de départements avec des valeurs faibles et quelques individus atypiques avec des valeurs très élevées. Les outliers[1] sont probablement des départements contenant des métropoles, ce qui est cohérent avec le Graphique 8 qui montre une minorité de départements avec métropole.

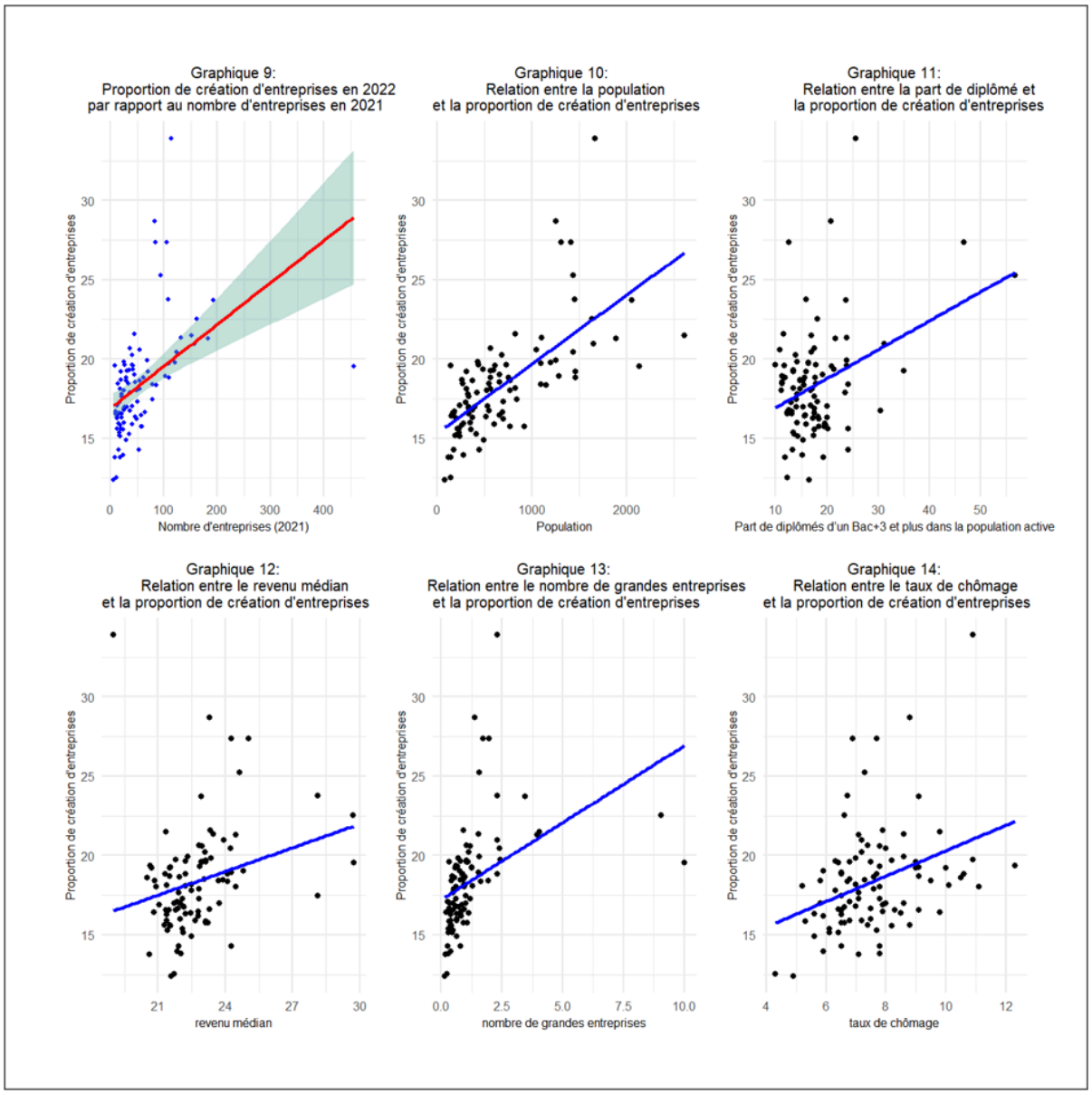
Toutes les variables, à l'exception du taux de chômage, présentent des individus atypiques, souvent concentrés sur des départements spécifiques (Paris, Hauts-de-Seine, Yvelines, etc). Ces départements jouent un rôle clé dans les disparités observées et justifient l'utilisation de transformations logarithmiques pour réduire l'influence de ces valeurs extrêmes sur les analyses. Cependant ces mêmes variables semblent asymétriques vers la droite, ce qui pourrait influencer l'analyse par la suite. Des tests de normalité seront également nécessaires pour confirmer ces ajustements.



Graphique 2: Carte de la France montrant les départements atypiques

[1] Valeurs aberrantes

2.2. Analyse descriptive bivariée



Graphique 3: Régressions linéaires simples illustrant les relations entre la proportion de création d'entreprises et les variables explicatives

Les régressions linéaires simples (Graphique 3) permettent d'explorer les relations entre la proportion de création d'entreprises et différentes variables explicatives, révélant des tendances cohérentes avec les attentes théoriques.

On observe une relation positive entre la proportion de création d'entreprises et toutes les autres variables explicatives. Cependant, on peut constater des effets différents. Une pente relativement faible (Graphiques 12 et 14) signifie que même de grandes variations du revenu annuel médian et du taux de chômage n'ont qu'un faible impact sur la création d'entreprise. En revanche, une pente relativement forte (Graphiques 9, 10 et 13) indique qu'une augmentation du nombre d'entreprises et de grandes entreprises et de la population dans un département entraîne une forte augmentation de la

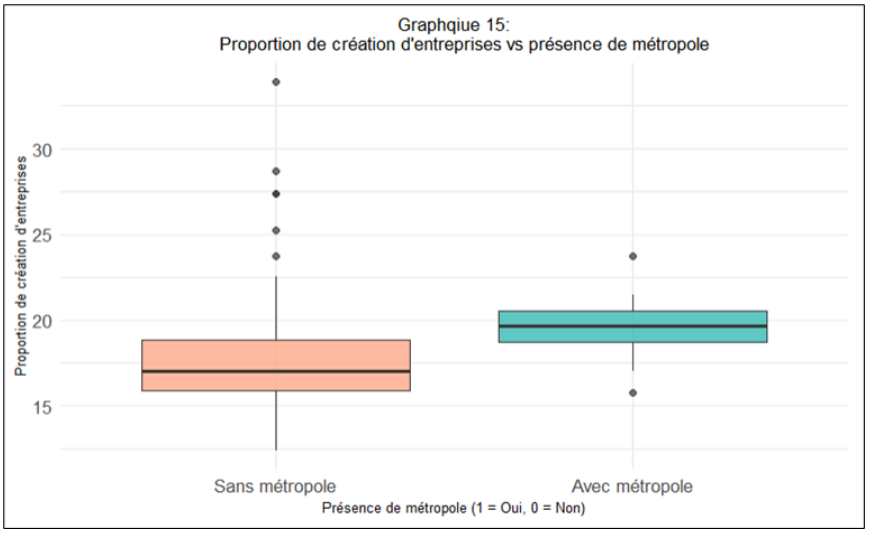
création d'entreprises. Ces variables peuvent être considérées comme déterminantes si elles sont statistiquement significatives. Concernant le revenu médian annuel (Graphique 11), la pente est modérée, c'est-à-dire que la variable exerce une influence mais pas dominante.

Les nuages de points (Graphiques 9, 10, 11 et 13) montrent une forte concentration des individus au niveau des faibles valeurs, avec quelques valeurs extrêmes. En effet, une grande partie de la population est regroupée dans des zones urbaines qui centralisent l'activité économique (comme l'Ile-de-France, les grandes métropoles régionales telles que Lyon, Marseille ou Toulouse, etc.), ces zones correspondent à une minorité de départements. En revanche, les départements ruraux ont peu d'entreprises et une population plus dispersée, ce qui explique la forte concentration des points aux faibles valeurs sur les graphiques. Quant à la relation entre le nombre d'entreprises, la population, la part de diplômés, le nombre de grandes entreprises et la proportion de création d'entreprises, elle a une légère tendance exponentielle. Une transformation linéaire est recommandée pour réduire les écarts entre les individus en compressant les valeurs extrêmes et en amplifiant les petites valeurs. Une linéarisation de la relation permet aux variables de mieux s'adapter au modèle linéaire.

Cependant, la part de diplômés, contrairement aux autres variables, à une plage de dispersion beaucoup plus faible car la variable est exprimée en pourcentage. Une transformation en logarithme n'est donc pas nécessaire dans ce cas.

Concernant le revenu médian annuel et le taux de chômage (Graphiques 12 et 14), les points sont très éloignés de la tendance, mettant en évidence une grande dispersion. Néanmoins, les points semblent symétriquement distribués au-dessus et en dessous de la droite de tendance, c'est-à-dire que les points sont répartis assez uniformément. Aucune transformation en logarithme n'est utile pour ces deux variables.

On constate que la répartition des points sur les graphiques 9, 10 et 13 est très similaire. Il est important de vérifier s'il n'y a pas de colinéarité entre le nombre d'entreprises, la population et le nombre de grandes entreprises.



Graphique 4: Boîte à moustache de la proportion de création d'entreprise en 2022 selon la présence de métropole dans les départements

Les départements avec une métropole ont une médiane plus élevée, suggérant que la présence d'une métropole est associée à une proportion plus importante de créations d'entreprises. Cependant, la dispersion est plus grande pour les départements sans métropole, avec des valeurs extrêmes plus marquées, ce qui peut indiquer des outliers dans la relation entre la présence d'une métropole et la proportion de création d'entreprises. En effet, la présence d'une métropole semble être un facteur associé à une proportion plus élevée et moins variable de créations d'entreprises.

III. Etude de la corrélation

L'étude de la corrélation de la régression linéaire permet d'identifier la multicolinéarité dans un modèle. La multicolinéarité concerne la corrélation entre plusieurs variables indépendantes. Lorsqu'elles sont fortement corrélées, les erreurs standards peuvent être plus élevées, ce qui diminue la signification statistique des variables et peut réduire la précision des prédictions. Identifier ces colinéarités aide à simplifier le modèle, en supprimant ou en combinant les variables explicatives corrélées. Il est important de s'assurer que chaque variable apporte une information indépendamment des autres à la régression linéaire.

3.1. Analyse de la régression linéaire multiple et des régressions linéaires simples

On rappelle le modèle linéaire :

$$pcENT_i = \beta_0 + \beta_1 nbENT_i + \beta_2 POP_i + \beta_3 DIPL_i + \beta_4 REV_i + \beta_5 gndENT_i + \beta_6 txCHOM_i + \beta_7 METRO_i + \varepsilon_i$$

Avant d'étudier la corrélation, on s'intéresse à la validité globale du modèle avec un test de Fisher, ainsi qu'à la valeur du R2-ajusté, et à la significativité des variables explicatives avec un test de Student pour identifier des potentiels problèmes de colinéarité.

Régression linéaire multiple

Mise en place des hypothèses du test de Fisher :

$H0$: *les variables explicatives n'apportent aucune information pour expliquer la variabilité de la variable dépendante*

$H1$: *Au moins un coefficient des variables explicatives n'est pas nul. Le modèle a un pouvoir explicatif global*

Mise en place des hypothèses du test de Student :

$H0$: *La variable explicative Xi, n'a aucun effet significatif sur la variable dépendante*

$H1$: *La variable explicative Xi a un effet significatif sur la variable dépendante*

	Dependent variable:
	pcENT
nbENT	-0.021**
	(0.010)
POP	0.006***
	(0.001)
REV	0.385*
	(0.208)
DIPL	0.075**
	(0.035)
gndENT	-0.078
	(0.360)
txCHOM	0.718***
	(0.188)
METRO	-1.289*
	(0.693)
Constant	0.305
	(5.464)
Observations	94
R^2	0.621
Adjusted R^2	0.590
Residual Std. Error	2.155 (df = 86)
F Statistic	20.097*** (df = 7; 86)
Note:	*p**p***p<0.01

Tableau 2 : Résumé du modèle linéaire multiple

Résultats du test de Fisher :
La F-statistique est de 20,097 avec une p-value de 1,008e-15, ce qui est largement inférieur au seuil de 1%. On rejette l'hypothèse nulle, donc le modèle est globalement significatif. Cela signifie qu'au moins une des variables explicatives a une influence significative sur la variable dépendante.

Le R2 ajusté est de 0,590. Cela nous indique que 59% de la variabilité de la variable dépendante, soit la proportion de création d'entreprises, est expliquée par les variables explicatives. Le modèle a un assez bon pour explicatif global.

Résultats du test de Student
Au seuil de 5%, les variables nbENT, POP, DIPL et txCHOM ont une p-value inférieur à 0,05. On rejette les hypothèses nulles, ainsi le nombre d'entreprises, la population, la part de diplômés et le taux de chômage sont statistiquement significatifs dans ce modèle. Elles ont un effet sur la proportion de création d'entreprises.

Cependant, les autres variables telles que REV, gndENT, METRO et la constante présentent des p-values supérieures à 0,05. On a échoué à rejeter les hypothèses nulles, les effets du revenu annuel médian, du nombre de grandes entreprises, de la présence de métropoles ne sont pas suffisamment forts pour être considérés comme significatifs.

Régressions linéaires simples

On s'intéresse à la régression linéaire simple pour comprendre la relation entre la variable dépendante et les variables explicatives individuellement.

Predictor <chr>	Coefficient <dbl>	Std_Error <dbl>	t_value <dbl>	P_value <dbl>
nbENT	0.02635992	0.0052559140	5.015288	2.564364e-06
POP	0.00434556	0.0004911835	8.847122	6.082230e-14
REV	0.49706157	0.1955438605	2.541944	1.269579e-02
DIPL	0.18229630	0.0474980470	3.837975	2.273734e-04
gndENT	0.96983145	0.2186786218	4.434962	2.545913e-05
txCHOM	0.80102204	0.2216355039	3.614141	4.912313e-04
METRO	1.71096360	0.8044830867	2.126786	3.611416e-02

Tableau 3 : Résumé des coefficients des régressions linéaires simples

Au seuil de 5%, toutes les variables explicatives ont une p-value inférieur à 0,05. On rejette les hypothèses nulles, Individuellement les variables sont toutes significatives statistiquement, c'est-à-dire qu'elles ont une influence sur la variable dépendante. Cependant, une fois qu'on les combine en un seul modèle, les variables perdent leur significativité statistiquement.

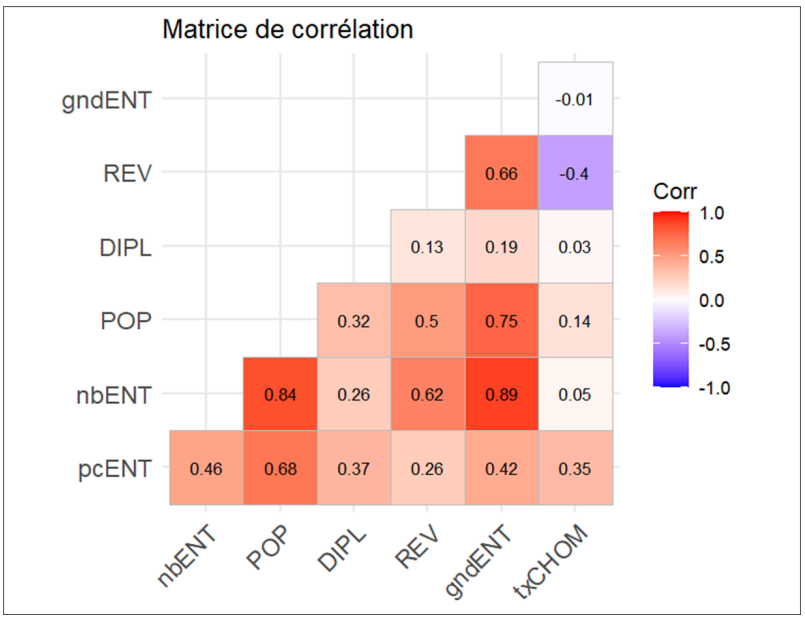
Cette différence peut provenir d'un problème de **multicolinéarité**. Dans ce cas, le modèle a du mal à différencier la contribution respective de chaque variable sur la variable dépendante.

3.2. Analyse de la corrélation

Plusieurs méthodes peuvent être utilisées pour détecter la multicolinéarité dans une régression linéaire. Nous allons appliquer la matrice de corrélation et le Facteur d'Inflation de la Variance (VIF), afin d'identifier et d'approfondir l'analyse de la corrélation.

Matrice de corrélation

Le coefficient de corrélation mesure l'intensité et la direction de la relation linéaire entre deux variables et sa valeur varie entre -1 et +1. Une valeur absolue proche de 1, indique une forte corrélation positive ou négative entre deux variables. Une valeur proche de 0 signifie qu'il y a une très faible, voire aucune corrélation.



Graphique 5 : Matrice de corrélation des variables

Selon la matrice de corrélation, les coefficients du nombre d'entreprises (nbENT), de la population (POP) et du nombre de grandes entreprises (gndENT) sont supérieurs à 0,8. Cela suggère que, à mesure que la population augmente, le nombre d'entreprises augmente également. En effet, il est logique que dans les zones plus peuplées, la concentration d'entreprises soit plus élevée. De même qu'un département avec plus d'entreprises pourrait aussi en avoir plus de grandes entreprises.

On peut soupçonner un problème de multicolinéarité du nombre d'entreprises avec la population et le nombre de grandes entreprises. Les coefficients estimés pour ces variables peuvent être moins fiables et significatifs, ce qui pourrait rendre le modèle moins performant et plus difficile à interpréter.

Facteur d'Inflation de la Variance (VIF)

Le Variance Inflation Factor est un indicateur clé pour détecter la colinéarité dans un modèle de régression. Un VIF supérieur à 1 indique un certain niveau de colinéarité entre les variables explicatives. En général, un VIF entre 1 et 5 indique qu'il y a une faible colinéarité avec les autres variables du modèle. Cela ne relève pas de problème majeur de colinéarité pour ces variables, puisqu'elles n'introduisent pas d'erreurs standard importantes ou d'instabilité dans les estimations des coefficients du modèle. En revanche, un VIF entre 5 et 10, traduit d'une colinéarité modérée à élevée entre la variable et les autres régresseurs.

Variance Inflation Factors						
nbENT	POP	REV	DIPL	gndENT	txCHOM	METRO
7.603	4.229	2.606	1.178	5.499	1.545	1.741

Square Root of Variance Inflation Factors						
nbENT	POP	REV	DIPL	gndENT	txCHOM	METRO
2.757	2.057	1.614	1.085	2.345	1.243	1.319

Tableau 4 : Facteur d'Inflation de la Variance (VIF) et Racine carré du Facteur d'Inflation de la Variance

Le résultat du VIF confirme notre observation faite avec la matrice de corrélation. Un VIF de 7,9 de la variable nbENT, indique qu'il y a une colinéarité assez forte entre le nombre d'entreprises et les autres régresseurs. On observe aussi une colinéarité modérée entre gndENT (nombre de grandes entreprises) et les autres régresseurs (VIF de 5,4). Concernant les autres variables, le VIF est inférieur à 5 donc il y a une faible colinéarité mais n'indique pas de problème majeur dans le modèle.

La racine carrée de VIF indique de combien de fois l'écart-type est modifié par rapport à la situation dans laquelle la variable Xj serait corrélée aux autres régresseurs. Par exemple, l'écart-type du coefficient de la variable du nombre d'entreprises est 2.81 fois plus grand qu'il ne le serait si le nombre d'entreprises n'était pas corrélé aux autres variables.

3.3. Sélection du modèle

On constate une multicolinéarité suffisamment importante entre les variables nbENT, POP et gndENT dans le modèle estimé. On peut ainsi envisager une modification du modèle en supprimant certaines des variables corrélées.

Critère d'Information d'Akaike (AIC)

L'AIC est une mesure utilisée pour évaluer la qualité d'un modèle statistique en fonction de la fonction de vraisemblance et du nombre de paramètres. Il sert à comparer différents modèles et à sélectionner celui qui s'ajuste le mieux aux données. On cherche à obtenir l'AIC le plus faible.

L'opération consiste à partir du modèle complet, de retirer une variable avec l'AIC le plus élevé et de d'observer le retrait qui entraîne la plus forte baisse de l'AIC. Si le retrait d'une variable n'entraîne pas une diminution de l'AIC, on s'arrête. Sinon, on recommence le processus de retrait.

AIC modèle complet
AIC
151.982

AIC modèle complet	
variable	AIC
nbENT	156.24869614741
POP	191.469068994755
DIPL	156.798892488236
REV	155.647933378092
gndENT	152.033464609707
txCHOM	166.762622870986
METRO	155.694818972352

AIC modèle sans POP
AIC
191.469

AIC modèle sans POP	
variable	AIC
nbENT	191.770634097135
DIPL	202.303664353724
REV	195.771380694335
gndENT	191.645517705249
txCHOM	211.294008743419
METRO	191.774413480224

AIC modèle sans nbENT
AIC
156.249

AIC modèle sans nbENT	
variable	AIC
POP	191.770634097135
DIPL	160.051790985257
REV	158.753366763277
gndENT	160.387604392393
txCHOM	169.045537857172
METRO	162.229361378439

AIC modèle sans gndENT
AIC
152.033

AIC modèle sans gndENT	
variable	AIC
nbENT	160.387604392393
POP	191.645517705249
DIPL	157.041035443491
REV	155.733918316397
txCHOM	166.775381345138
METRO	155.719192269457

Le retrait d'une variable fait augmenter l'AIC, on conserve le modèle complet :

$pcENT_i = \beta 0 + \beta 1 * nbENT_i + \beta 2 * POP_i + \beta 3 * DIPL_i + \beta 4 * REV_i + \beta 5 * gndENT_i + \beta 6 * txCHOM_i + \beta 7 * METRO_i$

Tableau 5 : Résultats des AIC

Il est pertinent de conserver le modèle complet, car il minimise l'AIC. Bien que la multicolinéarité entre les variables nbENT, POP et gndENT puisse rendre les coefficients individuels plus difficiles à interpréter, cela n'impacte pas nécessairement la performance globale du modèle à expliquer la variable dépendante (pcENT). L'AIC privilégie un ajustement optimal de l'ensemble des variables. En effet, le nombre d'entreprises est une variable clé pour expliquer la création d'entreprises, malgré qu'elle soit corrélée à d'autres variables explicatives. La présence de toutes les variables permet d'exploiter pleinement l'information disponible, même si certaines sont redondantes.

Par la suite, le R2-ajusté et la signification individuelle des variables pourront être améliorés en effectuant des transformations en logarithme.

On retient le modèle complet :

$$pcENT_i = \beta_0 + \beta_1 nbENT_i + \beta_2 POP_i + \beta_3 DIPL_i + \beta_4 REV_i + \beta_5 gndENT_i + \beta_6 txCHOM_i + \beta_7 METRO_i + \varepsilon_i$$

IV. Estimation du modèle

4.1. Justification de la spécification

La spécification d'une régression linéaire consiste à déterminer comment les relations entre les variables sont-elles modélisées et appliquer les transformations nécessaires (par exemple en logarithme). Une bonne spécification est primordiale pour que le modèle puisse être valide et interprétable. L'analyse bivariée (section 2.2) a permis d'identifier les relations des variables et de justifier les transformations en logarithme. Ainsi, les variables nbENT, POP et gndENT sont transformées en logarithmes pour linéariser les relations.

On décide d'évaluer trois modèles : le modèle complet sans transformation, un modèle niveau-log avec la variable pcENT non transformée et un modèle log-log avec pcENT transformée. Comparer les modèles permet d'identifier lequel offre le meilleur ajustement aux données.

Modèles étudiés

Modèle 1: niveau-niveau

$$pcENT_i = \beta_0 + \beta_1 nbENT_i + \beta_2 POP_i + \beta_3 DIPL_i + \beta_4 REV_i + \beta_5 gndENT_i + \beta_6 txCHOM_i + \beta_7 METRO_i + \varepsilon_i$$

Modèle 2: niveau-log

$$pcENT_i = \beta_0 + \beta_1 log(nbENT_i) + \beta_2 log(POP_i) + \beta_3 DIPL_i + \beta_4 REV_i + \beta_5 log(gndENT_i) + \beta_6 txCHOM_i + \beta_7 METRO_i + \varepsilon_i$$

Modèle 3: log-log

$$log(pcENT_i) = \beta_0 + \beta_1 log(nbENT_i) + \beta_2 log(POP_i) + \beta_3 DIPL_i + \beta_4 REV_i + \beta_5 log(gndENT_i) + \beta_6 txCHOM_i + \beta_7 METRO_i + \varepsilon_i$$

Les résultats sont présenté dans le tableau suivant :

	Dependent variable:		
	pcENT		log(pcENT)
	(1)	(2)	(3)
nbENT	-0.021**		
	(0.010)		
POP	0.006***		
	(0.001)		
log(nbENT)		-3.573***	-0.216***
		(1.053)	(0.047)
log(POP)		4.998***	0.278***
		(1.148)	(0.052)
DIPL	0.075**	0.098***	0.004***
	(0.035)	(0.034)	(0.002)

REV	0.385*	0.257	0.021**
	(0.208)	(0.212)	(0.010)
gndENT	-0.078		
	(0.360)		
log(gndENT)		1.432*	0.074**
		(0.828)	(0.037)
txCHOM	0.718***	0.740***	0.044***
	(0.188)	(0.190)	(0.009)
METRO1	-1.289*	-1.018	-0.028
	(0.693)	(0.661)	(0.030)
Constant	0.305	-12.659	1.072***
	(5.464)	(8.369)	(0.377)
Observations	94	94	94
R^2	0.621	0.643	0.709
Adjusted R^2	0.590	0.614	0.686
Residual Std. Error (df = 86)	2.155	2.091	0.094
F Statistic (df = 7; 86)	20.097***	22.107***	29.970***
Note:			*p**p***p<0.01

Tableau 6 : Comparaison des résumé des modèles linéaires multiples

Les trois modèles linéaires sont globalement significatifs puisque la p-value du test de Fisher est inférieur à 0,01. Cependant, la part de variabilité de la création d'entreprise expliquée par le modèle est bien plus élevée avec le modèle 3 (R2-ajusté = 0.686). Un R2-ajusté assez élevé indique que le modèle capture bien la relation entre les variables explicatives et la variable dépendante.

Le résumé montre que toutes les variables du modèle 3, à l'exception de METRO, sont statistiquement significatives à un niveau de confiance de 95%. En comparaison, les modèles 1 et 2 présentent seulement 4 variables significatives au seuil de 5%.

On déduit que le modèle log-log a un meilleur ajustement aux données que les deux autres modèles et que la transformation en logarithme de la variable dépendante est nécessaire. On choisit donc le modèle 3

4.2. Etude de la robustesse du modèle

4.2.1. Etude de l'Hétéroscédasticité

L'hétéroscédasticité dans une régression linéaire peut entraîner des problèmes d'estimation des coefficients, ainsi que des tests statistiques invalides. En d'autres termes, un modèle qui souffre d'hétéroscédasticité signifie que la variance des résidus des variables sont différentes. Le test de White qui s'appuie du test de Breusch-Pagan permet de vérifier la présence d'hétéroscédasticité dans le

modèle ou non. L'intérêt du test de White est qu'il est plus général car il inclut aussi des relations qui sont non linéaires comparé au test de Breusch-Pagan. Dans le cas où l'hétéroscédasticité est linéaire, les deux tests donnent des résultats similaires.

Test de White

Mise en place des hypothèses du test de White :

$H0$: $Homoscédasticité\ des\ résidus$

$H1$: $Présence\ d'hétéroscédasticité$

Résultats

```
BP = 34.647, df = 11, p-value = 0.0008597
```

La p-value étant inférieure à 5 %, on rejette l'hypothèse nulle d'homogénéité de la variance des résidus au seuil de 5 %. Les aléas présentent donc de l'hétéroscédasticité, ce qui peut rendre les estimations des coefficients inefficaces et compromettre la validité des tests statistiques. Il est essentiel de corriger cette hétéroscédasticité afin d'assurer des résultats fiables et robustes.

Correction de l'hétéroscédasticité

On applique les corrections de White et de Newey-West qui ajustent les erreurs standard pour l'hétéroscédasticité sans tenir compte de la taille de l'échantillon, tandis que Newey-West corrige également pour l'autocorrélation. Dans notre cas, avec des données transversales, cette dernière est moins pertinente. Les coefficients restent inchangés, mais la variance est modifiée, affectant les tests de significativité. Cette méthode ne rend pas les paramètres biaisés, mais elle ajuste leur variance, ce qui peut rendre certaines variables moins ou plus significatives.

Les résultats de la correction :

Comparaison de l'estimation MCO			
	Variable dépendante :'proportion de création d'entreprise		
	MCO	MCOWhite	MCONewey
log(POP)	0.104***	0.104**	0.104**
	(0.038)	(0.043)	(0.044)
DIPL	0.003**	0.003**	0.003**
	(0.002)	(0.002)	(0.001)
REV	0.004	0.004	0.004
	(0.010)	(0.013)	(0.014)
log(gndENT)	0.052	0.052	0.052
	(0.041)	(0.041)	(0.036)
txCHOM	0.033***	0.033***	0.033***
	(0.009)	(0.009)	(0.008)
METRO	-0.059*	-0.059	-0.059
	(0.032)	(0.037)	(0.045)

Constant	1.874***	1.874***	1.874***
	(0.369)	(0.388)	(0.373)
Observations	94		
R^2	0.639		
Adjusted R^2	0.614		
Residual Std. Error	0.104 (df = 87)		
F Statistic	25.672*** (df = 6; 87)		
Note:			*p**p***p<0.01

Tableau 7: Comparaison de l'estimation MCO avec celles des corrections de White et de Newey-West

L'ajustement des variances à l'aide de la correction de White entraîne de légères augmentations et diminutions des erreurs standard.Ces ajustements reflètent la prise en compte de l'hétéroscédasticité, en augmentant ainsi la précision des tests de significativité notamment pour les variables log(nbENT), log(POP), DIPL et txCHOM. Au contraire, les variables comme REV, log(gndENT) et METRO perdent de leur significativité sur la variable dépendante.

Dans l'ensemble, ces corrections montrent que les estimations MCO standards sous-évaluent l'incertitude associée aux coefficients. En ajustant les erreurs standard, elles permettent des tests plus fiables, bien que certaines variables puissent devenir moins significatives en raison de l'augmentation des marges d'erreur pour certaines d'entre elles.

4.2.2. Test de normalité des résidus

Deuxièmement, il est essentiel de tester la normalité des résidus, puisqu'elle affecte la validité des tests statistiques comme ceux de Fisher et de Student. Le test de Shapiro-Wilk et le QQ-plot sont utilisés pour détecter la normalité des résidus. Si les résidus sont normalement distribués, cela renforce la validité du modèle.

Test de Shapiro-Wilk

Mise en place des hypothèses :

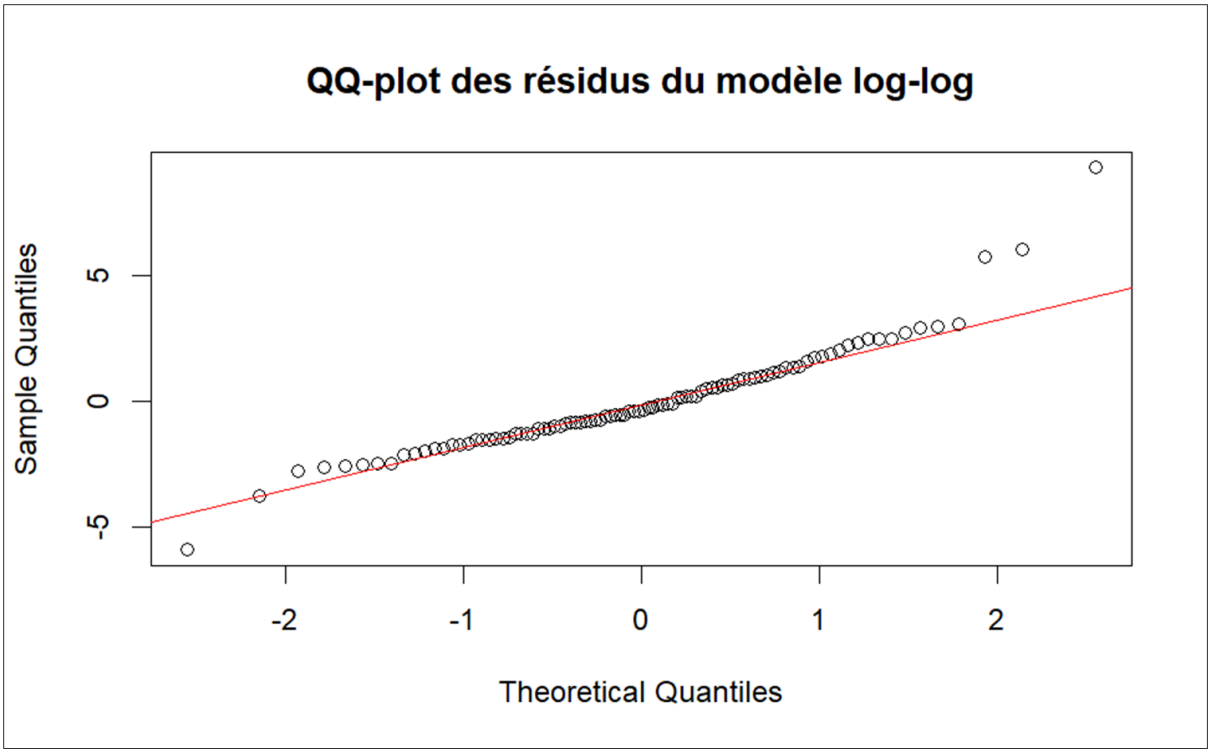
$H0$: $Normalité\ des\ résidus$

$H1$: $Anormalité\ des\ résidus$

Résultats

```
W = 0.93929, p-value = 0.0002806
```

La p-value est inférieur à 0,05. On rejette l'hypothèse nulle, on ne peut pas considérer que les résidus sont normalement distribués. Sachant que notre échantillon n'est pas assez grand pour que l'approximation des tests reste valide sans la normalité des résidus, On peut suggérer un problème de spécification du modèle ou alors la présence de valeurs aberrantes.



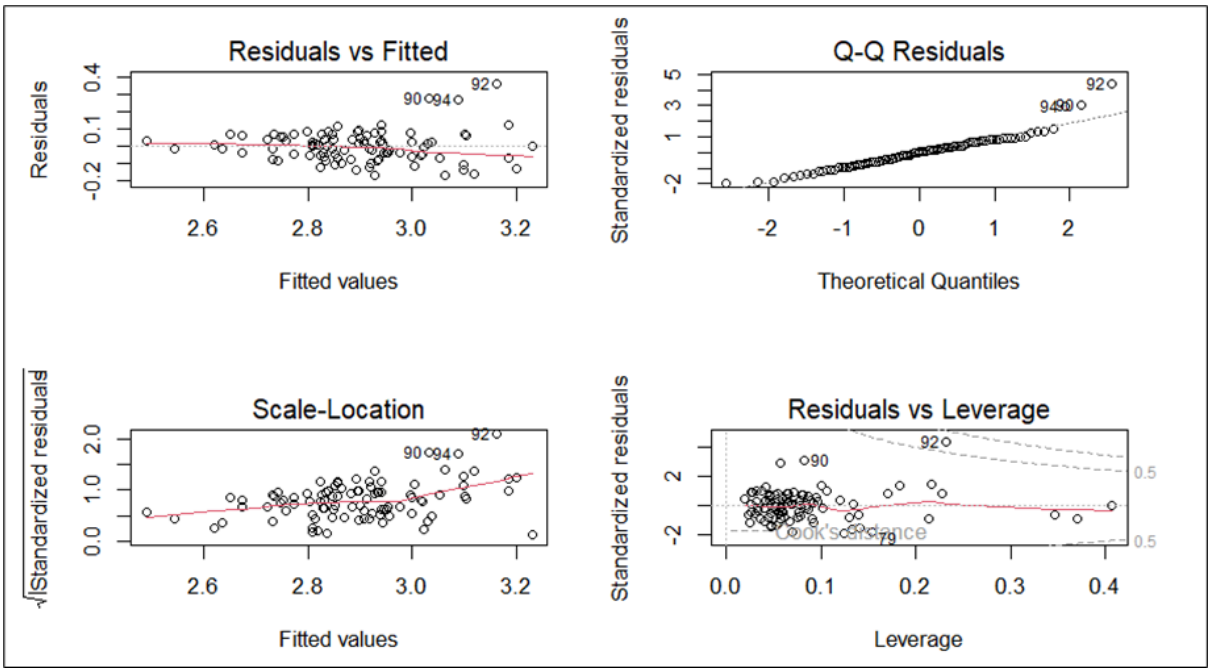
Graphique 6: QQ-plot des résidus pour vérifier la normalité dans le modèle log-log

Le graphique ci-dessus représente le QQ-plot des résidus du modèle log-log. La ligne rouge représente la distribution normale théorique, et les points correspondant aux résidus devraient s'aligner sur cette ligne si l'hypothèse de normalité était respectée. Globalement, les points suivent la ligne rouge, ce qui suggère une certaine tendance à la normalité. Cependant, on observe que plusieurs points s'écartent significativement de la ligne, en particulier aux extrémités.

Ces points éloignés sont des valeurs aberrantes ou individus atypiques. Ils posent problème dans le cadre du modèle car ils enfreignent l'hypothèse de normalité des résidus, essentielle pour assurer la validité statistique des résultats. Ces valeurs aberrantes peuvent indiquer des observations influentes qui peuvent fausser les estimations des coefficients, ou encore une mauvaise spécification du modèle.

Bien que les problèmes d'hétéroscédasticité aient été corrigés, la présence de ces valeurs atypiques doit être prise en compte pour améliorer la robustesse du modèle.

4.2.3. Etude des points influents



Graphique 7: Graphique des résidus standardisés en fonction du levier avec les distances de Cook

L'analyse des graphiques de diagnostic révèle des informations importantes concernant la qualité de l'ajustement du modèle. Le graphique "Residuals vs Fitted" montre que les résidus sont globalement répartis autour de zéro, ce qui suggère une absence de tendance systématique. Cependant, certains points, notamment les observations numérotées 90, 92 et 94, apparaissent comme atypiques et pourraient indiquer des cas influents. Ces mêmes points se distinguent également dans le Q-Q plot, où ils s'écartent de la droite de référence, remettant en question l'hypothèse de normalité pour ces observations extrêmes.

Le graphique "Scale-Location" met en évidence une légère augmentation de la variance des résidus pour des valeurs ajustées plus élevées, suggérant une possible hétéroscédasticité. Les points 90, 92 et 94 se démarquent à nouveau par leur influence notable. Enfin, le graphique "Residuals vs Leverage" permet de quantifier cette influence à l'aide de la distance de Cook. Le point 92, correspondant au département de la Seine-Saint-Denis(93), présente une distance de Cook proche de 1, indiquant qu'il exerce une influence significative sur les coefficients estimés du modèle. Les points 90, département de l'Essonne (91), et 94 département du Val-d'Oise (95), bien qu'ayant une distance de Cook inférieure à 0.5, nécessitent une attention particulière.

4.2.4. Analyse de robustesse de la régression linéaire sans les points influents

Il est essentiel d'évaluer l'impact de ces observations atypiques, afin de garantir la robustesse de l'analyse. Cette section permet de vérifier la robustesse de la régression après l'exclusion des points influents.

	Dependent variable:
	log(pcENT)
log(nbENT)	-0.231***
	(0.035)
log(POP)	0.249***
	(0.039)
DIPL	0.005***
	(0.001)
REV	0.035***
	(0.008)
log(gndENT)	0.060**
	(0.028)
txCHOM	0.046***
	(0.006)
METRO	0.034
	(0.023)
Constant	0.929***
	(0.291)
Observations	91
R^2	0.777
Adjusted R^2	0.758
Residual Std. Error	0.070 (df = 83)
F Statistic	41.289*** (df = 7; 83)
Note:	*p**p***p<0.01

Tableau 8 : Résumé du modèle 3 sans les points influents

Validité globale du modèle

Le test de Fisher confirme la validité globale du modèle : avec une p-value inférieure à 1%, l'hypothèse nulle est rejetée, indiquant que le modèle est globalement satisfaisant. Par ailleurs, le coefficient de détermination ajusté (R^2 ajusté) est de 75,8 %, nettement supérieur au modèle sans l'exclusion des individus influents (R2 ajusté de 68,6 %). Cela montre que 75,8 % de la variabilité de la proportion de création d'entreprises est expliquée par les variables incluses dans le modèle. Cela souligne un excellent pouvoir explicatif..

Significativité statistique des variables

Les résultats montrent que les variables log(nbENT), log(POP), DIPL, REV et log(txCHOM) sont significatives au seuil de 1 %, témoignant de leur impact important sur la variable expliquée. La variable log(gndENT) est également significative, mais au seuil de 5 %. En revanche, la variable METRO n'a pas d'influence statistiquement significative sur la variable à expliquer.

Hétéroscédasticité

Résultats du test de White :

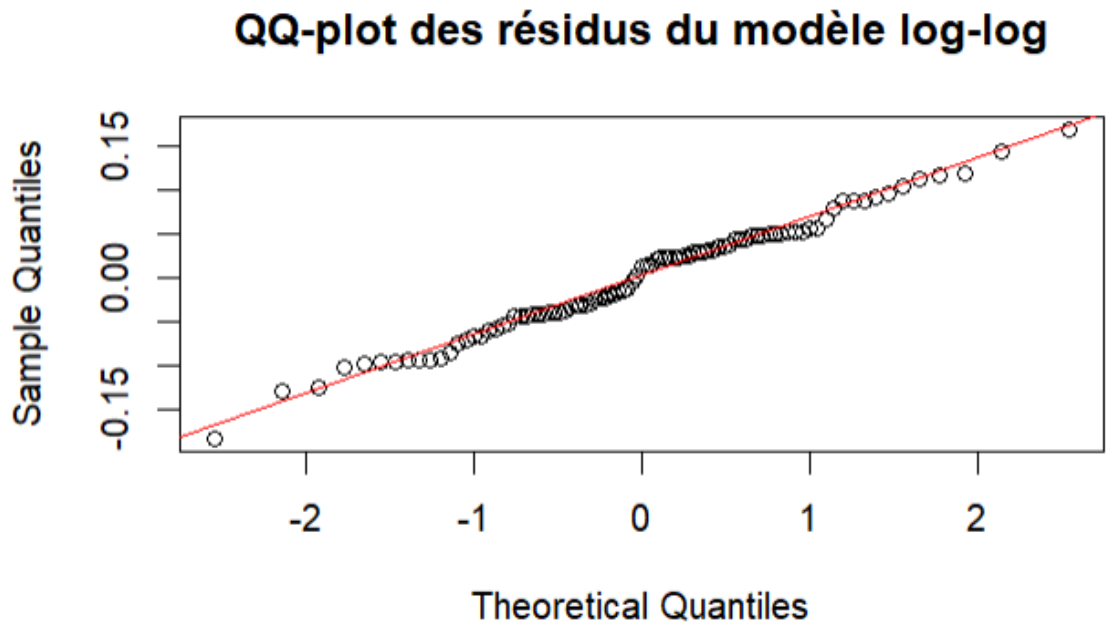
```
BP = 14.31, df = 11, p-value = 0.2163
```

Le test de White, avec une p-value supérieure à 5 %, indique que le modèle ne souffre pas d'hétéroscédasticité. Ainsi, l'hypothèse d'homoscédasticité est respectée, garantissant la validité des estimations des écarts-types.

Normalité des résidus

Résultats du test de Shapiro-Wilk :

```
W = 0.98883, p-value = 0.6378
```



Graphique 8: QQ-plot des résidus du modèle log-log sans les points influents

Le test de Shapiro-Wilk, avec une p-value supérieure à 5 %, confirme que les résidus suivent une distribution normale. Cela valide l'hypothèse de normalité, indispensable pour interpréter correctement les résultats et réaliser des inférences fiables.

Après avoir exclu les points influents, le modèle ajusté est robuste et présente des résultats fiables. Les variables explicatives ont, pour la plupart, une influence significative, et les tests statistiques montrent que les hypothèses fondamentales du modèle sont respectées. Le modèle sans les points influents est clairement meilleur que le modèle initial.

4.3. Autres tests

4.3.1. Test de changement structurel

Le Test de Chow va nous permettre de vérifier si les coefficients de la régression sont significativement différents entre le groupe de départements qui ont une métropole et le groupe de département qui n'a pas de métropole. Ce test permet de s'assurer que le modèle est stable sur l'ensemble des données.

Test de Chow :

$$H0: \ les\ coefficients\ sont\ identiques\ entre\ les\ deux\ groupes$$
$$H1: \ les\ coefficients\ diffèrent\ entre\ les\ deux\ groupes$$

Résultats

```
f(efp) = 0.93305, p-value = 0.9676
```

La p-value est supérieur à 5%, on accepte l'hypothèse nulle, il n'y a pas de preuve statistique d'une rupture structurelle dans le modèle. Les coefficients de régression semblent donc être stables sur l'ensemble des données.

4.3.2. Vérification de la spécification du modèle

Le RESET test (Regression Equation Specification Error Test) à détecter des erreurs de spécification dans un modèle de régression linéaire. Il vérifie si le modèle est correctement spécifié en testant l'inclusion de termes non linéaires des variables explicatives (par exemple puissances, produits ou logarithme) dans la régression.

RESET test :

$$H0: Le\ modèle\ initial\ est\ correctement\ spécifié$$
$$H1: \ Il\ existe\ des\ erreurs\ de\ spécification\ dans\ le\ modèle$$

Résultats

```
RESET = 0.75308, df1 = 2, df2 = 81, p-value = 0.4742
```

La p-value est supérieure à 5 %, nous acceptons l'hypothèse nulle selon laquelle le modèle est correctement spécifié. Cela signifie que les relations incluses dans le modèle sont appropriées et cohérentes avec les données disponibles.

Le modèle peut donc être considéré comme fiable pour interpréter les déterminants structurels de la création d'entreprises dans les départements étudiés, tout en tenant compte des ajustements réalisés pour renforcer sa validité statistique.

V. Conclusion

Cette étude a permis d'analyser les facteurs influençant la proportion de création d'entreprises dans les départements français en 2022. Grâce à un modèle économétrique log-log robuste sans la présence d'individus influents, nous avons identifié des relations significatives entre certaines variables explicatives et le dynamisme entrepreneurial.

L'équation estimée est la suivante :

$$\log\left(\widehat{pcENT}_i\right) = 0.929 - 0.231\,\log\left(nbENT_i\right) + 0.249\,\log\left(POP_i\right) + 0.005\,DIPL_i + 0.035\,REV_i + 0.060\,\log\left(gndENT_i\right) + 0.046\,\log\left(txCHOM_i\right) + 0.034\,METRO_i$$

Dans l'analyse des facteurs influençant la création d'entreprises, plusieurs tendances se dégagent. Tout d'abord, il semble que les départements où le nombre d'entreprises existantes est déjà important présentent une légère baisse de l'entreprenariat, une augmentation de 1 % du nombre d'entreprises existantes étant associée à une diminution de 0,231 % de la proportion de création d'entreprises, probablement en raison d'une saturation des opportunités. A l'inverse, un département dynamique, illustré par une population plus élevée, stimule l'entrepreneuriat : une hausse de 1 % de la population entraîne une augmentation de 0,249 % des nouvelles entreprises. De même, la qualification de la main d'œuvre joue un rôle essentiel, avec une augmentation d'un point de pourcentage de la part des diplômés dans un département conduit une hausse de 0,5 % de la proportion de création d'entreprises.

Par ailleurs, un revenu médian annuel plus élevé est indicatif d'un meilleur pouvoir d'achat, et agit comme un moteur : chaque hausse de 1 000 € de revenu est associée à une augmentation de 0,35 % des initiatives entrepreneuriales. Les grandes entreprises aussi ont un effet positif, avec une augmentation de 1 % de leur nombre conduisant à une hausse de 0,060 % de la création d'entreprises. De plus, bien que contre-intuitif, une hausse de 1 % du taux de chômage est associée à une augmentation de 0,046 % des créations, ce qui pourrait révéler une dynamique d'auto-emploi dans les zones en difficulté. Enfin, La présence ou non d'une métropole dans un département n'a pas d'effet statistiquement significatif sur la proportion de création d'entreprises, indiquant que d'autres facteurs structurels locaux jouent un rôle plus déterminant.

Limites et perspectives

Bien que ce modèle soit considéré comme robuste dans son ensemble, il présente plusieurs limites qui méritent d'être examinées pour mieux comprendre ses résultats et ses implications. Tout d'abord, il souffre de l'omission de certaines variables explicatives importantes. En effet, des facteurs tels que les spécificités culturelles locales, les politiques publiques propres à certaines régions ou encore la présence d'infrastructures stratégiques n'ont pas été intégrés dans le modèle. Ces éléments pourraient pourtant jouer un rôle déterminant dans l'explication des variations observées dans les données, ce qui limite la portée des conclusions tirées.

Ensuite, des problèmes d'endogénéité sont également présents. Il est possible que certaines des relations entre les variables explicatives et la variable dépendante soient influencées par des causalités inversées ou des corrélations non causales. Par exemple, une variable censée expliquer un

phénomène pourrait en réalité en être influencée, ce qui biaise l'interprétation des coefficients estimés. Ces problèmes, s'ils ne sont pas correctement corrigés, peuvent réduire la validité des résultats obtenus et conduire à des conclusions erronées.

Enfin, la fiabilité des résultats est également affectée par des erreurs de mesure. Bien que les données utilisées proviennent de sources officielles, elles peuvent comporter des imprécisions ou des biais liés aux méthodes de collecte ou de déclaration. Ces erreurs peuvent se répercuter sur la qualité des estimations du modèle et altérer sa capacité à représenter fidèlement la réalité qu'il cherche à modéliser. Ces limites soulignent l'importance de rester prudent dans l'interprétation des résultats et de considérer ces éventuelles imperfections dans les analyses futures.

Recommandations

Pour aller plus loin, une analyse dynamique prenant en compte l'évolution temporelle des données permettrait de mieux comprendre les tendances de création d'entreprises. De plus, des approches alternatives comme les modèles structurels pourraient améliorer la robustesse des estimations en intégrant les interactions complexes entre variables.

En conclusion, cette étude fournit une base solide pour éclairer les décideurs publics dans leur stratégie de réduction des disparités territoriales. Les politiques visant à améliorer le niveau de qualification, à soutenir les initiatives dans les zones à faible revenu et à renforcer les synergies entre entreprises devraient être proposées pour promouvoir un développement économique plus équilibré.

VI. Bibliographie

Claude Fouquet. L'inquiétant coup de frein des créations d'entreprises en France. Les Echos, octobre 2024
https://www.lesechos.fr/economie-france/conjoncture/coup-de-mou-des-creations-dentreprises-en-septembre-2126280

s.d. Data.gouv. Nombre de grandes entreprises en 2021
https://annuaire-entreprises.data.gouv.fr/rechercher?terme=&cp_dep_label=Ain+%2801%29&cp_dep_type=dep&cp_dep=01&fn=&n=&dmin=&dmax=&type=&label=&etat=A&sap=&naf=&nature_juridique=&tranche_effectif_salarie=&categorie_entreprise=GE

Denis Carré Nadine Levratto Mounir Amdaoud Luc Tessie. Synthèse Emploi et territoires : regards croisés sur les 22 métropoles. Economix, janvier 2021
https://hal.science/hal-03092526/document

s.d. INSEE. Création d'entreprise (en nombre) en 2022
https://statistiques-locales.insee.fr/#c=indicator&view=map2

s.d. INSEE. Unités légales (en nombre) en 2021
https://statistiques-locales.insee.fr/#c=indicator&i=ree_stocks.enntotnn&s=2021&t=A01&view=map2

s.d. INSEE. Population municipale en 2021
https://statistiques-locales.insee.fr/#c=indicator&i=pop_depuis_1876.pop&s=2021&t=A01&view=map2

s.d. INSEE. Part des diplômés d'un BAC+3 ou BAC+4 dans la pop. non scolarisée de 15 ans ou + 2021
https://statistiques-locales.insee.fr/#c=indicator&i=rp.pt_sup34&s=2021&t=A01&view=map2
s.d. INSEE. d'un BAC+5 ou plus dans la pop. non scolarisée de 15 ans ou + 2021
https://statistiques-locales.insee.fr/#c=indicator&i=rp.pt_sup5&s=2021&t=A01&view=map2

s.d. INSEE. Médiane du niveau de vie (en euro) 2021
https://statistiques-locales.insee.fr/#c=indicator&i=filosofi.med&s=2021&t=A01&view=map2

INSEE. taux de chômage par département en 2021, juin 2022
https://www.insee.fr/fr/statistiques/6453710?sommaire=6453776#tableau-figure2_radio2

Nadine Levratto, Denis Carré et Messaoud Zouikri. Dynamique des territoires et création d'entreprises : une analyse des départements français en 2008. Economix, juillet 2013
https://shs.hal.science/halshs-00840365/

Noora Shrestha. Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 2020, Vol. 8, No. 2, 39-42. https://pubs.sciepub.com/ajams/8/2/1/

J. Scott Long, Pravin K. Trivedi. SOME SPECIFICATION TESTS FOR THE LINEAR REGRESSION MODEL. Sociological Methods and Research 21:161-204. Reprinted in Bollen and Long, 1993. https://journals.sagepub.com/doi/10.1177/0049124192021002003