Mathématiques et Informatiques Appliquées aux Sciences Humaines et Sociales (MIASHS)

Parcours : MQME

Rapport du Projet

Analyse des déterminants départementaux du taux de pauvreté (ODD_DEP)

Étudiants : Perrin Kibiti & Essossinam Alayi
Formation : Master 1 - MQME

Encadrant : Camille Sabbah, Chef de département

Lille, le 19 décembre 2025

Table des matières

Chapitre 1

Introduction

La pauvreté territoriale constitue un enjeu majeur pour les politiques publiques. Ce rapport s'intéresse aux déterminants du taux de pauvreté à partir d'un ensemble d'indicateurs socio-économiques, résidentiels et sociaux. L'objectif est de comprendre quels facteurs sont les plus fortement associés à la variation du taux de pauvreté entre départements, en mobilisant une analyse en composantes principales (ACP) et un modèle de régression linéaire multiple.

Concrètement, nous avons d'abord exploré la base de données (ODD_DEP) pour voir quelles informations elle contient, comment sont réparties les variables et si certains chiffres paraissent anormaux. Nous avons ensuite nettoyé la base, en gérant les valeurs manquantes, en vérifiant la présence de valeurs extrêmes et, si nécessaire, en adaptant certaines variables pour qu'elles soient utilisables dans l'analyse.

Dans un deuxième temps, nous avons réalisé une analyse en composantes principales (ACP) pour résumer l'information, repérer les grandes dimensions qui structurent les départements et identifier les indicateurs les plus liés au taux de pauvreté. Enfin, à partir de ces résultats, nous avons construit un modèle de régression linéaire afin de mesurer plus précisément l'effet des différentes variables sur le taux de pauvreté et de pouvoir interpréter les résultats.

Chapitre 2

Données et préparation

2.1 Source des données

Les données utilisées dans ce travail viennent de l'INSEE. Elles rassemblent différents indicateurs socio-économiques, de logement et de contexte social, mesurés pour chaque département. Le fichier considéré dans cette étude est nommé ODD_DEP a été fourni au format CSV.

La base initiale contient 70 902 entrées (lignes) et 77 variables (colonnes). Parmi elles, le taux de pauvreté, qui est au centre de notre analyse, et toute une série d'indicateurs qui décrivent la population, l'emploi, les revenus, les conditions de logement ou encore certaines dimensions sociales. Dans la suite du rapport, nous présentons plus en détail la définition du taux de pauvreté retenue et la description des principales variables utilisées.

2.2 Nettoyage et traitement des données

Dans un premier temps, nous avons vérifié la qualité des données. Concrètement, nous avons regardé s'il y avait des valeurs manquantes, des doublons ou chiffres clairement incohérents, en examinant les statistiques de base et quelques graphiques simples. Pour cette étude, n'avous considéré que l'année "2021".

Nous avons ensuite traité les valeurs manquantes. Quand une observation était incomplète, nous l'avons retirée de la base, nous avons choisi de conserver les variables concernées, afin de garder une analyse facile à interpréter.

Enfin, nous avons mis la base en forme pour l'analyse : vérification que chaque variable avait le bon type (numérique, catégorielle), et simplification des noms de variables pour qu'ils soient plus lisibles dans les tableaux et les graphiques.

Le nettoyage s'est finalement fait en deux étapes : la première consistait à obtenir un tableau de dimension reduite par rapport à la base initiale, ainsi nous avons obtenu une base de 96 lignes et 399 colonnes (Réf. A.1).

La seconde étape consistait à réaliser une ACP sur un jeu de données de 30 variables incluant le taux de pauvreté (Réf. A.2). Pour ce qui concerne la régression linéaire, nous avons effectué une sélection de 10 variables jugées plus pertinentes, afin de pouvoir estimer un modèle par moindres carrés ordinaires (MCO) aussi équilibré que possible (Réf. A.3).

Le tableau ci-dessous présente la description des variables sélectionnées.

N°	Base individu	Description des variables
1	Taux_pvt_total	Taux de pauvreté monétaire (60 % du niveau de vie médian) total et par tranche d'âge des individus (de 0 à 29 ans, 30 à 39 ans, de 40 à 49 ans, de 50 à 59 ans, de 60 à 74 ans, 75 ans ou plus).
2	Part_pop75	Part des 75 ans ou plus dans la population totale.
3	Prg_dechets	Programme ou indicateur lié à la gestion/collecte des déchets (fréquence, couverture ou types de programme).
4	Enerc2e_classiq_ter	Part ou quantité d'énergie de type classique (énergie fossile).
5	Taux_crea_etab	Taux de créations d'établissements dans le secteur concerné.
6	Part_depl_dom_trav	Part des déplacements domicile-travail réalisés sans transport collectif (voiture personnelle, vélo, marche).
7	Parc_pl_critair2	Part des voitures particulières et des poids lourds classés critère 2 fonctionnant au gazole dans l'ensemble du parc.
8	Infrac_tx_cambrio_	Taux d'infractions de type cambriolages (nombre pour 1 000 / 10 000 habitants).
9	Reemploi_asso	Nombre de structures (associations) chargées du réemploi.
10	Log_hlm_phab	Nombre de logements sociaux (HLM) par habitant ou pour 10 000 habitants.
11	Part_critair_	Part des véhicules particuliers en classement Critère 1 ou 0 (véhicules les moins polluants).
12	Locaux_raccor	Nombre ou part de locaux raccordés à un réseau (électricité, eau, assainissement, fibre).
13	Sal_hor_net_hom_4	Salaire horaire net des hommes dans la catégorie socioprofessionnelle 4.
14	Part_emp_ess	Part des emplois relevant de l'économie sociale et solidaire (coopératives, associations, mutuelles, etc.).
15	Nb_etab_seveso_sb	Nombre d'établissements industriels à risque (classés Seveso seuil bas).

N°	Base individu	Description des variables
16	Enerc2e_prepar_res	Indicateur lié à la consommation d'énergie dans les logements en situation de précarité énergétique.
17	Ges_polluant_pfc	Quantité d'émissions de gaz polluants de type PFC (perfluorocarbones ; gaz à effet de serre).
18	Nb_etab_seveso_sh	Nombre d'établissements industriels à risque (classés Seveso seuil haut).
19	Emp_ess_coop	Nombre d'emplois dans les coopératives relevant de l'économie sociale et solidaire.
20	Sal_hor_net_hom_5	Salaire horaire net des hommes pour la catégorie socioprofessionnelle 5.
21	Ecart_tx_emp_f_h	Écart entre le taux d'emploi des femmes et le taux d'emploi des hommes âgés de 55 à 64 ans.
22	Reemploi_ent	Quantité de biens réutilisés ou réemployés via les entreprises (hors associations).
23	Quantite_dechets_org	Quantités de déchets d'origine animale ou végétale produites ou collectées.
24	Taux_lgmt_suroc	Part des logements en situation de sur-occupation au regard de la taille du ménage.
25	Quantite_dd_trait	Quantité de déchets dangereux traités dans des installations adaptées.
26	Ecorener_ind	Indicateur de l'énergie renouvelable consommée ou produite dans l'industrie.
27	Vol_prevel_sout_aep	Volume prélevé dans les eaux souterraines destiné à l'alimentation en eau potable (AEP).
28	Infrac_tx_degrad	Taux d'infractions de type dégradations (vandalisme, détérioration de biens publics ou privés).
29	CO2_emissions_bio	Émissions de CO_2 liées à la biomasse (combustion de matières organiques/renouvelables).
30	Infrac_tx_violences_	Taux d'infractions pour violences intrafamiliales (nombre de cas pour 1 000/10 000 habitants).

TABLE 2.1 – Description des variables de la base de données

2.3 Exploration de la base de données

Dans cette section, nous proposons une étude descriptive des variables retenues pour l'application de l'ACP. Il s'agit de présenter les premières statistiques descriptives (moyennes, écarts-types, minima, maxima) et, le cas échéant, quelques graphiques permettant de visualiser la distribution des principales variables autour de leur moyennes respectvives (Réf. A.2).

```
summary(data_desc)
taux_pvt_total   part_pop75_     sal_hor_net_homme_cs_4 sal_hor_net_homme_cs_5
Min.   : 9.10   Min.   : 5.180   Min.   :14.94          Min.   :11.00
1st Qu.:12.57   1st Qu.: 9.265   1st Qu.:15.99          1st Qu.:11.59
Median :14.80   Median :10.915   Median :16.45          Median :11.76
Mean   :14.88   Mean   :10.841   Mean   :16.53          Mean   :11.81
3rd Qu.:16.23   3rd Qu.:12.495   3rd Qu.:16.86          3rd Qu.:12.00
Max.   :28.40   Max.   :15.030   Max.   :21.85          Max.   :13.03
log_hlm_phab_    part_emp_ess_   taux_crea_etab_ infrac_tx_violences_intrafam_
Min.   : 234.4   Min.   : 5.82   Min.   :11.88   Min.   :1.228
1st Qu.: 487.3   1st Qu.:10.46   1st Qu.:14.54   1st Qu.:1.762
Median : 656.2   Median :12.18   Median :16.26   Median :2.050
Mean   : 665.3   Mean   :12.31   Mean   :16.30   Mean   :2.065
3rd Qu.: 780.0   3rd Qu.:13.34   3rd Qu.:17.34   3rd Qu.:2.345
Max.   :1408.7   Max.   :26.50   Max.   :22.24   Max.   :3.267


ecart_tx_emp_f_h_55_64 taux_lgmt_suroccupes_
Min.   :-11.160        Min.   : 1.020
1st Qu.: -4.630        1st Qu.: 1.558
Median : -2.490        Median : 2.075
Mean   : -3.065        Mean   : 3.080
3rd Qu.: -1.492        3rd Qu.: 2.953
Max.   :  1.900        Max.   :18.800
infrac_tx_cambriolages_
Min.   : 1.112
1st Qu.: 3.074
Median : 3.996
Mean   : 4.262
3rd Qu.: 5.039
Max.   :10.955
```

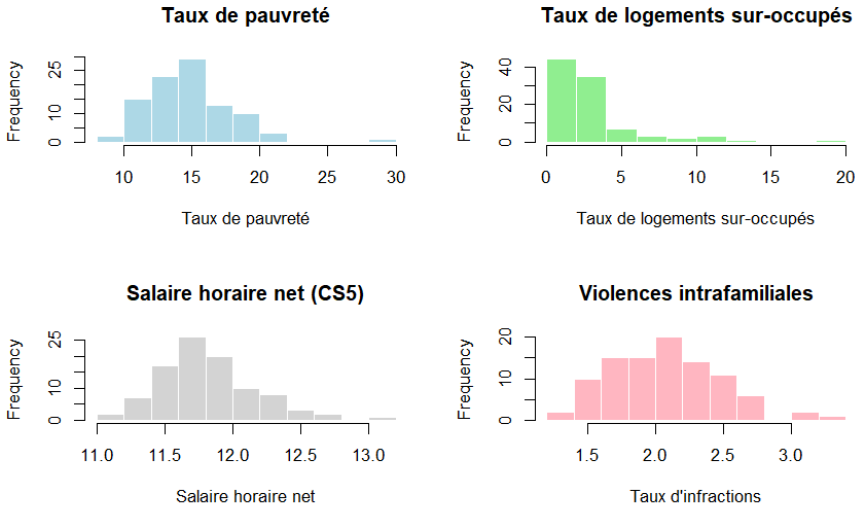

FIGURE 2.1 – Distribution de quelques variables principales

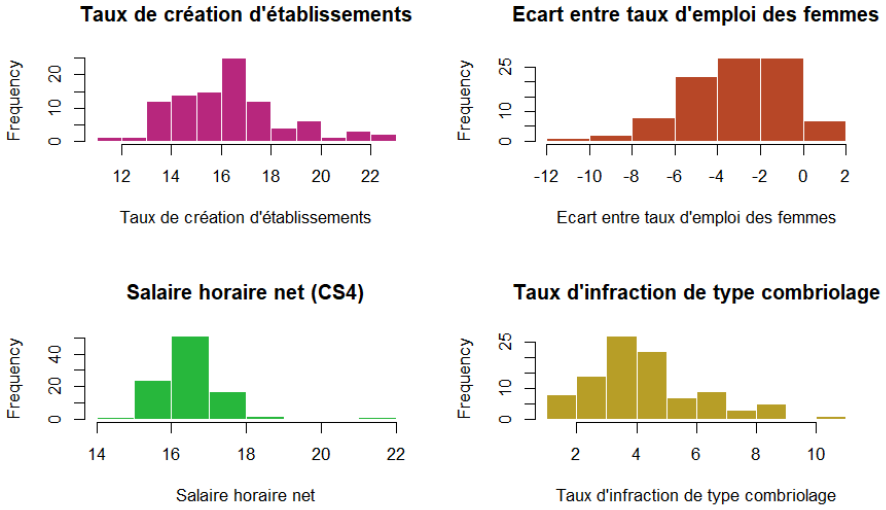

FIGURE 2.2 – Distribution de quelques variables principales

L'observation des moyennes et des écarts-types montre que les variables sont réparties de façon régulière autour de leur moyenne, et que leurs distributions sont globalement normales. Cette situation est très favorable pour l'analyse en composantes principales : la normalité des variables garantit la fiabilité des résultats et facilite l'interprétation statistique des axes factoriels. Autrement dit, la qualité des données analysées permet d'obtenir une synthèse solide et représentative de la structure globale du jeu de données (ODD_DEP).

Chapitre 3

Analyse en composantes principales

3.1 Objectifs de l'ACP

Nous avons mis en place un algorithme permettant à la fois de vérifier les corrélations entre variables et de supprimer celles dont la variance est nulle, afin de nous concentrer sur les données les plus pertinentes. Juste après cette étape, nous avons procédé à la réalisation d'une ACP préliminaire afin de mieux identifier les variables contribuant le plus à la formation du premier axe.

Notre objectif ici est de simplifier et résumer nos données en quelques grands axes, pour mieux comprendre les tendances et voir ce qui distingue vraiment les observations les unes des autres. C'est donc une façon de rendre les données complexes plus faciles à lire, à comparer et à interpréter, pour ensuite tirer des conclusions plus claires.

3.2 Résultats principaux

L'application de l'ACP nous a permis de faire des observations et analyses suivantes :

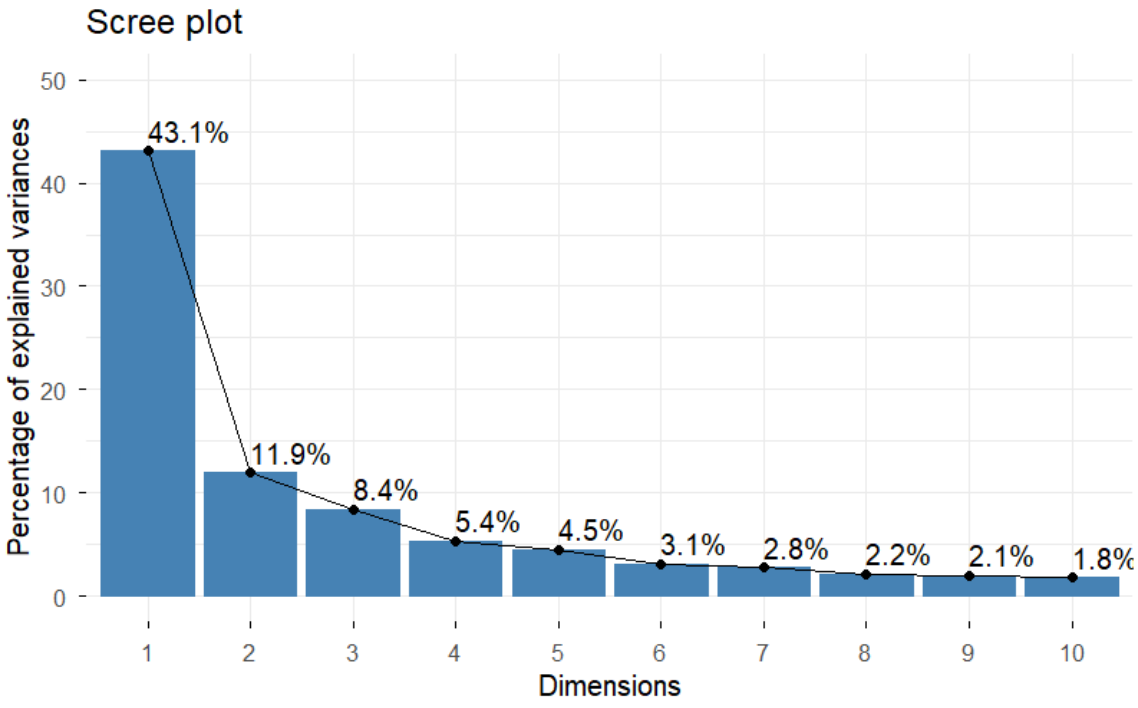

FIGURE 3.1 – Scree Plot

1. **Choix des axes principales**

   D'après le output, nous remarquons que les cinq premières valeurs propres sont supérieures à 1, ce qui nous permet de conserver les composantes associées à ces valeurs propres d'après la règle de Kaiser. Cependant, d'après le graphique ci-dessus, nous retenons les composantes principales suivantes :

   Composante 1 : 43,1% de variance de la variabilité du nuage des individus est expliqueé.

   Composante 2 : 11,9% de variance de la variabilité du nuage des individus est expliquée.

   Composante 3 : 8,4% de variance de la variabilité du nuage des individus est expliquée.

   Ainsi 63,4% (soit 43,1% + 11,9% + 8,4%) de la variance cumulative est expliquée par les trois premières composantes lesquelles par conséquent expliquent quasiment toute l'information engendrée par la base de données.

   La première composante résume la majorité de l'information et les trois premiers axes captent pratiquement toute la structure. De plus, on remarque la formation d'un coude au niveau de la quartième composante (5,4%), ce qui nous conduit au choix des trois axes pour la suite notre analyse d'après la règle de Coude.

2. **Interprétation des axes par rapport à la projection des variables sur le plan (1,2)**

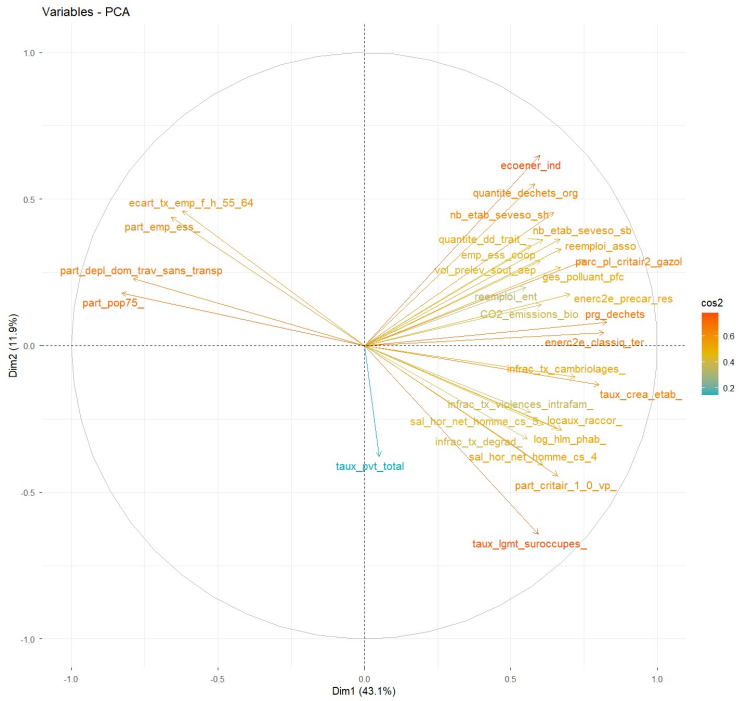

FIGURE 3.2 – Plan (1,2)

Les deux axes principaux de l'ACP (Dim1 et Dim2) correspondent aux directions dans lesquelles les données varient le plus.

Axe 1 (Dim1, 43.10%)

Sur le côté négatif, nous avons des départements résidentiels, avec davantage de personnes âgées, plus de déplacements domicile-travail sans transports collectifs et des structures d'emploi. Par contre du côté positif, nous avons des départements urbanisés et industrialisés, où l'on retrouve beaucoup d'activités économiques, de déchets, de pollution, de sites industriels à risque, de précarité énergétique et de délinquance (cambriolages, dégradations).

L'axe 1 oppose des départements très urbanisés, pollués et marqués par la précarité et la délinquance à des départements plutôt résidentiels, plus âgés et dépendants de la voiture

Axe 2 (Dim2, 11.94%)

Sur le coté positif, nous avons les départements qui concentrent surtout des problèmes liés à l'environnement et aux installations industrielles (pollutions, déchets, sites Seveso, enjeux énergétiques collectifs). Par contre sur le côté négatif, nous avons les départements où les problèmes dominants sont le logement et le quotidien des habitants, avec sur-occupation des logements, salaires bas et forte petite délinquance.

Par conséquent,l'axe 2 oppose donc des départements plutôt marqués par des enjeux environnementaux/industriels à des départements centrés sur des problèmes de logement, sur-occupation et insécurité quotidienne.

3. **Interprétation des axes par rapport à la projection des variables sur le plan (1,3)**

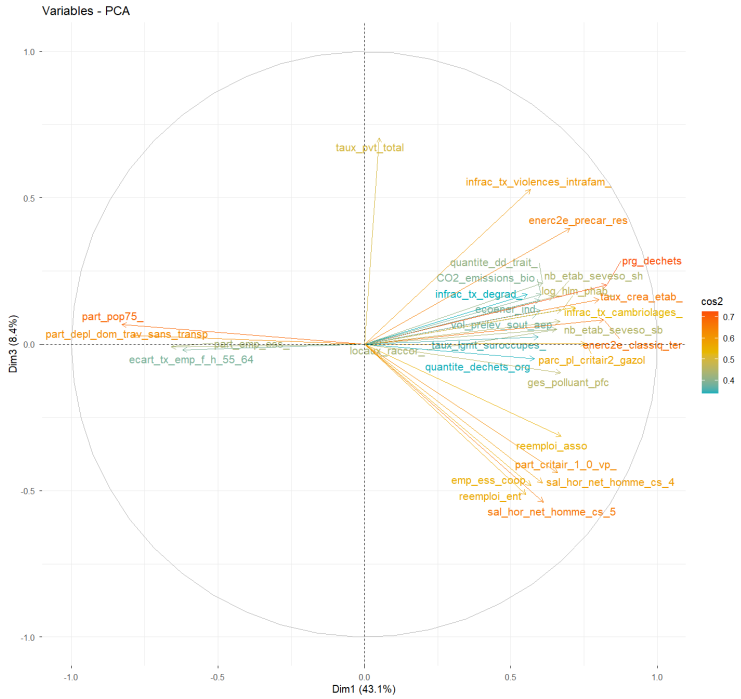

FIGURE 3.3 – Plan (1,3)

Axe 3 (Dim3, 8,4%)

Sur le côté positif de l'axe, nous avons des départements marqués par un fort taux de pauvreté (précarité), des violences intrafamiliales et des risques environnementaux. Par contre sur le côté négatif, il y a des départements où l'on observe davantage de recyclage/réemploi, des contextes un peu plus protecteurs ou organisés socialement.

En tenant compte de l'interprétation des trois axes, nous remarquons que le plan (1,2) résume plus de la moitié de l'information. Il représente clairement les différences entre types de départements. Cependant, l'axe 3 peut éventuellement nous orienter sur les nuances concernant certaines formes de précarité.

4. **Projection des individus (départements) sur le plan (1,2)**

Sur le graphique des individus, chaque point représente un département, la position indique à quel "profil" il correspond selon les deux axes :

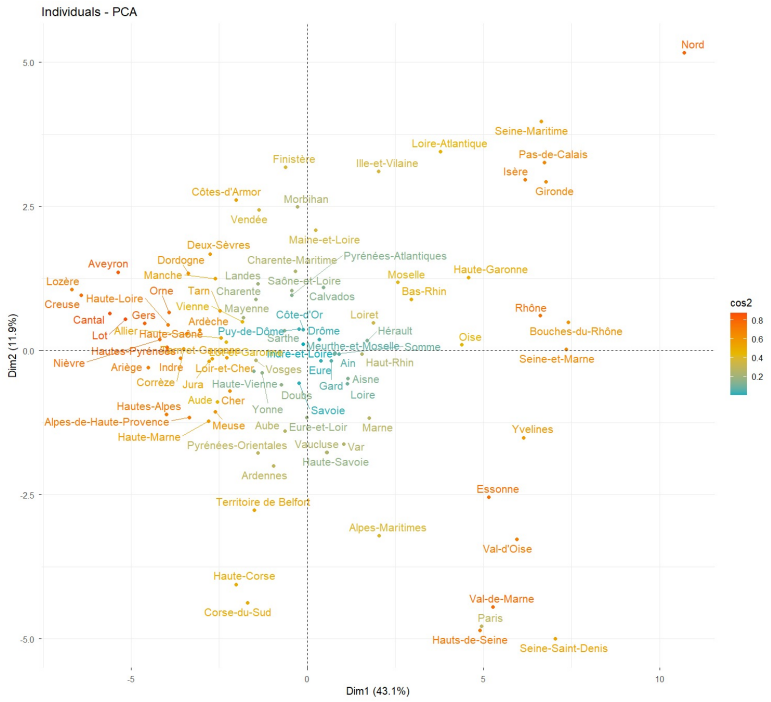

FIGURE 3.4 – Projection des individus

A gauche, nous voyons des départements très ruraux et peu denses comme le Cantal, la Lozère, la Creuse, l'Aveyron, les Alpes-de-Haute-Provence. Par contre à droite, nous avons des départements très urbains : Paris, la petite couronne, mais aussi Rhône, Haute-Garonne, Bouches-du-Rhône, Gironde, Nord... bref, les grands pôles métropolitains.

Cependant, plusieurs départements de l'Ouest/Nord-Ouest comme Finistère, Côtes-d'Armor, Loire-Atlantique, Vendée, Maine-et-Loire ainsi que le Nord ont un profil particulier qui les rassemble entre eux. Par ailleurs plus bas, il y a la Corse et quelques départements du Sud-Est cpmme Alpes-Maritimes, Var, Vaucluse qui se détachent avec un comportement très spécifique par rapport aux autres.

La majorité des départements sont regroupés au centre : ils ont un profil assez moyen en terme de représentation, ni très urbain ni très rural, sans forte particularité sur ces deux axes. Quelques départements comme Paris et petite couronne, Nord, Corse, Alpes-Maritimes sortent vraiment du lot : ce sont donc ceux qui ont le profil le plus atypique et qui sont le mieux décrits par ce plan.

5. **Biplot de l'ACP**

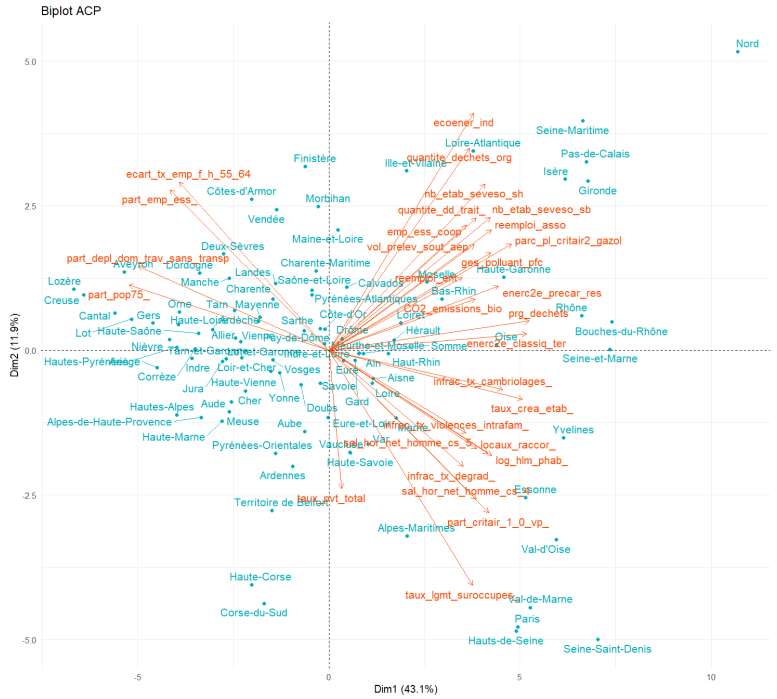

FIGURE 3.5 – Biplot de l'ACP

Les départements très urbains et les départements ruraux ne vivent pas du tout les mêmes réalités, et les problèmes auxquels ils font face ne sont pas les mêmes.

Dans les départements ruraux, comme la Lozère, la Creuse, le Cantal ou la Haute-Loire, il y a moins d'habitants et la population est plus âgée. Ce sont des départements plus calmes, moins industrialisés, où l'on rencontre moins de problèmes sociaux typiques des grandes villes, comme une forte délinquance ou une grosse pression sur le logement.

À l'inverse, les départements urbains sont plus industrialisés, donc plus pollués, et attirent beaucoup de monde grâce à leur dynamisme économique et à leurs infrastructures. Ils concentrent une population importante, ce qui fait apparaître de nombreux problèmes sociaux. Parmi eux, certains départements comme le Nord, la Seine-Maritime, le Pas-de-Calais ou la Gironde sont surtout marqués par des enjeux environnementaux liés à l'industrie. D'autres, comme la Corse, les Alpes-Maritimes, Paris et sa banlieue ou le département de Belfort, souffrent davantage de la forte densité de population : difficultés à se loger, tensions sociales, délinquance et violence.

Chapitre 4

Modélisation par régression linéaire

4.1 Spécification du modèle avec constante

A partir de la modélisation qui est faite ici, sous souhaitons expliquer les variations du taux de pauvreté (`taux_pvt_total`) entre départements à partir d'un ensemble de variables socio-économiques, résidentielles et sociales retenues à l'issue de l'exploration des données et de l'ACP. Le modèle estimé est une régression linéaire multiple de la forme :

$$\text{taux\_pvt\_total}_i = \beta_0 + \sum_{k=1}^{10} \beta_k X_{ki} + \varepsilon_i,$$

où ε_i représente le terme d'erreur, supposé centré, de variance constante et indépendamment distribué.

Matriciellement, le modèle s'écrit :

$$Y = X\beta + \varepsilon$$

avec

$$Y = \begin{pmatrix} \text{taux\_}pvt\_total^{(1)} \\ \vdots \\ \text{taux\_}pvt\_total^{(n)} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_1^{(1)} & X_2^{(1)} & X_3^{(1)} & X_6^{(1)} & X_7^{(1)} & \cdots & X_{10}^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_1^{(n)} & X_2^{(n)} & X_3^{(n)} & X_6^{(n)} & X_7^{(n)} & \cdots & X_{10}^{(n)} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

où, pour tout Département i :

$$
\begin{aligned}
X_1^{(i)} &= \text{part\_pop75\_}^{(i)},\\
X_2^{(i)} &= \text{sal\_hor\_net\_homme\_cs\_4}^{(i)},\\
X_3^{(i)} &= \text{sal\_hor\_net\_homme\_cs\_5}^{(i)},\\
X_4^{(i)} &= \text{ecart\_tx\_emp\_f\_h\_55\_64}^{(i)},\\
X_5^{(i)} &= \text{taux\_lgmt\_suroccupes\_}^{(i)},\\
X_6^{(i)} &= \text{log\_hlm\_phab\_}^{(i)},\\
X_7^{(i)} &= \text{part\_emp\_ess\_}^{(i)},\\
X_8^{(i)} &= \text{taux\_crea\_etab\_}^{(i)},\\
X_9^{(i)} &= \text{infrac\_tx\_violences\_intrafam\_}^{(i)},\\
X_{10}^{(i)} &= \text{infrac\_tx\_cambriolages\_}^{(i)}.
\end{aligned}
$$

Les hypothèses de Gauss–Markov restent inchangées avec cette notation compacte.
Avec les hypothèses classiques (Gauss–Markov) associées au modèle linéaire :

1. Linéarité :

$$\mathbb{E}[taux\_pvt\_total^{(i)} \mid X] = \beta_0 + \beta_{\text{part\_pop75\_}}\,\text{part\_pop75\_}^{(i)} + \cdots$$

$$+\beta_{\text{cambriolages}}\,\text{infrac\_tx\_cambriolages\_}^{(i)}.$$

2. Exogénéité : $\mathbb{E}[\varepsilon_i \mid X] = 0$ pour tout i.
3. Absence de multicolinéarité parfaite : les colonnes de X sont linéairement indépendantes (donc $X^\top X$ est inversible).
4. Homoscédasticité : $\mathrm{V}(\varepsilon_i \mid X) = \sigma^2$ pour tout i.
5. Indépendance des erreurs : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X) = 0$ pour $i \neq j$.

Plus concrètement, le modèle MCO estimé dans R est le suivant :

$$\text{taux\_pvt\_total} = \beta_0+\beta_1\text{part\_pop75\_}+\beta_2\text{sal\_hor\_net\_homme\_cs\_4}+\beta_3\text{sal\_hor\_net\_homme\_cs\_5}$$

$$+\beta_4\text{ecart\_tx\_emp\_f\_h\_55\_64} + \beta_5\text{taux\_lgmt\_suroccupes\_} + \beta_6\text{log\_hlm\_phab\_}$$

$$+\beta_7\text{part\_emp\_ess\_} + \beta_8\text{taux\_crea\_etab\_} + \beta_9\text{infrac\_tx\_violences\_intrafam\_}+$$

$$\beta_{10}\text{infrac\_tx\_cambriolages\_}$$

Les variables explicatives ont été choisies pour couvrir les principaux déterminants théoriques de la pauvreté (structure démographique, niveau de rémunération, inégalités d'emploi, logement, fragilités sociales), en s'appuyant sur les résultats de l'ACP et en veillant à limiter la redondance entre indicateurs (multicolinéarité).

4.2 Résultats du modèle (A.3)

Voici les résultats (output) de l'estimation du modèle par MCO :

```
Call:
lm(formula = taux_pvt_total ~ part_pop75_ + sal_hor_net_homme_cs_4 +
sal_hor_net_homme_cs_5 + ecart_tx_emp_f_h_55_64 + taux_lgmt_suroccupes_ +
log_hlm_phab_ + part_emp_ess_ + taux_crea_etab_ + infrac_tx_violences_intrafam_ +
infrac_tx_cambriolages_, data = df)

Residuals:
Min      1Q  Median      3Q     Max
-3.2695 -1.1723 -0.0989  0.9037  4.2393

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)                  81.160908  11.211375   7.239 1.87e-10 ***
part_pop75_                   0.258840   0.167110   1.549 0.125117
sal_hor_net_homme_cs_4       -0.068401   0.379457  -0.180 0.857377
sal_hor_net_homme_cs_5       -6.157054   0.881645  -6.984 5.98e-10 ***
ecart_tx_emp_f_h_55_64       -0.437077   0.114843  -3.806 0.000266 ***
taux_lgmt_suroccupes_         0.528271   0.094924   5.565 2.99e-07 ***
log_hlm_phab_                 0.001278   0.001153   1.108 0.271020
part_emp_ess_                 0.071288   0.087817   0.812 0.419185
taux_crea_etab_              -0.324208   0.164108  -1.976 0.051448 .
infrac_tx_violences_intrafam_  1.502882   0.605128   2.484 0.014971 *
infrac_tx_cambriolages_       0.523768   0.140213   3.736 0.000338 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.687 on 85 degrees of freedom
Multiple R-squared:  0.7241, Adjusted R-squared:  0.6917
F-statistic: 22.31 on 10 and 85 DF,  p-value: < 2.2e-16
```

Qualité globale du modèle

- Le $R^2 = 0{,}7241$ signifie que le modèle explique environ 72% de la variabilité du taux de pauvreté entre Départements.
- Le test de Fisher est très significatif (p-value $< 2{,}2 \times 10^{-16}$), donc au moins une des variables explique significativement le taux de pauvreté.

Constante

- L'ordonnée à l'origine (intercept) vaut 81,16 (significative au seuil 0,1%).
- C'est le taux de pauvreté « de base » prédit quand toutes les variables explicatives valent 0 (valeur surtout interprétée comme niveau moyen d'ancrage du modèle, pas forcément réaliste).

Variables non significatives (p-value $> 0{,}05$) Elles n'ont pas d'effet clairement détectable dans ce modèle :

- `part_pop75_` ($p = 0{,}125$) : la part de 75 ans et plus n'est pas significativement liée au taux de pauvreté, toutes choses égales par ailleurs.
- `sal_hor_net_homme_cs_4` ($p = 0{,}858$) : le salaire horaire des ouvriers/employés qualifiés (cs4) n'a pas d'effet significatif.
- `log_hlm_phab_` ($p = 0{,}271$) : la part de logements sociaux n'est pas significative dans ce modèle avec constante.
- `part_emp_ess_` ($p = 0{,}420$) : la part d'emplois dans l'économie sociale et solidaire n'est pas significative.
- `taux_crea_etab_` ($p = 0{,}051$) : la création d'établissements est à la limite de la significativité (p juste au-dessus de 0,05), effet négatif mais à interpréter avec prudence.

Variables significatives

Effets (toutes choses égales par ailleurs) sur le taux de pauvreté :

- `sal_hor_net_homme_cs_5` : coefficient estimé $\widehat{\beta_3} = -6{,}16$, $p < 0{,}001$. Quand le salaire horaire moyen des cadres/professions les plus qualifiées (cs5) augmente d'une unité, le taux de pauvreté diminue fortement en moyenne. Des Départements mieux rémunérés en haut de l'échelle sont associés à un moindre taux de pauvreté.
- `ecart_tx_emp_f_h_55_64` : coefficient $\widehat{\beta_4} = -0{,}44$, $p < 0{,}001$. Plus l'écart de taux d'emploi femmes–hommes 55–64 ans est élevé, plus le taux de pauvreté est faible, ce qui est contre-intuitif et doit être interprété comme un effet de structure ou de corrélation entre variables plutôt qu'un lien causal direct.
- `taux_lgmt_suroccupes_` : coefficient $\widehat{\beta_5} = 0{,}53$, $p < 0{,}001$. Plus la part de logements sur-occupés augmente, plus le taux de pauvreté augmente. La sur-occupation est donc un bon indicateur de précarité résidentielle.
- `infrac_tx_violences_intrafam_` : coefficient $\widehat{\beta_9} = 1{,}50$, $p \approx 0{,}015$. Les départements où le taux de violences intrafamiliales est plus élevé ont, en moyenne, un taux de pauvreté plus élevé.
- `infrac_tx_cambriolages_` : coefficient $\widehat{\beta_{10}} = 0{,}52$, $p < 0{,}001$. Les départements avec davantage de cambriolages ont aussi un taux de pauvreté plus élevé.

Synthèse

Dans ce modèle avec constante, les facteurs les plus nettement associés à la pauvreté territoriale sont les bas salaires en haut de l'échelle (cs5), la sur-occupation des logements et les indicateurs de climat social tendu (violences intrafamiliales, cambriolages), tandis que plusieurs variables démographiques ou structurelles ne ressortent pas comme significatives.

4.3 Modélisation par régression linéaire sans constante

4.3.1 Spécification du modèle

Dans cette variante, l'objectif reste d'expliquer les variations du taux de pauvreté monétaire (`taux_pvt_total`) entre départements à partir des mêmes variables socio-économiques, résidentielles et sociales, mais en estimant un modèle sans constante. Le modèle estimé est alors une régression linéaire multiple de la forme :

$$\text{taux\_pvt\_total}_i = \sum_{k=1}^{10} \beta_k X_{ki} + \varepsilon_i,$$

où ε_i représente le terme d'erreur, supposé centré, de variance constante et indépendamment distribué.

Ici, nous supposons les mêmes hypothèses de Gauss-Markov, mais appliquées à un modèle où l'ordonnée à l'origine est fixée à 0.

On considère le modèle

$$Y = X^*\beta^* + \varepsilon,$$

où X^* est la matrice des variables explicatives sans colonne de constantes et β^* le vecteur des coefficients associés.

Hypothèses de Gauss–Markov pour le modèle sans constante

— Linéarité conditionnelle :

$$\mathbb{E}[Y \mid X^*] = X^*\beta^*.$$

— Exogénéité :

$$\mathbb{E}[\varepsilon \mid X^*] = 0.$$

— Absence de multicolinéarité parfaite : les colonnes de X^* sont linéairement indépendantes, donc $(X^*)^\top X^*$ est inversible.

— Homoscédasticité et indépendance :

$$V(\varepsilon \mid X^*) = \sigma^2 I_n.$$

Sous ces hypothèses, l'estimateur des moindres carrés

$$\hat{\beta}^* = \left((X^*)^\top X^*\right)^{-1} (X^*)^\top Y$$

est le meilleur estimateur linéaire sans biais (propriété BLUE) pour ce modèle sans constante.

Spécificité du modèle sans constante

La différence essentielle avec le modèle avec constante est l'hypothèse implicite

$$\mathbb{E}[Y \mid X^* = 0] = 0,$$

c'est-à-dire que l'hyperplan de régression est contraint à passer par l'origine.

Plus concrètement, le modèle MCO sans constante estimé dans R s'écrit :

$$\text{taux\_pvt\_total} = \beta_1 \text{part\_pop75\_} + \beta_2 \text{sal\_hor\_net\_homme\_cs\_4} + \beta_3 \text{sal\_hor\_net\_homme\_cs\_5} +$$

$$\beta_4 \text{ecart\_tx\_emp\_f\_h\_55\_64} + \beta_5 \text{taux\_lgmt\_suroccupes\_} + \beta_6 \text{log\_hlm\_phab\_} +$$

$$\beta_7 \text{part\_emp\_ess\_} + \beta_8 \text{taux\_crea\_etab\_} + \beta_9 \text{infrac\_tx\_violences\_intrafam\_} +$$

$$\beta_{10} \text{infrac\_tx\_cambriolages\_}$$

L'absence de constante revient à imposer que, lorsque toutes les variables explicatives valent zéro, le taux de pauvreté prédit par le modèle soit lui aussi nul. Cette spécification n'est justifiée que si ce point d'origine a un sens dans le contexte empirique.

4.3.2 Résultats du modèle sans constante (A.4)

Voici le output :

```
        Residuals:
    Min      1Q  Median      3Q     Max
-4.9365 -1.3555 -0.2215  1.1749  5.0887

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
part_pop75_                    1.026410   0.163266   6.287 1.30e-08 ***
sal_hor_net_homme_cs_4         0.219996   0.476989   0.461 0.645806
sal_hor_net_homme_cs_5        -1.638726   0.787080  -2.082 0.040311 *
ecart_tx_emp_f_h_55_64        -0.414604   0.145110  -2.857 0.005359 **
taux_lgmt_suroccupes_          0.428355   0.118711   3.608 0.000517 ***
log_hlm_phab_                  0.002437   0.001443   1.688 0.095005 .
part_emp_ess_                  0.371654   0.097831   3.799 0.000270 ***
taux_crea_etab_                0.161968   0.189270   0.856 0.394515
infrac_tx_violences_intrafam_  3.189899   0.705896   4.519 1.97e-05 ***
infrac_tx_cambriolages_        0.336698   0.174195   1.933 0.056543 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.133 on 86 degrees of freedom
Multiple R-squared:  0.9823, Adjusted R-squared:  0.9803
F-statistic: 477.9 on 10 and 86 DF,  p-value: < 2.2e-16
```

4.3.3 Qualité globale du modèle sans constante

Le modèle sans constante ajuste très fortement les données, avec un R^2 de 0,9823 et un R^2 ajusté de 0,9803. Cela signifie que, dans ce cadre particulier (sans constante), le modèle explique presque toute la variabilité du taux de pauvreté entre départements. Attention toutefois : pour un modèle sans constante, le R^2 n'a plus exactement la même interprétation que dans un modèle avec constante, et a

tendance à être artificiellement élevé, car la variance totale est calculée par rapport à zéro et non plus par rapport à la moyenne du taux de pauvreté.

L'erreur standard résiduelle est de 2,133 points de taux de pauvreté, ce qui signifie qu'en moyenne les écarts entre les valeurs observées et les valeurs prédites sont de l'ordre de 2 points. C'est un ajustement correct, mais légèrement moins précis que dans le modèle avec constante, où cette erreur était plus faible.

4.3.4 Interprétation des coefficients significatifs

Dans un modèle sans constante, chaque coefficient mesure directement la contribution moyenne de la variable au niveau du taux de pauvreté, si l'on considère que toutes les autres sont maintenues constantes à zéro (ce qui est une situation rarement réaliste, mais c'est l'interprétation formelle).

- **`part_pop75_`** ($\hat{\beta}_1 \approx 1,03$, $p < 0,001$) : une augmentation de 1 point de pourcentage de la part des 75 ans ou plus dans la population est associée à une hausse d'environ 1 point du taux de pauvreté, toutes choses égales par ailleurs. Ce coefficient est nettement plus élevé que dans le modèle avec constante, ce qui montre que l'effet de la structure démographique est davantage « chargé » lorsque l'on supprime l'ordonnée à l'origine.
- **`sal_hor_net_homme_cs_5`** ($\hat{\beta}_3 \approx -1,64$, $p \approx 0,04$) : le salaire horaire net des hommes en catégorie socioprofessionnelle 5 conserve un effet négatif significatif : plus ce salaire augmente, plus le taux de pauvreté diminue, même si l'intensité apparente de cet effet est moins forte qu'avec constante. Cela confirme le rôle protecteur des emplois qualifiés et bien rémunérés.
- **`ecart_tx_emp_f_h_55_64`** ($\hat{\beta}_4 \approx -0,41$, $p < 0,01$) : comme dans le modèle avec constante, le coefficient est négatif et significatif : un écart plus défavorable aux femmes est associé à un taux de pauvreté plus faible, ce qui reste contre-intuitif. Cela renvoie probablement à des effets de structure (départements où les hommes seniors sont très insérés sur le marché du travail) et à la colinéarité avec d'autres variables.
- **`taux_lgmt_suroccupes_`** ($\hat{\beta}_5 \approx 0,43$, $p < 0,001$) : une hausse d'un point de pourcentage du taux de logements sur-occupés est associée à une augmentation d'environ 0,4 point du taux de pauvreté. L'effet reste clairement positif et significatif, ce qui confirme le lien entre précarité résidentielle et pauvreté.

— **`part_emp_ess_`** ($\hat{\beta}_7 \approx 0,37$, $p < 0,001$) : dans ce modèle, la part des emplois relevant de l'économie sociale et solidaire devient significative et positive. Les départements où ce secteur pèse davantage ont, toutes choses égales par ailleurs, des taux de pauvreté plus élevés. Cela peut traduire le fait que l'économie sociale et solidaire est plus développée dans des zones déjà fragiles, où elle joue un rôle de « filet de sécurité » plutôt que de moteur de prospérité.

— **`infrac_tx_violences_intrafam_`** ($\hat{\beta}_9 \approx 3,19$, $p < 0,001$) : Ce coefficient est très élevé : ceci revient à dire qu'une augmentation d'une unité du taux de violences intrafamiliales est associée à une hausse d'environ 3 points du taux de pauvreté. Cela renforce le diagnostic d'un couplage très fort entre vulnérabilités sociales extrêmes et pauvreté.

— **`infrac_tx_cambriolages_`** ($\hat{\beta}_{10} \approx 0,34$, $p \approx 0,06$)

L'effet reste positif et proche de la significativité (environ 10 %), indiquant que les départements plus touchés par les cambriolages tendent à être aussi plus pauvres, même si la preuve statistique est ici un peu plus fragile que dans le modèle avec constante.

— **`log_hlm_phab_`** ($\hat{\beta} \approx 0,0024$, $p \approx 0,10$)

La part de logements HLM a un coefficient positif et marginalement significatif. Comme précédemment, ce résultat est à interpréter avec prudence, car le parc social peut à la fois concentrer des ménages pauvres et amortir certaines formes de précarité.

— **Variables non significatives : `sal_hor_net_homme_cs_4`, `taux_crea_etab_`**

Le salaire de la catégorie 4 et le taux de création d'établissements ne sont pas significatifs dans ce modèle. Leur effet net sur la pauvreté n'est donc pas clairement identifié ici.

4.3.5 Diagnostic spécifique au modèle sans constante

Le R^2 extrêmement élevé (0,98) et la statistique de Fisher très grande (477,9, $p < 2,2 \times 10^{-16}$) pourraient laisser penser que le modèle est « parfait ». En réalité, pour un modèle sans constante, ces indicateurs ont tendance à surévaluer la qualité d'ajustement, car la variance totale de la variable dépendante est calculée par rapport à zéro et non par rapport à sa moyenne. Plusieurs auteurs recommandent donc de ne pas interpréter trop littéralement le R^2 dans ce cas et de se concentrer davantage sur l'analyse des résidus et la comparaison avec le modèle avec constante.

L'erreur standard résiduelle plus élevée que dans le modèle avec constante montre d'ailleurs que, en termes d'écart moyen entre valeurs observées et prédites, la spécification sans constante fait légèrement moins bien. Cela va dans le sens de l'idée générale : sauf raison théorique forte de forcer le modèle à passer par l'origine, il est en général préférable de conserver une constante et de considérer le modèle sans constante comme un test de robustesse plutôt que comme la spécification principale.

Chapitre 5

Discussion

Les deux modèles de régression racontent la même histoire : la pauvreté territoriale n'est pas liée à un seul facteur, mais à un ensemble de dimensions qui se combinent. Les résultats mettent surtout en avant le rôle des revenus du travail, des conditions de logement et du contexte social local dans les différences de pauvreté entre départements.

D'abord, le marché du travail apparaît comme un élément central. Dans les deux modèles, quand le salaire horaire des emplois les plus qualifiés augmente, le taux de pauvreté diminue en moyenne : les départements où les emplois en haut de l'échelle sont mieux payés sont aussi ceux où la pauvreté est moins fréquente. À l'inverse, le résultat concernant l'écart de taux d'emploi entre femmes et hommes de 55 à 64 ans est plus difficile à interpréter : le coefficient est négatif, ce qui va à l'encontre de l'intuition, et renvoie sans doute à des spécificités locales ou à des corrélations avec d'autres variables plutôt qu'à un effet direct.

Les conditions de logement jouent également un rôle important. Dans les deux spécifications, une part plus élevée de logements sur-occupés est clairement associée à des taux de pauvreté plus forts : vivre dans un logement trop petit pour la taille du ménage apparaît ainsi comme un bon indicateur de précarité. En revanche, la part de logements sociaux a un effet moins net : elle n'est pas significative dans le modèle avec constante et n'apparaît que faiblement dans le modèle sans constante, ce qui reflète le double visage du parc HLM, à la fois concentrateur de ménages modestes et outil de protection via des loyers plus abordables.

Les indicateurs de fragilités sociales et d'insécurité (violences intrafamiliales, cambriolages) sont eux aussi fortement liés à la pauvreté. Les départements où ces phénomènes sont plus fréquents sont généralement ceux où la pauvreté est plus élevée, et cette association est encore renforcée dans le modèle sans constante. Cela ne signifie pas que la pauvreté « cause » directement ces violences ou ces délits, mais plutôt que ces difficultés se concentrent dans les mêmes Départements.

Enfin, le modèle sans constante modifie le poids apparent de certaines variables, comme la part d'emplois dans l'économie sociale et solidaire, qui devient significative alors qu'elle ne l'était pas avec constante. Cela tient surtout à la manière dont le modèle est écrit : en supprimant la constante, une partie de la variabilité est mécaniquement transférée vers les autres coefficients, sans que cela garantisse un lien plus solide sur le plan substantiel. Pour cette raison, il est plus raisonnable de considérer le modèle avec constante comme modèle principal, et le modèle sans constante comme un exercice de vérification et de comparaison plutôt que comme base de l'interprétation finale.

Chapitre 6

Conclusion

Ce travail, réalisé dans le cadre de notre projet académique, visait à mieux comprendre les déterminants départementaux du taux de pauvreté en France. L'analyse a combiné exploration et nettoyage des données, analyse en composantes principales pour résumer l'information, puis régression linéaire multiple afin de quantifier les liens entre pauvreté et caractéristiques des départements.

Les résultats mettent en évidence le rôle central des revenus du travail et de la qualité de l'emploi, ainsi que l'importance des conditions de logement pour expliquer les écarts de pauvreté territoriale. Ils montrent aussi que la pauvreté s'accompagne fréquemment de fragilités sociales plus larges, notamment à travers les associations observées avec les violences intrafamiliales et les cambriolages, ce qui souligne l'enchevêtrement des dimensions économiques, résidentielles et sociales de la vulnérabilité.

Plusieurs prolongements seraient utiles. Sur le plan des données, disposer de séries temporelles ou de découpages géographiques plus fins permettrait d'étudier plus précisément la dynamique et la localisation de la pauvreté. Sur le plan méthodologique, le recours à des modèles spatiaux ou à des approches explicitement orientées vers l'inférence causale offrirait une meilleure compréhension des mécanismes à l'œuvre. Enfin, l'articulation de ce type d'analyse statistique avec des démarches qualitatives ou de terrain apparaît essentielle pour transformer ces constats en pistes opérationnelles pour les politiques publiques de lutte contre la pauvreté et les inégalités territoriales

Bibliographie

[1] Insee. *L'essentiel sur la pauvreté*. 2025. URL : https://www.insee.fr/fr/statistiques/5759045.

[2] Denis Fougère, Francis Kramarz et Julien Pouget. Crime and Unemployment in France. Paris School of Economics, rapport de recherche, 2003. URL : http://piketty.pse.ens.fr/files/Fougereetal2003.pdf.

Annexe A

Codes R

Voici tous les extraits de code R utilisés pour la réalisation de ce travail sont présentés dans cette section, pour l'importation des données, le nettoyage, l'ACP et la régression linéaire.

A.1 Importation et Nettoyage de la base

```
#DEBUT DE PROGRAMME
# Chargement du fichier CSV
dat0 <- read.csv("ODD_DEP.csv", header = TRUE, sep = ";")
# Affichage
print(dat0)
View(dat0)
# Vérification des dimensions
dim(dat0)
# Sélection des colonnes d'intérêt: libgeo, variable, sous_champ, A2021
dat01 <- dat0[, c("libgeo", "variable", "sous_champ", "A2021")]
print(dat01)
View(dat01)
# Création d'une nouvelle colonne: "variable_souschamp"
dat01$variable_souschamp <- paste(dat01$variable, dat01$sous_champ, sep = "_")
print(dat01)
View(dat01)
# Conservation des colonnes utiles: libgeo, variable_souschamp, A2021
dat01 <- dat01[, c("libgeo", "variable_souschamp", "A2021")]
print(dat01)
dim(dat01)
View(dat01)
# Suppression des doublons
dat1 <- dat01[!duplicated(dat01), ]
View(dat1)
dim(dat1)
# Transformation des lignes de la colonne "variable_souschamp" en colonnes
dat1_tranpose <- reshape(
```

```
 dat1,
 timevar = "variable_souschamp",
 idvar = "libgeo",
 direction = "wide"
 )
 # Nettoyage des noms de colonnes (enlève le préfixe "A2021.")
 colnames(dat1_tranpose) <- gsub("^A2021\\.", "", colnames(dat1_tranpose))
 print(dat1_tranpose)
 View(dat1_tranpose)
 # Conservation des 96 premiers départements (métropole)
 dat1_tranpose <- dat1_tranpose[1:96, ]
 View(dat1_tranpose)
 # Vérification des colonnes avec des valeurs manquantes
 na_cols <- which(colSums(is.na(dat1_tranpose)) > 0)
 print(na_cols)
 # Suppression des colonnes avec des valeurs manquantes
 dat10 <- dat1_tranpose[, colSums(is.na(dat1_tranpose)) == 0]
 rownames(dat10)<-NULL
 View(dat10)
 dim(dat10)
 # Vérification de l'existence de valeurs manquantes
 which(colSums(is.na(dat10)) > 0)

 # FIN DU PROGRAMME
```

A.2 Sélection des variables & ACP

```
#Importation des librairies
library(FactoMineR)
library(factoextra)
library(ggplot2)
library(caret)

#Chargement du fichier
D<-read.csv(" ://mon_fichier.csv", header = TRUE, sep = ",", fileEncoding = "latin1")
print(D)
rownames(D)<-D$libgeo
D<-D[,!(names(D) %in%"libgeo")]
View(D)
colnames(D)

#Variable à garder "taux_pvt_total"
```

```
var_keep_name <- "taux_pvt_total"

#----------------------Algorithme-------------------------
#Extraire les variables numeric
D01 <- D[sapply(D, is.numeric)]
dim(D01)

#Retirer temporairement la variable à garder pour la sélection
D02 <- D01[, !(names(D01) %in% var_keep_name)]
dim(D02)

#Retirer les variables dont les variances respectives sont nulles
#-- Répérage des variables avec des variances nulles
var_zero <- which(apply(D02,2,var,na.rm = TRUE) == 0)
print(var_zero)
names(D02)[var_zero]

#-- Suppression de ces variables
D03 <- D02[,-var_zero, drop = FALSE]
dim(D03)

#Calcul de la matrice de corrélation sur l'ensemble des données
M_cor <- cor(D03,use = "pairwise.complete.obs")
View(M_cor)

#-- Verification des valeurs manquantes dans la matrice de correlation

any(is.na(M_cor))

#Retrait des variables fortement corrélées
to_remove <- findCorrelation(M_cor, cutoff = 0.8)
D03 <- D03[,-to_remove, drop = FALSE]

#Réintégration des variables à garder
D10 <- cbind(D03,D01[,var_keep_name, drop = FALSE])
dim(D10)
View(D10)

#ACP préliminaire sur toutes les variables pour verifier les
#varibles qui contribuent le plus à la formation du premier axe
#L'objectif est de récuperer les 29 variables qui contribuent le plus à la création
#de l'axe 1 et de les considéser comme variables à retenir

res_pre <- PCA(D10, scale.unit = TRUE, graph = FALSE)
```

```
#-- Extraction des contributions des variables à l'axe 1
contrib_axe1 <- res_pre$var$contrib[,1]

#--Variables ordonnées par importance de contribution
ordered_var <- names(sort(contrib_axe1, decreasing = TRUE))
print(ordered_var)

#-- Sélections des 29 premières variables
var_final <- c(var_keep_name, ordered_var[1:29])
D11 <- D10[,var_final]

# Etude descriptive des variables---------------

#1 - La moyenne de chacune des variables
moy <- apply(D11,2,mean)
print(moy)

#2 - Hétérogénéité des variables
#----boxplot
boxplot(D11)
#----Ecart type
ecart_t <- apply(D11,2,sd)
print(ecart_t)

#Histogrammes pour chaque variable numérique d'un data frame ---

# df = votre data frame

#Créer un dossier pour enregistrer les images
dir_name <- "..//Pictures/"
if (!dir.exists(dir_name)) {
dir.create(dir_name)
}

for (var in names(D11)) {

#Vérifier que la variable est numérique
if (!is.numeric(D11[[var]])) next

x <- D11[[var]]
x <- x[!is.na(x)]   # retirer les NA

#Nom de fichier
```

```
file_path <- file.path(dir_name, paste0("hist_", var, ".png"))

# Ouvrir le fichier image
png(file_path, width = 800, height = 600)

#Histogramme
hist(x,
col = "skyblue",
border = "white",
main = paste("Histogramme de", var),
xlab = var,
probability = TRUE)   # densité pour visualiser la dispersion

#Moyenne, médiane, min et max
abline(v = mean(x),   col = "blue",  lwd = 2)         # moyenne
abline(v = median(x), col = "red",   lwd = 2, lty = 2) # médiane
abline(v = range(x),  col = "green", lwd = 2, lty = 3) # min/max

#Légende
legend("topright",
legend = c("Moyenne", "Médiane", "Min/Max"),
col = c("blue", "red", "green"),
             lty = c(1, 2, 3),
             lwd = 2)

dev.off()
    }

#--------------ACP finale---------------

#1 - La matrice de correlation
x_corr <- cor(D11)
View(x_corr)

#2 - Projection des variables
res <- PCA(D11, scale.unit = TRUE, graph = FALSE)
plot(res, choix = "var")
#fviz_var

#Qualité de représentation
representation <- res$var$cos2
print(representation)

#Visualisation de l'inertie
```

```
fviz_eig(res, addlabels = TRUE, ylim = c(0, 50))

#Afficher les valeurs propres (Variance) ainsi que leurs pourcentage cumulé
res$eig

#Afficher les composantes principales: les coordonnées
c_principales <- res$ind$coord
print(c_principales)

#Gardons les deux premiers axes
print(c_principales[,1:2])

#Visualisation des variables (1,2)
fviz_pca_var(res,
                 col.var = "cos2",
                 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                 repel = TRUE)

#Visualisation des variables (1,3)
fviz_pca_var(res,
                 col.var = "cos2",
                 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                 repel = TRUE, axes=c(1,3))


#Visualisation des variables (2,3)
fviz_pca_var(res,
                 col.var = "cos2",
                 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                 repel = TRUE, axes=c(2,3))

#Visualisation des individus
fviz_pca_ind(res,
                 col.ind = "cos2",
                 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                 repel = TRUE)

#Projection simultannée des individus et des variables sur le plan (1,2)
fviz_pca_biplot(res,
                    repel=TRUE,
                    col.var = "#FC4E07",
                    col.ind = "#00AFBB")+
      ggtitle("Biplot ACP")+
      theme_minimal()
```

A.3 Estimation par MCO du modèle avec constante

```
df <- read.csv("data_restr.csv")
df
# Régression linéaire
mco <- lm(taux_pvt_total~ part_pop75_ + sal_hor_net_homme_cs_4 +
sal_hor_net_homme_cs_5 + ecart_tx_emp_f_h_55_64 + taux_lgmt_suroccupes_ +
log_hlm_phab_ + part_emp_ess_ + taux_crea_etab_
+ infrac_tx_violences_intrafam_ + infrac_tx_cambriolages_, data = df)

summary(mco)
```

A.4 Estimation par MCO du modèle sans constante

```
df <- read.csv("data_restr.csv")
df
# Régression linéaire
mco_sans_const <- lm(taux_pvt_total ~ part_pop75_ +  sal_hor_net_homme_cs_4 +
sal_hor_net_homme_cs_5 + ecart_tx_emp_f_h_55_64 + taux_lgmt_suroccupes_ +
log_hlm_phab_ + part_emp_ess_ +  taux_crea_etab_ + infrac_tx_violences_intrafam_ +
infrac_tx_cambriolages_ - 1, data = df)

summary(mco_sans_const)
```