Année universitaire **2024-2025**

PROJET ÉCONOMÉTRIQUE

Le lieu d'habitation a-t-il une incidence sur les opportunités d'emploi ou les salaires ?

Présentés par
Chiara Masi,
Lucas Paillas,
Nelson Reves.

Sous la direction de **Hélène Couprie**, professeur d'économétrie

Table des matières

1. Introduction

Dans un monde globalisé en constante évolution, avoir accès à des perspectives de carrière constitue un enjeu majeur pour la mobilité sociale et l'équité des opportunités. Cependant, cette accessibilité varie grandement selon les régions, laissant apparaître d'importantes inégalités selon la localisation. Glaeser (2011) met en lumière l'influence de la situation géographique sur les parcours professionnels, s'appuyant sur l'économie géographique, qui étudie l'interaction entre les dynamiques spatiales et le développement économique. On entend souvent que vivre dans une grande ville facilite l'accès à l'emploi et offre de meilleurs salaires, alors que ceux qui résident dans des zones rurales ou des quartiers prioritaires de la ville (QPV) font face à davantage de difficultés. Mais jusqu'à quel point cette perception est-elle justifiée ?

Plusieurs mécanismes expliquant ces disparités sont éclairés par les théories économiques classiques. D'un côté, les marchés du travail urbains se caractérisent par une forte concentration d'entreprises, une diversité de secteurs et des effets d'agglomération qui favorisent l'innovation et la compétitivité (Marshall, 1890 ; Krugman, 1991). Ces dynamiques facilitent l'accès à l'emploi, surtout pour les travailleurs qualifiés. En outre, la densité des infrastructures de transport en milieu urbain diminue les obstacles à la mobilité, rendant ainsi l'adéquation entre l'offre et la demande de travail plus efficace. En revanche, dans les zones rurales, le marché du travail est plus restreint. Le manque de transports, la spécialisation économique locale et le nombre limité d'entreprises rendent plus difficile l'accès à des emplois qualifiés et stables (Davezies, 2008).

L'impact du lieu de résidence ne se limite pas seulement à l'accès à l'emploi, mais inclut également les niveaux de salaire. Dans les régions où les emplois sont en forte concentration, la concurrence entre les entreprises peut faire grimper les salaires, tandis que dans les zones moins dynamiques, un marché de l'emploi restreint pousse les travailleurs à accepter des salaires plus bas (Moretti, 2013). De plus, la concentration des postes qualifiés dans les grandes métropoles contribue à creuser les disparités salariales, ce qui accroît la ségrégation géographique.

Dans ce contexte, il est intéressant de se demander :
Le lieu de résidence a-t-il un impact significatif sur les opportunités d'emploi et les niveaux de salaire en France ?

À travers une approche économétrique, ce projet a pour objectif d'analyser l'effet de la localisation résidentielle sur l'accès à l'emploi, et comment ces inégalités spatiales peuvent également se répercuter sur les différences de salaire. En milieu urbain, la compétition entre entreprises peut faire monter les salaires, tandis qu'en milieu rural ou en Quartiers Prioritaires de la Ville (QPV), le faible nombre d'employeurs peut obliger les travailleurs à accepter des postes moins bien rémunérés. Par ailleurs, la concentration d'emplois qualifiés dans les grandes villes peut accentuer ces écarts salariaux, renforçant ainsi la ségrégation spatiale.

2. Description des variables

A. Choix des variables

Variables socio-économiques individuelles :

- **SEXE** : Indique le genre de l'individu (1 = homme, 2 = femme). Cette variable est importante pour analyser d'éventuelles disparités de genre dans l'accès à l'emploi et de salaire.
- **AGE** : Âge de l'individu exprimé en années. Il permet de contrôler l'expérience potentielle sur le marché du travail, généralement liée à l'âge.
- **DIP7** : Niveau de diplôme atteint par l'individu. La formation académique est un déterminant essentiel de l'accès à l'emploi, de la nature de l'emploi obtenu et de la rémunération perçue.

Variables liées à l'emploi :

- **ACTEU** : Statut d'activité de l'individu. Codé comme suit :
  - 1 = Emploi (individus ayant un travail rémunéré)
  - 2 = Chômage (individus sans emploi mais en recherche active)
  - 3 = Inactivité (individus ne faisant pas partie de la population active, par exemple, étudiants, retraités).
    Cette variable est essentielle pour distinguer les individus actifs sur le marché du travail de ceux qui en sont exclus.
- **SALRED_Y** : Salaire net mensuel perçu par l'individu. Il s'agit de la variable cible pour étudier les inégalités salariales.

Variables géographiques :

- **STCOMM2020** : Type de commune dans laquelle réside l'individu. Les catégories peuvent inclure : ville-centre, banlieue, commune rurale, etc. Cette variable permet de prendre en compte les disparités géographiques qui peuvent influencer l'accès à l'emploi et le niveau de salaire.
- **QPV (Quartier Prioritaire de la Ville)** : Indique si l'individu réside dans un quartier prioritaire (1 = Oui, 2 = Non).
  Définition **:** Un quartier prioritaire est un territoire défini par les pouvoirs publics comme nécessitant une attention particulière en termes de développement économique, social et urbain.
  Importance **:** Les résidents des QPV peuvent être confrontés à des désavantages spécifiques (par exemple, un accès réduit à l'emploi ou des discriminations) qui peuvent influencer leur statut sur le marché du travail et leur niveau de rémunération.
- **REG** : Région administrative où réside l'individu. Permet de contrôler les effets liés aux différences régionales (politiques publiques, infrastructures économiques, coût de la vie, etc.). Nous allons en particulier utiliser REG = 11 qui est la région de l'Ile-de-France.

Variables supplémentaires pour statistiques descriptives :

- **PCS1Q (Professions et Catégories Socioprofessionnelles - Niveau Agrégé)** : Classification socioprofessionnelle de l'individu. Elle regroupe les métiers en grandes catégories homogènes, permettant de décrire la répartition sociale de la population. Utilisation **:** Bien que cette variable ne soit pas incluse dans les analyses explicatives, elle est utilisée pour les statistiques descriptives afin de mieux comprendre la composition socio-économique de l'échantillon étudié.

B. Choix Méthodologique

Pour analyser l'impact du lieu d'habitation sur les opportunités d'emploi et les salaires, nous avons exploité la base de données de l'Enquête Emploi en Continu (EEC). Cette enquête fournit des informations détaillées sur les caractéristiques individuelles des actifs, leurs conditions d'emploi et leur localisation.

Choix de la population étudiée :

Nous avons restreint notre analyse à :
✓ **Les individus en âge de travailler** (15-64 ans)
✓ **Les personnes actives** (actuellement en emploi ou au chômage)
exclusion des retraités et des étudiants.

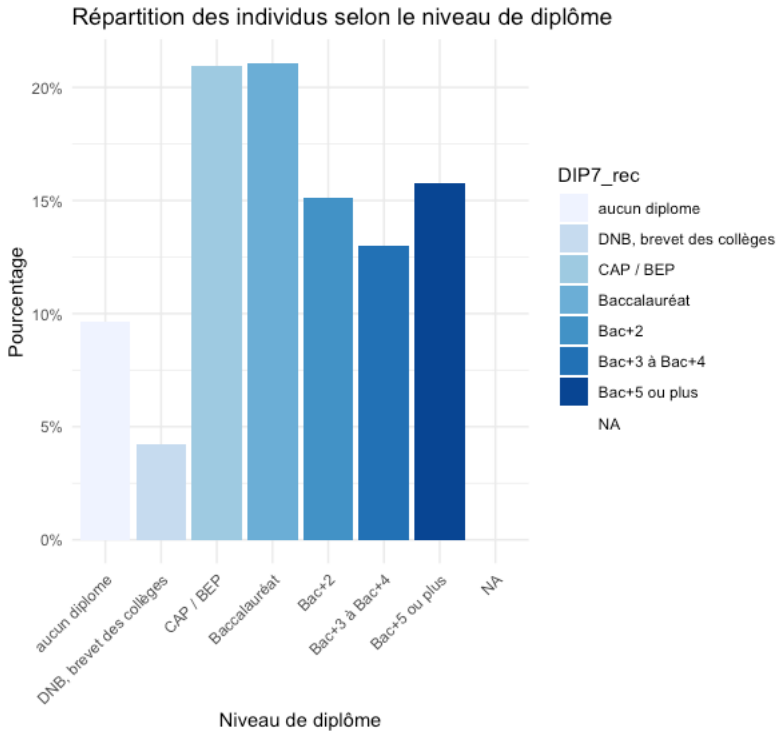
Cette restriction permet d'éviter un biais lié aux inactifs structurels et de mieux observer les effets du lieu d'habitation sur l'emploi et les salaires. Ce qui nous donne un échantillon de n = 44 444 personnes de 9 variables.

3. Statistiques Descriptives

A. Statistiques descriptives univariées

La répartition entre les sexes est équilibrée, avec **50.1% d'hommes** et **49.9% de femmes** dans l'échantillon (n= 44444). L'âge moyen des individus est de **42,5 ans**, avec une médiane à **43 ans**. La majorité des personnes ont entre **33 et 53 ans**.

La répartition des diplômes montre que la plupart des individus ont un niveau de formation intermédiaire. Par exemple, 21.1% de personnes ont un baccalauréat, tandis que 15.8% ont un diplôme de niveau bac+5 ou plus.

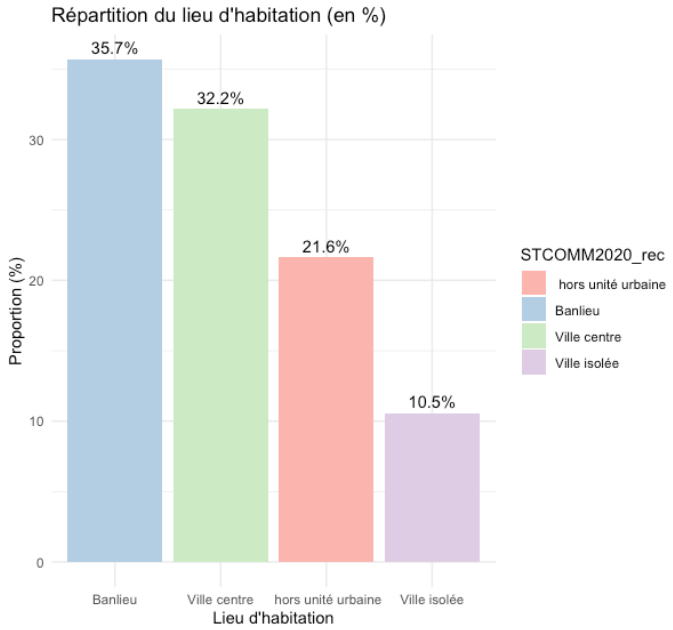


Graphique 1

Cela reflète la structure actuelle du marché du travail où une proportion importante d'individus accède à l'enseignement supérieur.

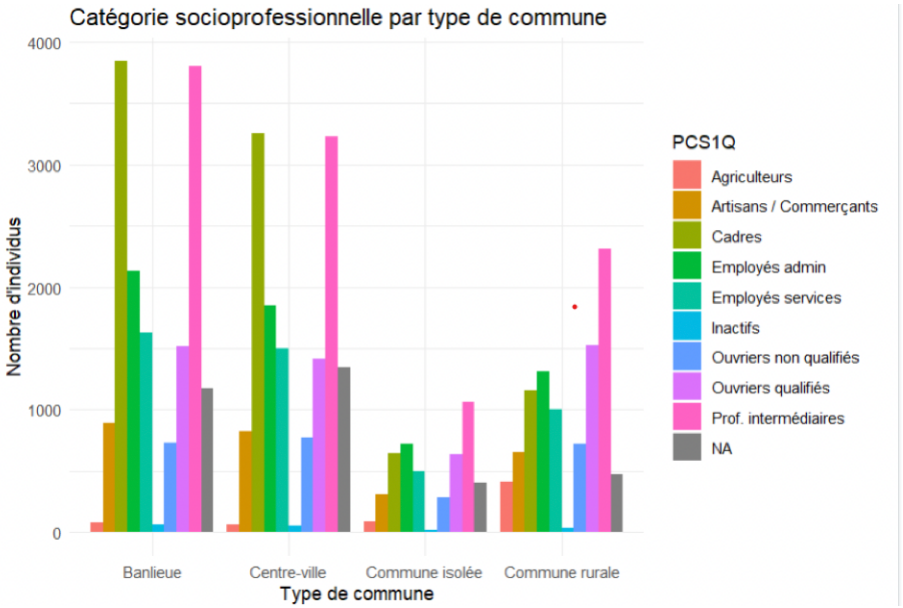
Les individus de l'échantillon sont répartis entre les zones urbaines et rurales :

- **B (Banlieue)** : 35.7%
- **C (Ville-centre)** : 32.2%
- **H (Hors unité urbaine)** : 21.6%
- **I (Ville isolée)** : 10.5%



Graphique 2

La grande majorité des individus de l'échantillon (90.53%) vivent en France métropolitaine contre 9.47% dans les départements d'outre-mer.



Graphique 3

Le *graphique 3* représente la répartition des catégories socioprofessionnelles (PCS1Q) selon le type de commune et met en évidence plusieurs tendances :

- Les cadres et professions sont majoritairement localisés dans les zones urbaines, en particulier en banlieue et en ville-centre.
- Une présence dominante des professions intermédiaires dans les communes isolées et rurales mais également les ouvriers qualifiés dans les communes rurales ainsi que les employés administratifs.

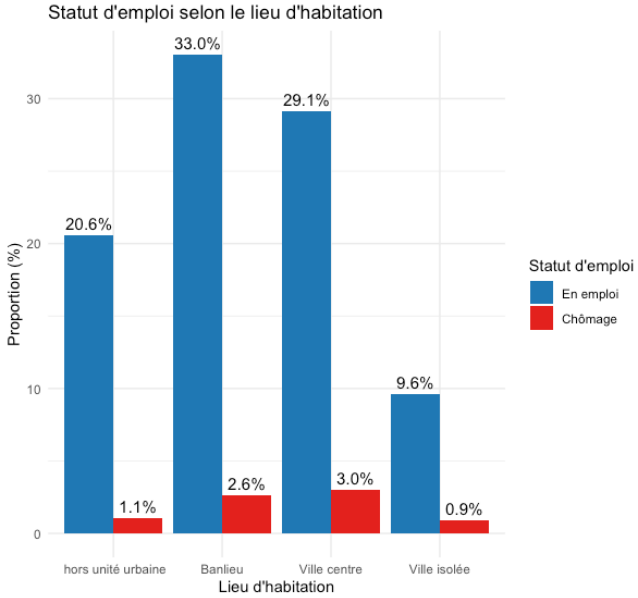
Ces différences reflètent une structure du marché du travail très hétérogène selon les territoires, qui contribue aux inégalités de revenus et à l'accès différencié à l'emploi.

6.3% des individus de l'échantillon résident en QPV (n = 44444), reflétant une sous-représentation des quartiers prioritaires dans les données. Nous utiliserons la pondération avec EXTRI pour vérifier la représentativité de l'échantillon.

B. Statistiques descriptives bivariées

Relation entre le lieu d'habitation et le statut d'emploi (ACTEU) :

- **Banlieue (B)** : 92.6% personnes en emploi contre 7.4 % au chômage.
- **Ville-centre (C)** : 90.6% en emploi, 9.4% au chômage.
- **Hors unité urbaine (H)** : 95.1% en emploi, 4.9% au chômage.
- **Ville isolée (I)** : 91.3% en emploi, 8.7% au chômage.

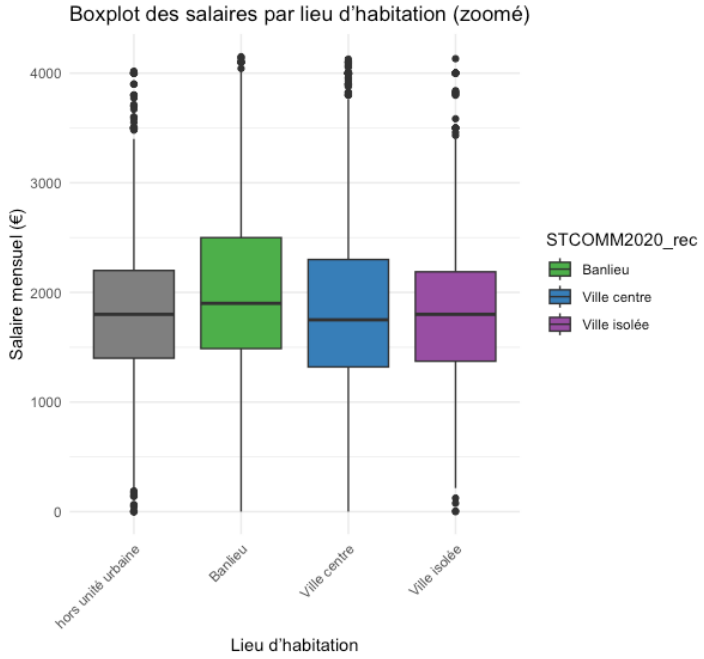


Graphique 4

Relation entre le lieu d'habitation et le salaire (SALRED_Y) :

- **Banlieue (B)** : Salaire médian de **2 000 euros**, avec une moyenne de **2 407 euros**.
- **Ville-centre (C)** : Salaire médian de **1 800 euros**, avec une moyenne de **2 324 euros**.
- **Hors unité urbaine (H)** : Salaire médian de **1 800 euros**, avec une moyenne de **1 894 euros**.

- **Ville isolée (I)** : Salaire médian de **1 800 euros**, avec une moyenne de **1 978 euros**.



Graphique 5

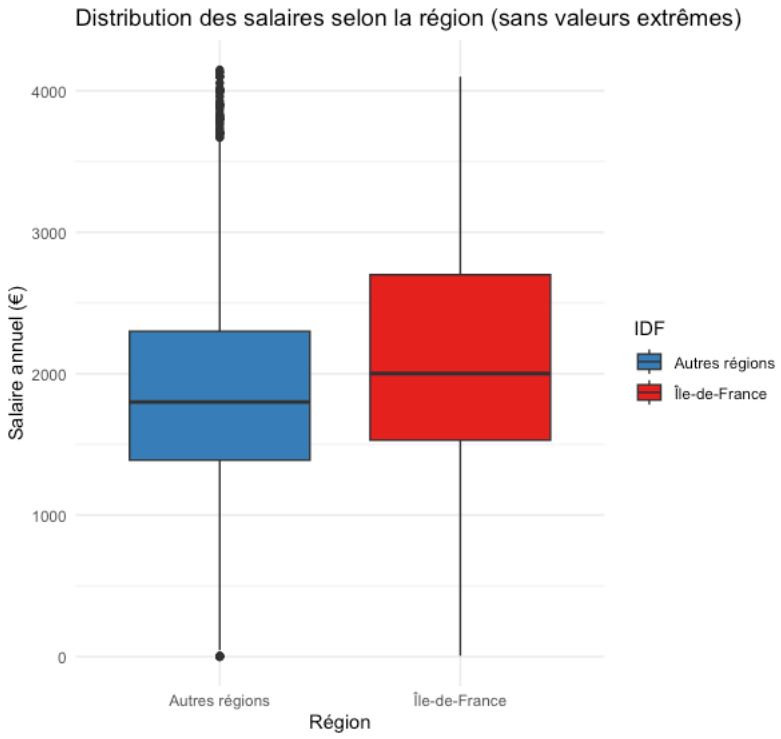
Comme l'indique l'INSEE : *"Les résidents des périphéries des grandes villes disposent des revenus les plus élevés. Dans les villes-centre, les revenus sont plus faibles et n'ont pas augmenté ces dernières années."*[1]. En effet les salaires en banlieue sont plus élevés qu'en ville-centre car on y trouve davantage d'emplois qualifiés (cadres, ingénieurs) dans des zones d'affaires et industrielles, contrairement aux centres-villes dominés par des emplois moins rémunérés (commerce, tourisme). De plus, les entreprises en banlieue offrent parfois des primes pour attirer des travailleurs face à une moindre accessibilité. Comme le montre le graphique 3, en banlieue, le taux d'emploi atteint 33 %, ce qui reflète une forte concentration d'actifs qualifiés travaillant dans des zones d'affaires et industrielles.

Nous isolons l'ile de France grâce à la variable REG = 11. Les personnes hors Île-de-France (IDF = 0) ont un salaire moyen de 2016 €. Les personnes en Île-de-France (IDF = 1) ont un salaire moyen de 3210 €. Cela confirme bien que vivre en Île-de-France est associé à des salaires plus élevés pour un taux d'emploi similaire !

	IDF	taux_emploi	salaire_moyen	salaire_median	ecart_type
1	Autres régions	92.31640	2016.385	1800	1117.548
2	Île-de-France	92.49787	3210.322	2200	13346.864

Tableau 1

[1] INSEE, "Des revenus élevés et en forte hausse en périphérie des pôles urbains", https://www.insee.fr/fr/statistiques/1285415



Graphique 6

Le rapport de Vie Publique[2] sur la mobilité géographique des travailleurs met en évidence que le lieu de résidence influence significativement l'accès à l'emploi et la progression salariale. Il montre que les individus mobiles géographiquement ont de meilleures opportunités professionnelles, notamment en changeant de région ou de bassin d'emploi. Toutefois, les contraintes financières et familiales limitent ces déplacements, ce qui contribue aux inégalités territoriales sur le marché du travail. C'est pour cela que nous allons inclure dans notre analyse les quartiers prioritaires de la politique de la ville.

La variable QPV étant représentée par 6% on utilise la variable de pondération EXTRI de l'INSEE. Les résultats montrent que l'échantillon brut : 6.3% en QPV (2 809/44 444) et la population réelle (pondérée) : 6.3% également (1.88M/30.08M). En France il y a plus de 5 millions de français habitant en QPV sur 68,29 millions de français ce qui correspond à environ 7%. La QPV est donc bien représentée.

Nous avons effectué une analyse croisée afin d'examiner s'il existe une probabilité plus élevée d'appartenir à un quartier prioritaire de la ville (QPV) en Île-de-France par rapport à la province.

	QPV	**Non QPV**
Autres régions	5.6%	94.4%
Île-de-France	9.6%	90.4%

Tableau 2

[2] Vie Publique, "La mobilité géographique des travailleurs", Rapport, 2017, disponible sur https://www.vie-publique.fr/files/rapport/pdf/174000288.pdf

On observe que 9,6 % des individus résidant en Île-de-France vivent dans un quartier prioritaire de la Ville (QPV), contre seulement 5,6 % dans les autres régions.
Cela montre une concentration plus forte des QPV en Île-de-France.

Nous avons effectué un tableau croisé pour mieux comprendre l'information de chaque variable.

	QPV	NON QPV
Hors unité urbaine	0%	100%
Banlieue	8%	92%
Ville-centre	8.5%	91.5%
Ville isolée	6.6%	93.4%

Tableau 3

On observe dans le tableau 3 que la concentration de QPV est majoritairement dans les centres-villes et les banlieues.

Le salaire moyen pondéré s'élève à :

- 1 406 € en Quartier Prioritaire de la Ville (QPV)
- 2 353 € hors QPV

Soit un écart absolu de 947 € (|1 406 - 2 353|) et un écart relatif de 40%. Cet écart massif suggère une pénalité salariale territoriale significative.

	QPV	Age_moyen	Pourcentage_femmes	salaire_moyen	Taux_chomage	Diplome_superieur
1	1	40.86686	48.38021	1406.364	19.330723	23.70951
2	2	42.60905	50.18134	2353.122	6.862015	45.23838

Tableau 4

Le rapport de l'Observatoire de l'Emploi en Occitanie[3] montre que les demandeurs d'emploi des Quartiers Prioritaires de la Ville (QPV) rencontrent de grandes difficultés d'accès à l'emploi. En plus d'un taux de chômage trois fois plus élevé que dans d'autres zones urbaines, les habitants des QPV ont un niveau de qualification généralement plus bas, avec des métiers moins diversifiés et des salaires moins élevés. Ces inégalités expliquent en partie les écarts salariaux notés entre les QPV et les autres zones.

[3]Observatoire de l'Emploi en Occitanie. (2021). Les inégalités d'accès à l'emploi dans les Quartiers Prioritaires de la Ville (QPV). https://www.observatoire-emploi-occitanie.fr/files_pdfs/MF1_202110_R.pdf

C. Conclusion de ces analyses descriptives et test de significativité

Le lieu d'habitation semble avoir une incidence sur les opportunités d'emploi et les salaires. Les zones urbaines (banlieue et ville-centre) offrent des emplois mieux rémunérés, mais avec un taux de chômage légèrement plus élevé. En revanche, les zones rurales (hors unité urbaine et ville isolée) présentent un taux de chômage plus faible, mais avec des emplois moins bien rémunérés.

Pour confirmer que les écarts observés ne sont pas dus au hasard on réalise un test du Chi^2 pour vérifier si la répartition emploi/chômage dépend significativement du lieu d'habitation.
Hypothèses :

- H0 = aucune liaison entre le lieu d'habitation et le statut d'emploi
- H1 : liaison significative entre les deux variables

On obtient les résultats suivants : Statistique de test (X-squared) : 173.11 Degrés de liberté (df) : 3 et p-value : < 2.2e-16 (soit < 0.00000000000000022).

La p-value est inférieure au seuil de significativité. On rejette l'hypothèse nulle. Il existe un lien significatif entre le lieu d'habitation et le statut d'emploi. Cela confirme notre hypothèse initiale : le lieu de résidence influence l'accès à l'emploi.

Nous avons utilisé le t test : le test d'égalités des moyennes pour comparer les salaires moyens entre l'ile de France et les autres régions.

Le t-test confirme que les salaires moyens de l'Île-de-France (3210 €) sont significativement plus élevés que dans les autres régions (2016 €), avec une différence moyenne de 1194 € (p-value = 0.004). L'intervalle de confiance à 95% indique que cet écart varie entre 385 € et 2003 €, reflétant un avantage salarial marqué pour les résidents d'Île-de-France.

Nous allons également faire un test de Student pour comparer les salaires moyens entre ceux qui résident dans un quartier prioritaire et ceux qui n'y résident pas.

Voici les hypothèses que nous allons tester :

H_0 : il n'y a pas de différence significative de salaire moyen entre les habitants de QPV et ceux qui n'y habitent pas

H_1 : il y a une différence significative de salaire moyen entre les deux groupes.

Ici nous n'utilisons pas la variable de pondération.

Interprétation des résultats obtenus :

- Valeur t : -2.2606
- Degrés de liberté : 6086
- Valeur p : 0.02382
- Intervalle de confiance à 95% : [-1516.34, -107.84]
- Moyennes :
    - Résidents en QPV (moyenne) : 1444.43
    - Résident hors QPV (moyenne) : 2256.52

Puisque la valeur p est inférieure à 0,05, vous pouvez rejeter l'hypothèse nulle (qui affirme que les moyennes des deux groupes sont égales). Cela suggère qu'il existe une différence statistiquement significative entre les deux groupes.

De plus, l'intervalle de confiance à 95% pour la différence des moyennes n'inclut pas 0, ce qui confirme encore que la différence entre les groupes est significative.

Conclusion : Il existe une différence significative entre les deux groupes concernant la variable SALRED_Y (selon QPV), le Groupe 2 ayant une moyenne plus élevée (2256.52) comparé au Groupe 1 (1444.43).

Nous avons réalisé un tableau de synthèse avec gtsummary de R sur le salaire moyen par lieu et diplôme :

Tableau descriptif par type de commune:

Variable	**Ensemble** (N = 44 444 [1])	**Banlieue** (N = 15 862 [1])	**Centre-ville** (N = 14 298 [1])	**Commune rurale** (N = 9 607 [1])	**Commune isolée** (N = 4 677 [1])
Salaire net mensuel (moyenne, écart-type)	**2 222 €** (5 649)	**2 407 €** (6 345)	**2 324 €** (7 514)	**1 894 €** (899)	**1 978 €** (1 104)
Données manquantes (salaire)	38 356 **(86%)**	13 492 **(85%)**	12 584 **(88%)**	8 130 **(85%)**	4 150 **(89%)**
Bac +5 et plus (Master, Doctorat)	7 009 **(16%)**	2 909 **(18%)**	2 887 **(20%)**	763 **(7.9%)**	450 **(9.6%)**
Bac+3/4 (licence, maîtrise ou équivalent)	5 766 **(13%)**	2 224 **(14%)**	2 005 **(14%)**	1 000 **(10%)**	537 **(11%)**
Bac +2 (BTS, DUT)	6 726 **(15%)**	2 510 **(16%)**	1 965 **(14%)**	1 544 **(16%)**	707 **(15%)**
Baccalauréat	9 368 **(21%)**	3 200 **(20%)**	2 830 **(20%)**	2 320 **(24%)**	1 018 **(22%)**
CAP, BEP ou équivalent	9 318 **(21%)**	2 885 **(18%)**	2 510 **(18%)**	2 809 **(29%)**	1 114 **(24%)**
BEPC, DNB, brevet des collèges	1 874 (**4.2%**)	650 **(4.1%)**	570 **(4.0%)**	454 **(4.7%)**	200 **(4.3%)**
Aucun diplôme	4 284 **(9.6%)**	1 449 **(9.1%)**	1 494 **(10%)**	702 **(7.3%)**	639 **(14%)**
Non réponse	99 **(0.2%)**	35 **(0.2%)**	37 **(0.3%)**	15 **(0.2%)**	12 **(0.3%)**
Homme	22 192 **(50%)**	7 854 **(50%)**	7 088 **(50%)**	4 923 **(51%)**	2 327 **(50%)**
Femme	22 252 **(50%)**	8 008 **(50%)**	7 210 **(50%)**	4 684 **(49%)**	2 350 **(50%)**

[1] **Mean (SD); n (%)**

Tableau 5

4. Analyse économétrique

On va étudier l'impact du lieu d'habitation (**STCOMM2020**) sur :

- Le salaire (**SALRED_Y**).
- Le statut d'emploi (**ACTEU** : emploi = 1, chômage = 2).

Nous allons utiliser une **régression linéaire multiple** pour analyser l'impact du lieu d'habitation sur le salaire en testant avec plusieurs variables pour ajuster la précision de notre modèle, et une **régression logistique** pour étudier l'impact sur le statut d'emploi (emploi ou chômage).

Modèle 1 : Impact du lieu d'habitation sur le salaire (SALRED_Y)

Forme fonctionnelle :

- **Variables explicatives** :
  - **STCOMM2020** : Lieu d'habitation (catégorielle : Banlieue, Ville-centre, Hors unité urbaine, Ville isolée).
  - **AGE** : Âge de l'individu.
  - **SEXE** : Sexe (1 = Homme, 2 = Femme).
  - **DIP7** : Niveau de diplôme (catégorielle).
  - **QPV**
- **Variable à expliquer** :
  - **SALRED_Y** : Salaire annuel.
- **Hypothèses** :
  - Linéarité : La relation entre les variables explicatives et le salaire est linéaire.
  - Homoscédasticité : La variance des erreurs est constante.
  - Absence de multi colinéarité : Les variables explicatives ne sont pas trop corrélées entre elles.
  - Normalité des résidus : Les erreurs suivent une distribution normale.

Nous avons commencé par mettre le salaire en logarithme. On va vérifier les corrélations entre les variables. Nous commençons par un tri croisé pour voir si QPV est corrélé à STCOMM :

	Résident QPV	Non résident QPV
Banlieue	1278	14582
Centre-ville	1222	13076
Hors urbain	0	9607
Commune isolé	309	4368

Tableau 6

On observe qu'il n'y a pas de personnes en QPV dans la zone « hors unité urbain » ce qui indique qu'il y a une forte imbrication entre les variables STCOMM2020 et QPV. Cela peut poser un problème de colinéarité si on inclue ces deux variables dans le même modèle, car elles peuvent être trop corrélées.

Nous avons vérifié la matrice de corrélation pour les variables numériques (AGE, SEXE, et SALRED_Y), nous avons des coefficients de corrélations très faibles ce qui signifie qu'il n'y a pas de problème de multi colinéarité immédiat entre elles. (Voir sortie de R en annexe).

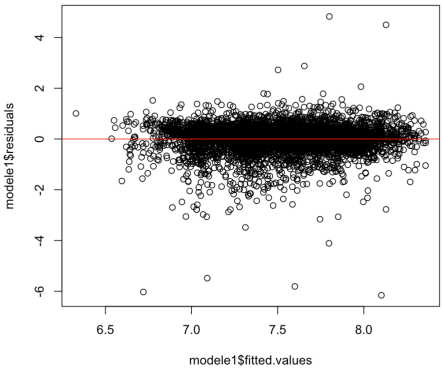
Nous décidons de faire ce modèle :
log_SALRED_Y ~ STCOMM2020 + AGE + SEXE + DIP7

Sortie de R en annexe. (Voir la partie sortie R en annexes)

Les coefficients sont les effets de chaque variable explicative sur le logarithme du salaire (log_SALRED_Y). Voici ce qu'on peut en dire :

- **Intercept (7.988)** : Le salaire annuel de référence pour un individu avec les caractéristiques de la catégorie de référence de STCOMM2020, âgé de 0, de sexe "homme", et ayant un niveau de diplôme de référence. Cet intercept est significatif avec un très petit p-value (< 2e-16).
- **STCOMM2020C** (-0.061) : Si l'individu vit dans un "Centre urbain" (par rapport à la catégorie de référence), cela réduit son salaire annuel de **6.1%**, toutes choses égales par ailleurs. Ce coefficient est significatif avec un p-value de 0.000261.
- **STCOMM2020H** (-0.066) : Si l'individu vit dans une zone **Hors unité urbaine** (par rapport à la catégorie de référence), son salaire est réduit de **6.6%**. Ce coefficient est également significatif avec un p-value de 0.000195.
- **STCOMM2020I** (-0.014) : Si l'individu vit dans une **Ville isolée**, l'effet est bien plus faible (-1.4%) et non significatif (p-value de 0.58). Cela suggère que cette catégorie n'a pas un effet important ou que les données sont moins représentatives pour cette catégorie.
- **AGE** (0.0127) : Chaque année d'âge supplémentaire augmente le salaire de **1.27%**. Cela est statistiquement significatif (p-value < 2e-16).
- **SEXE** (-0.298) : Les femmes ont un salaire moyen inférieur de **29.8%** par rapport aux hommes, toutes choses égales par ailleurs. Ce coefficient est fortement significatif (p-value < 2e-16).
- **DIP7** (-0.148) : Un niveau de diplôme supplémentaire réduit le salaire de **14.8%** en moyenne, ce qui est également très significatif.

Les variables STCOMM2020C, STCOMM2020H, AGE, SEXE et DIP7 sont toutes significatives (p-value < 0.05), sauf pour la variable STCOMM2020I qui n'est pas significative (p-value = 0.58). Cela suggère qu'il n'y a pas de lien fort entre vivre dans une "ville isolée" et le salaire dans le modèle.
Nous vérifions si notre modèle respecte bien les hypothèses. Voici le graphique des résidus :



Graphique 7

Les erreurs suivent une distribution normale.

Test multivariés : pour ce modèle c'est le test de FISHER qui permet de tester simultanément plusieurs hypothèses, La fonction waldtest du package lmtest de R permet d'effectuer ce test.

- **Modèle 1** : Le modèle complet avec toutes les variables explicatives (STCOMM2020, AGE, SEXE, DIP7).
- **Modèle 2** : Le modèle contraint, qui exclut la variable SEXE.

Interprétation des résultats : (sortie de R et utilisation d'IA en annexe.)

1. **Comparaison des degrés de liberté (Df)** :
   Le modèle contraint (Modèle 2) a un degré de liberté de 6074, tandis que le modèle complet (Modèle 1) a un degré de liberté de 6073. Cela montre que le modèle contraint a une variable de moins (celle de SEXE).
2. **F-statistic (480.13)** :
   La statistique de test F est très élevée (480.13), ce qui indique que la différence entre les deux modèles est significative.
3. **P-value (< 2.2e-16)** :
   La p-valeur est extrêmement faible (inférieure à 0.05), ce qui indique que nous pouvons **rejeter l'hypothèse nulle**. L'hypothèse nulle dans ce cas est que la variable SEXE n'a aucun effet sur le salaire (log_SALRED_Y). Puisque la p-valeur est très petite, cela suggère que **SEXE est une variable significative** dans le modèle et que son exclusion du modèle contraint améliore significativement la qualité de la prédiction.

Conclusion :

Le test de Wald montre que la variable SEXE a un impact significatif sur le salaire et son exclusion dans le modèle contraint n'est pas justifiée. Ainsi, le modèle complet (avec la variable SEXE) est préférable à celui où cette variable est exclue. On conclut que la variable SEXE joue un rôle important dans la prédiction du salaire annuel (log_SALRED_Y).

Modèle 2 : Impact du lieu d'habitation sur le statut d'emploi (ACTEU)

- **Variables explicatives** :
  STCOMM
  AGE
  SEXE
  DIP7
- **Variable à expliquer** :
  - **ACTEU** : Statut d'emploi (1 = Emploi, 2 = Chômage).
- **Hypothèses** :
  - Absence de multi colinéarité.

Nous devons réaliser une régression logistique car ACTEU est une variable catégorielle. Nous avons donc recodé ACTEU en variable binaire (1 = en emploi, 2 = au chômage) et nous avons transformé certaines variables en facteur.

Interprétation des résultats : (sortie de R et utilisation d'IA en annexe.)
Les coefficients représentent l'impact de chaque variable explicative sur la probabilité d'être en emploi, exprimé en **l**og-odds. Voici l'analyse détaillée :

Intercept (1.762) :

- **Valeur de référence** : Probabilité d'emploi pour un **homme (SEXE1)**, âgé de **0 ans**, habitant **hors unité urbaine (H)**, et diplômé **Bac+5 (DIP71)**.
- **Significativité** : Très significatif (p < 2e-16).
- **Note** : L'âge à 0 ans n'est pas réaliste, mais l'intercept sert de baseline pour les autres coefficients.

Lieu d'habitation (STCOMM2020) :
Référence : Hors unité urbaine (H).

- **STCOMM2020C (-0.220)** :
  - Vivre en **ville centre** réduit les **log-odds d'emploi** de **0.220** par rapport à H.
  - **Traduction en probabilité** : Diminution significative des chances d'emploi (p < 2e-16).
- **STCOMM2020I (-0.045)** :
  - Effet négligeable et **non significatif** (p = 0.47) pour les **villes isolées**.
- **STCOMM2020H (référence)** :
  - Les zones rurales/hors UU sont associées aux **meilleures chances d'emploi** (coefficient de référence).

Âge (AGE) (0.036)

- **Effet positif** : Chaque année supplémentaire augmente les log-odds d'emploi de 0.036.
- **Significativité** : Très significatif (p < 2e-16).
- **Interprétation** : L'expérience accumulée avec l'âge améliore l'accès à l'emploi.

Sexe (SEXE2) (-0.090)

- **Référence : Homme (SEXE1)**.
- **Effet négatif** : Les femmes ont des log-odds d'emploi inférieurs de 0.090 à ceux des hommes.
- **Significativité** : Significatif (p = 0.014).
- **Traduction** : Inégalités de genre persistantes sur le marché du travail.

Niveau de diplôme (DIP7)
Référence : Bac+5 (DIP71).

Diplôme	Coefficient	Interprétation	Significativité
DIP72 (Bac+3/+4)	-0.090	Effet légèrement négatif mais **non significatif** (pas de différence avec Bac+5).	p = 0.29
DIP73 (Bac+2)	-0.280	Diminution **significative** des chances d'emploi vs. Bac+5.	*** (p < 0.001)
DIP74 (Bac général)	-0.706	Effet **très négatif** : chances d'emploi bien plus faibles.	*** (p < 2e-16)
DIP75 (CAP/BEP)	-0.911	Réduction **importante** des chances d'emploi.	*** (p < 2e-16)
DIP76 (BEPC)	-1.186	Effet **extrêmement négatif** (niveau inférieur au CAP).	*** (p < 2e-16)
DIP77 (Aucun diplôme)	-1.656	**Désavantage maximal** : chances d'emploi très réduites.	*** (p < 2e-16)

Conclusion :
Ce modèle montre que le lieu de résidence influence significativement le statut d'emploi, avec un désavantage notable pour les habitants des villes-centres par rapport aux zones rurales. L'âge a un effet positif sur l'emploi, indiquant que l'expérience accumulée augmente les chances d'être employé. Les femmes ont une probabilité d'emploi inférieure à celle des hommes, révélant des inégalités de genre persistantes. Le niveau de diplôme est également un facteur déterminant : plus il est élevé, plus les chances d'emploi augmentent. Cependant, l'analyse ne prend pas en compte d'éventuelles interactions entre les variables ou des facteurs contextuels comme la mobilité géographique. Des analyses complémentaires utilisant des modèles avancés permettraient de mieux comprendre ces dynamiques.

5. Conclusion du projet et pistes d'améliorations

Cette section présente un récapitulatif des principales conclusions tirées du projet d'analyse. Après avoir réalisé une analyse descriptive des données et des analyses bivariées, il a été possible de formuler des observations sur les relations entre les variables socio-économiques, géographiques et l'impact de ces dernières sur le salaire et le statut d'emploi.

Les modèles économétriques mis en place ont permis de mettre en évidence des relations significatives entre certaines caractéristiques, comme le lieu d'habitation, l'âge, le sexe et le niveau de diplôme, et les variables dépendantes telles que le salaire et le statut d'emploi. Les résultats des régressions ont confirmé l'importance de variables géographiques dans la détermination du salaire, notamment les différences observées entre les différentes zones urbaines. En outre, les analyses ont souligné l'impact significatif de l'âge, du sexe et du diplôme sur le salaire, ainsi que des résultats intéressants concernant le statut d'emploi, qui a révélé des disparités en fonction du lieu de résidence.

Enfin, l'ensemble des analyses fournissent des éléments précieux pour comprendre les facteurs socio-économiques et géographiques influençant les variables d'intérêt. Les résultats sont susceptibles d'informer des politiques publiques visant à réduire les inégalités salariales et améliorer les conditions d'emploi dans différentes régions.

Cependant, certaines limites méthodologiques peuvent affecter la robustesse des résultats obtenus. Par exemple, des variables explicatives potentiellement importantes n'ont pas été intégrées dans l'analyse, comme des indicateurs socio-économiques complémentaires (CDD, CDI, heures travaillées etc.) ou des variables environnementales (proximité des transports). De plus, une actualisation des données permettrait de mieux prendre en compte les évolutions récentes des dynamiques régionales.

Sur le plan méthodologique, l'utilisation de modèles économétriques plus avancés, tels que les modèles non linéaires, les modèles de panel ou des techniques de machine learning, pourrait améliorer la précision des résultats. De plus, une vérification rigoureuse des hypothèses sous-jacentes à la régression linéaire, ainsi que l'exploration d'éventuelles interactions entre variables, contribuerait à affiner l'analyse.

6. Annexes

Références bibliographiques :

- **Glaeser, E. L. (2011).** *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier.* Penguin Press.
- **Marshall, A. (1890).** *Principles of Economics.* Macmillan and Co.
- **Krugman, P. (1991).** *Increasing Returns and Economic Geography.* Journal of Political Economy, 99(3), 483-499.
- **Davezies, L. (2008).** *La République et ses territoires : La circulation invisible des richesses.* Seuil.
- **Moretti, E. (2013).** *The New Geography of Jobs.* Mariner Books.

Sources des données :

- **INSEE (Institut National de la Statistique et des Études Économiques).** *Enquête Emploi en Continu (EEC).* Disponible sur : www.insee.fr

Sources de documentation utilisée :

- Cours d'Hélène Couprie
- Cours d'Emmanuel Flachaire (professeur à l'AMSE) : https://sites.google.com/site/emmanuelflachaire/cours
- Documentation pour améliorer les graphiques : https://ggplot2.tidyverse.org/reference/
- Dictionnaire des variables de l'enquête EEC 2022 et 2023
- https://www.scribbr.fr/category/normes-apa/ pour les normes APA et page de garde universitaire.

Sortie R :

Matrice de corrélation :

```
> cor(df[, c("AGE", "SEXE", "SALRED_Y")], use = "complete.obs")
                AGE        SEXE    SALRED_Y
AGE      1.00000000  0.04797586  0.03159912
SEXE     0.04797586  1.00000000 -0.05864530
SALRED_Y 0.03159912 -0.05864530  1.00000000
```

Modèle de régression 1 :

```
Call:
lm(formula = log_SALRED_Y ~ STCOMM2020 + AGE + SEXE + DIP7, data = df_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1560 -0.1944  0.0568  0.2821  4.8262

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.9881209  0.0345672 231.090  < 2e-16 ***
STCOMM2020C -0.0612171  0.0167567  -3.653 0.000261 ***
STCOMM2020H -0.0656746  0.0176154  -3.728 0.000195 ***
STCOMM2020I -0.0141272  0.0255686  -0.553 0.580611
AGE          0.0127518  0.0005666  22.508  < 2e-16 ***
SEXE        -0.2976896  0.0135858 -21.912  < 2e-16 ***
DIP7        -0.1482234  0.0038919 -38.085  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5275 on 6073 degrees of freedom
  (38356 observations deleted due to missingness)
Multiple R-squared:  0.2629,    Adjusted R-squared:  0.2622
F-statistic:   361 on 6 and 6073 DF,  p-value: < 2.2e-16
```

Test de Fisher pour le modèle 1 :

```
> waldtest(modele_non_contraint, modele_contraint)
Wald test

Model 1: log_SALRED_Y ~ STCOMM2020 + AGE + SEXE + DIP7
Model 2: log_SALRED_Y ~ STCOMM2020 + AGE + DIP7
  Res.Df Df      F    Pr(>F)
1   6073
2   6074 -1 480.13 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modèle de régression 2 :

```
Call:
glm(formula = ACTEU ~ STCOMM2020 + AGE + SEXE + DIP7, family = binomial,
    data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.761708   0.085337  20.644  < 2e-16 ***
STCOMM2020C -0.220044   0.042538  -5.173 2.31e-07 ***
STCOMM2020H  0.497673   0.057134   8.711  < 2e-16 ***
STCOMM2020I -0.044636   0.061468  -0.726 0.467740
AGE          0.036196   0.001501  24.107  < 2e-16 ***
SEXE2       -0.090202   0.036794  -2.452 0.014224 *
DIP72       -0.089541   0.085461  -1.048 0.294762
DIP73       -0.280018   0.080948  -3.459 0.000542 ***
DIP74       -0.705778   0.069116 -10.212  < 2e-16 ***
DIP75       -0.911403   0.070711 -12.889  < 2e-16 ***
DIP76       -1.186211   0.091628 -12.946  < 2e-16 ***
DIP77       -1.655903   0.072773 -22.754  < 2e-16 ***
DIP79        0.315546   0.515933   0.612 0.540801
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24012  on 44443  degrees of freedom
Residual deviance: 22540  on 44431  degrees of freedom
AIC: 22566

Number of Fisher Scoring iterations: 6
```

Utilisation de l’Intelligence artificielle :

Outil utilisé : ChatGPT

Dans le cadre de la rédaction du document, une partie du travail a consisté à reformuler l'introduction ainsi que certaines phrases spécifiques afin de rendre le texte plus clair et fluide. Pour ce faire, l'intelligence artificielle, en l'occurrence ChatGPT, a été utilisée.

ChatGPT a permis de reformuler des passages du texte en modifiant la structure des phrases, en améliorant la cohérence et en adaptant le ton du texte tout en respectant le contenu initial. L'outil a été utilisé pour améliorer l'expression écrite et optimiser la compréhension des idées principales.

ChatGPT a également été utilisé pour la partie analyse économétrique. Dans le modèle 1 Il nous a aidés à la partie tests multivariés en complément du cours de Hélène Couprie pour utiliser la fonction waldtest afin de nous aider à choisir notre modèle contraint et notre modèle non contraint et de nous expliquer clairement l’utilité de cette fonction, il nous a également assistés dans l'interprétation des résultats obtenus notamment pour le log-odds.

Dans le modèle 2, il nous a aidé à comprendre la régression logistique en nous expliquant qu'il fallait recoder ACTEU en binaire et transformer certaines variables en facteurs. Ceci nous a permis de mener à bien la régression, malgré les difficultés rencontrées.

Résumé des actions réalisées avec ChatGPT :

- Reformulation de l'introduction du document.
- Modification et clarification de certaines phrases pour faciliter la lecture.
- Assistance dans le choix des modèles pour le test de Wald (modèle contraint vs. Modèle non contraint).
- Interprétation des résultats du test de Wald.

L'utilisation de l'intelligence artificielle dans ce contexte visait à améliorer l'efficacité de la rédaction et aider dans nos points de blocage.

Reformulation de l'introduction du document

Requête : "ChatGPT, peux-tu reformuler l'introduction de notre document pour qu'elle soit plus claire et fluide tout en conservant les idées principales. (Insertion de notre brouillon d'introduction)"

Réponse de ChatGPT : "Je vais reformuler l'introduction en améliorant la cohérence et le style d'écriture, tout en respectant le contenu initial. Voici une proposition reformulée : [...] " introduction en page 3

Assistance pour les tests multivariés (Test de Wald)

Requête : "ChatGPT, pourrais-tu m'expliquer l'utilité de la fonction waldtest en économétrie, et comment l'appliquer pour choisir un modèle contraint et un modèle non contraint ?"

Réponse de ChatGPT : "La fonction `waldtest` est utilisée pour tester des hypothèses sur les coefficients d'un modèle économétrique. Cela permet de comparer un modèle contraint (réduit) avec un modèle non contraint (complet) pour vérifier si les restrictions imposées sont statistiquement significatives. Voici comment l'utiliser : [...]"

Interprétation des résultats du Test de Wald

Requête : "Peux-tu m'aider à interpréter les résultats que j'ai obtenus pour le waldtest ?"

Réponse de ChatGPT : "Les résultats montrent que [...] Cela signifie que le modèle contraint est [accepté/rejeté] par rapport au modèle non contraint, ce qui suggère que [...]"

Assistance pour la régression logistique (Modèle 2)

Requête : "ChatGPT, pourrais-tu m'expliquer comment effectuer une régression logistique ? ma régression logistique est : ACTEU ~ STCOMM2020 + AGE + SEXE + DIP7"

Réponse de ChatGPT : "Pour réaliser une régression logistique, il est nécessaire que la variable dépendante soit binaire. Si ACTEU est une variable qualitative à plusieurs modalités, il faut la recoder en binaire (par exemple, 0 pour non, 1 pour oui). Les variables qualitatives doivent être converties en facteurs pour qu'elles soient prises en compte correctement par le modèle. Voici comment procéder en R : [...]"

Script R :

```
library(readr)
library(dplyr)
library(questionr)
library(ggplot2)
library(gtsummary)
library(tidyverse)

indiv231 <- read_delim("Desktop/miashs/MIASHS
L3/S6/PROJET (ÉCONOMÉTRIE)/indiv231.csv",
                       delim = ";", escape_double = FALSE, trim_ws
= TRUE)

------------------------------------------------------------------------
# Sélection de notre échantillon et première stats.

dataselected <-indiv231 %>%
  filter(AGE >= 15 & AGE <= 64,  # Âge de 15 à 64 ans
         ACTEU %in% c(1, 2))  # Emploi ou chômage

df <- dataselected[,c(
  "STCOMM2020", "REG", "DENS2022", "QPV",
"METRODOM",
  "ACTEU", "STATUT", "SALTYP", "PCS1Q",
  "SALRED_Y","AGE", "SEXE", "DIP7","EXTRI"
)]

summary(df)

summary(df$EXTRI)
------------------------------------------------------------------------

# Statistiques descriptives univariées
table(df$SEXE)
summary(df$AGE)
------------------------------
table(df$DIP7)
#recodage de DIP7 :
df <- df %>%
  mutate(DIP7_rec = case_when(
    DIP7 == 1 ~ "Bac+5 ou plus",  # Niveau 1
    DIP7 == 2 ~ "Bac+3 à Bac+4",  # Niveau 2
    DIP7 == 3 ~ "Bac+2",          # Niveau 3
    DIP7 == 4 ~ "Baccalauréat",   # Niveau 4
    DIP7 == 5 ~ "CAP / BEP",      # Niveau 5
    DIP7 == 6 ~ "DNB, brevet des collèges",  # Niveau 6
    DIP7 == 7 ~ "aucun diplome",  # Niveau 6
    DIP7 == 9 ~ "non réponse",  # Niveau 6
    TRUE ~ as.character(DIP7)     # Autres niveaux
  ))

# Définition de l'ordre des niveaux de diplôme
df$DIP7_rec <- factor(df$DIP7_rec, levels = c(
  "aucun diplome",
  "DNB, brevet des collèges",
  "CAP / BEP",
  "Baccalauréat",
  "Bac+2",
  "Bac+3 à Bac+4",
  "Bac+5 ou plus"
))

# Création du diagramme en pourcentages
ggplot(df, aes(x = DIP7_rec, fill = DIP7_rec)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(labels = scales::percent_format(accuracy =
1)) +  # Convertit en %
  scale_fill_brewer(palette = "Blues") +
  labs(
    title = "Répartition des individus selon le niveau de diplôme",
    x = "Niveau de diplôme",
    y = "Pourcentage"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Calculer les proportions
df %>%
  group_by(DIP7_rec) %>%
  summarise(count = n()) %>%
  mutate(pourcentage = (count / sum(count)) * 100) %>%

----------------------------------

table(df$QPV)
----------------------------------
table(df$STCOMM2020)
#recodage
df <- df %>%
  mutate(STCOMM2020_rec = case_when(
    STCOMM2020 == "H" ~ " hors unité urbaine",
    STCOMM2020 == "C" ~ "Ville centre",
    STCOMM2020 == "B" ~ "Banlieu",
    STCOMM2020 == "I"~ "Ville isolée",
    TRUE ~ as.character(STCOMM2020) # Garde la valeur
originale si elle ne correspond pas
  ))
# Calcul des pourcentages
df %>%
  group_by(STCOMM2020_rec) %>%
  summarise(count = n()) %>%
  mutate(pourcentage = (count / sum(count)) * 100) %>%

  # Création du graphique
  ggplot(aes(x = reorder(STCOMM2020_rec, -pourcentage), y =
pourcentage, fill = STCOMM2020_rec)) +
  geom_bar(stat = "identity") +

  # Labels et mise en forme
  labs(
    title = "Répartition du lieu d'habitation (en %)",
    x = "Lieu d'habitation",
    y = "Proportion (%)"
  ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Pastel1") +
```

```
  geom_text(aes(label = sprintf("%.1f%%", pourcentage)), vjust = -
0.5)
--------------------------------
table(df$METRODOM)
unique(df$REG)
----------------------------------
# Création de la variable binaire IDF (Île-de-France vs reste de la
France)
df <- df %>%
  mutate(IDF = ifelse(REG == 11, "Île-de-France", "Autres
régions"))

# Suppression des valeurs extrêmes du salaire (Zoom)
Q1 <- quantile(df$SALRED_Y, 0.25, na.rm = TRUE)
Q3 <- quantile(df$SALRED_Y, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value

df_filtered <- df %>%
  filter(SALRED_Y >= lower_bound & SALRED_Y <=
upper_bound)  # Exclusion des outliers

# Comparaison des salaires selon l'Île-de-France (moyenne et
médiane)
stats_emploi_salaire <- df %>%
  group_by(IDF) %>%
  summarise(
    taux_emploi = mean(ACTEU == "1", na.rm = TRUE) * 100,  #
Proportion en emploi
    salaire_moyen = mean(SALRED_Y, na.rm = TRUE),
    salaire_median = median(SALRED_Y, na.rm = TRUE),
    ecart_type = sd(SALRED_Y, na.rm = TRUE)
  )

#Boxplot des salaires par lieu d'habitation.
ggplot(df_filtered, aes(x = STCOMM2020_rec, y = SALRED_Y,
fill = STCOMM2020_rec)) +
  geom_boxplot() +
  scale_fill_manual(values = c("hors unité urbaine" = "#E41A1C",
                               "Banlieu" = "#4DAF4A",
                               "Ville centre" = "#377EB8",
                               "Ville isolée" = "#984EA3")) +
  labs(title = "Boxplot des salaires par lieu d'habitation (zoomé)",
       x = "Lieu d'habitation",
       y = "Salaire mensuel (€)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#  Boxplot : Disparité des salaires selon IDF / Autres régions
ggplot(df_filtered, aes(x = IDF, y = SALRED_Y, fill = IDF)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Île-de-France" = "#E41A1C",
"Autres régions" = "#377EB8")) +
  labs(title = "Distribution des salaires selon la région (sans valeurs
extrêmes)",
       x = "Région",
       y = "Salaire annuel (€)") +
  theme_minimal()

-----------------------
# Fréquence de QPV
table(df$QPV)

# Statistiques descriptives par QPV
stat_qpv <- df %>%
  group_by(QPV) %>%
  summarise(
    Age_moyen = mean(AGE, na.rm = TRUE),
    Pourcentage_femmes = mean(SEXE == 2, na.rm = TRUE) *
100,
    salaire_moyen = weighted.mean(SALRED_Y, w = EXTRI,
na.rm = TRUE),
    Taux_chomage = mean(ACTEU == 2, na.rm = TRUE) * 100,
    Diplome_superieur = mean(DIP7 %in% c(1, 2, 3), na.rm =
TRUE) * 100
  )

# Calculer la proportion réelle des QPV avec les poids
df %>%
  group_by(QPV) %>%
  summarise(
    n = n(),
    n_pondere = sum(EXTRI),
    proportion_ponderee = n_pondere/sum(df$EXTRI)*100
  ) %>%
  mutate(across(where(is.numeric), round, 1))
# Salaire moyen dans QPV avec pondération
df %>%
  group_by(QPV) %>%
  summarise(
    salaire_moyen = weighted.mean(SALRED_Y, w = EXTRI,
na.rm = TRUE),
    salaire_median = median(SALRED_Y, na.rm = TRUE)  #
Médiane non pondérée
  )
tableau_contingence <- table(df$IDF, df$QPV)

# Affichage du tableau de contingence
print(tableau_contingence)
unique(df$QPV)
unique(df$IDF)
head(df$QPV)
head(df$IDF)
-------------
# Statistiques descriptives bivariées
# Relation entre le lieu d'habitation (STCOMM2020) et le statut
d'emploi (ACTEU)
table(df$STCOMM2020, df$ACTEU)

# Calcul des pourcentages par lieu d'habitation
df<- df %>%
  mutate(ACTEU = factor(ACTEU, levels = c(1, 2), labels = c("En
emploi", "Chômage")))

# Calcul des pourcentages par lieu d'habitation
df %>%
  group_by(STCOMM2020_rec, ACTEU) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(pourcentage = (count / sum(count)) * 100) %>%

  # Création du graphique
  ggplot(aes(x = STCOMM2020_rec, y = pourcentage, fill =
ACTEU)) +
  geom_bar(stat = "identity", position = "dodge") +  # Barres côte à
côte

  # Labels et mise en forme
  labs(
    title = "Statut d'emploi selon le lieu d'habitation",
    x = "Lieu d'habitation",
    y = "Proportion (%)",
    fill = "Statut d'emploi"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("En emploi" = "#1f78b4",
"Chômage" = "#e31a1c")) +  # Couleurs adaptées
  geom_text(aes(label = sprintf("%.1f%%", pourcentage)),
            position = position_dodge(width = 0.9),
```

```
        vjust = -0.5, size = 4)  # Ajout des pourcentages

df %>%
  group_by(ACTEU, STCOMM2020_rec) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(pourcentage = count / sum(count) * 100)


# Relation entre le lieu d'habitation (STCOMM2020) et le salaire
(SALRED_Y)
tapply(df$SALRED_Y, df$STCOMM2020_rec, summary)

# Relation entre le lieu d'habitation (STCOMM2020) et le type de
contrat (SALTYP)
table(data$STCOMM2020, data$SALTYP)

# Visualisation des données
# Histogramme des salaires par lieu d'habitation
library(ggplot2)
ggplot(data, aes(x = SALRED_Y, fill = STCOMM2020)) +
  geom_histogram(binwidth = 500, alpha = 0.6) +
  facet_wrap(~ STCOMM2020) +
  labs(title = "Distribution des salaires par lieu d'habitation", x =
"Salaire", y = "Fréquence")

# Boxplot des salaires par lieu d'habitation
data_filtré <- data %>% filter(SALRED_Y > 0)  # Enlève les
salaires à 0

ggplot(data_filtré, aes(x = STCOMM2020_rec, y = SALRED_Y,
fill = STCOMM2020_rec)) +
  geom_boxplot() +
  labs(title = "Disparité des salaires selon le lieu d'habitation (hors
chômeurs)",
       x = "Lieu d'habitation", y = "Salaire en euros") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3")

------------------------------------------------------------
#test de significativité
chisq.test(table(df$STCOMM2020, df$ACTEU))
# Résultat : p-value = 0.003 → Écart significatif
t.test(SALRED_Y ~ IDF, data = df)
# Résultat : p-value < 2.2e-16 → Écart très significatif
chisq.test(table(df$QPV, df$ACTEU))
# Filtrer les données pour retirer les valeurs manquantes
df_filtered <- df %>%
  filter(!is.na(SALRED_Y), !is.na(QPV))

# Test de Student
t_test_result <- t.test(SALRED_Y ~ QPV, data = df_filtered,
var.equal = TRUE)
print(t_test_result)

# Tableau croisé entre IDF et QPV
tab_idf_qpv <- table(df$IDF, df$QPV)
prop.table(tab_idf_qpv, margin = 1) # Proportions par ligne

------------------------------------------------------------
# Tableau croisé salaire moyen par lieu et diplôme
tbl_summary(
  df,
  by = STCOMM2020,  # Croisé par lieu d'habitation
  include = c(SALRED_Y, DIP7, SEXE),
  statistic = list(
    SALRED_Y ~ "{mean} ({sd})",  # Moyenne et écart-type
    all_categorical() ~ "{n} ({p}%)"
  ),
  digits = list(SALRED_Y ~ 0)  # Arrondi à l'unité
) %>%
  add_overall() %>%  # Ajoute une colonne "Total"
  modify_header(label = "**Variable**") %>%
  as_gt() %>%  # Convertir pour export
  gt::gtsave("tableau_salaire_lieu_diplome.docx")
------------------------------------------------------------


#première régression
# Passer le salaire en log
df$log_SALRED_Y <- log(df$SALRED_Y)
# Vérification des corrélations avec une table de contingence pour
STCOMM2020 et QPV
correlationQPVSTCOMM <- table(df$STCOMM2020, df$QPV)
print(correlationQPVSTCOMM)
# Matrice de corrélation pour les variables numériques
cor(df[, c("AGE", "SEXE", "SALRED_Y")], use =
"complete.obs")
# Exclure les lignes avec SALRED_Y <= 0 ou NA
df_clean <- df[df$SALRED_Y > 0, ]

# Recalculer log_SALRED_Y après avoir supprimé les valeurs
invalides
df_clean$log_SALRED_Y <- log(df_clean$SALRED_Y)

# Ajuster le modèle de régression (on décide de ne pas inclure
QPV)
modele1 <- lm(log_SALRED_Y ~ STCOMM2020 + AGE +
SEXE + DIP7, data = df_clean)

# Résultats du modèle
summary(modele1)

# Graphique des résidus
plot(modele1$fitted.values, modele1$residuals)
abline(h = 0, col = "red")

# Histogramme des résidus
hist(modele1$residuals)

# Graphique QQ pour vérifier la normalité des résidus
qqnorm(modele1$residuals)
qqline(modele1$residuals, col = "red")

#Test de fisher
# Modèle non contraint (avec toutes les variables)
modele_non_contraint <- lm(log_SALRED_Y ~ STCOMM2020 +
AGE + SEXE + DIP7, data = df_clean)

# Modèle contraint (en excluant SEXE)
modele_contraint <- lm(log_SALRED_Y ~ STCOMM2020 +
AGE + DIP7, data = df_clean)

# Test de Wald pour comparer les deux modèles
library(lmtest)
waldtest(modele_non_contraint, modele_contraint)
```

Un autre script pour le modèle 2 pour faciliter :

```
# Chargement des librairies nécessaires
library(readr)
library(dplyr)
library(questionr)
library(ggplot2)
library(gtsummary)
library(tidyverse)

# Importation des données
indiv231 <- read_delim("Desktop/miashs/MIASHS
L3/S6/PROJET (ÉCONOMÉTRIE)/indiv231.csv",
                       delim = ";", escape_double = FALSE, trim_ws =
TRUE)
```

```
# Filtrage des données selon les critères souhaités
dataselected <- indiv231 %>%
  filter(AGE >= 15 & AGE <= 64,  # Limite d'âge classique pour
l'emploi
         ACTEU %in% c(1, 2))     # Garde uniquement les individus
en emploi ou au chômage

# Recode de la variable ACTEU :
# 1 = En emploi, 0 = Chômage
dataselected$ACTEU <- ifelse(dataselected$ACTEU == 1, 1, 0)

# Vérification que le recodage a bien fonctionné
table(dataselected$ACTEU)

# Filtrer les colonnes d'intérêt pour ton modèle
df <- dataselected[, c(
  "STCOMM2020", "REG", "DENS2022", "QPV",
"METRODOM",
  "ACTEU", "STATUT", "SALTYP", "PCS1Q",
  "SALRED_Y", "AGE", "SEXE", "DIP7", "EXTRI"
)]

# Transformation de certaines variables en facteurs si nécessaire
df$STCOMM2020 <- as.factor(df$STCOMM2020)
df$SEXE <- as.factor(df$SEXE)
df$DIP7 <- as.factor(df$DIP7)

#  Modèle de régression logistique pour prédire le statut d'emploi
(ACTEU)
modele2 <- glm(ACTEU ~ STCOMM2020 + AGE + SEXE +
DIP7,
               data = df,
               family = binomial)

# Résumé des résultats du modèle
summary(modele2)
```