Projet Econométrie 2

Position sur le marché du travail et investigation sur
de potentielles inégalités liées à l'origine

Chargé de TD : Lucas Girard - lucas.girard@ensae.fr

Présentation Ce projet cherche à étudier et quantifier l'effet de différentes variables, à la fois individuelles et relatives à la zone résidentielle géographique des individus, sur leur position sur le marché du travail. Par ce dernier terme, on entend à la fois une marge extensive - actif occupé, chômeur, inactif - et une marge intensive - le salaire.

La principale variable explicative d'intérêt sera une variable individuelle catégorielle "origine", construite à partir des nationalités de naissance des parents [1]. Elle pourra également être considérée en interaction avec d'autres variables catégorielles comme par exemple "immi", l'indicatrice d'être immigré (c'est-à-dire être né étranger à l'étranger). L'objectif du projet consiste en effet à mesurer l'effet potentiel de la variable "origine" sur la position sur le marché du travail et à rechercher d'éventuelles inégalités dues à l'origine des individus. Dans un premier temps, on s'intéressera aux salaires des actifs occupés et discutera de l'existence d'éventuelles inégalités salariales. Dans un second temps, on étudiera la marge extensive, à savoir le fait d'être ou non chômeur parmi la population d'actifs.

Consignes générales Le projet n'est pas fondé sur un article de recherche précis et il présente un caractère exploratoire, laissant une certaine liberté aux élèves. Une attention particulière sera ainsi portée aux choix des modèles et variables retenues, aux quantités d'intérêt considérées, à l'interprétation et aux limites des estimations réalisées.
Le projet présente trois objectifs pédagogiques :

- appliquer certaines méthodes vues en cours (instruments, panel, tobit/sélection)
- illustrer la difficulté à répondre de manière raisonnable et suffisamment complète à une question empirique - existe-t-il des inégalités sur le marché du travail liées à l'origine des individus - en combinant données et modèles économétriques : quelles variables d'intérêt, quels modèles (et derrière quelles hypothèses), quelles variables explicatives, quelles observations ?
- découvrir une nouveauté théorique [2] : les panels lorsque la variable dépendante est binaire, et plus précisément les modèles "random effect probit" et "fixed effect logit".

1. Il s'agit d'un *proxy* fréquemment utilisé dans la littérature pour mesurer le groupe ethnique auquel appartient un individu. La variable "origine" prend ici neuf modalités : France, Europe du Nord, Europe du Sud, Europe de l'Est, Maghreb, reste de l'Afrique, Proche-Orient, Asie du Sud-Est (Cambodge, Laos, Vietnam), reste du monde.

2. Des références sont fournies (*cf. infra*) et la partie théorique du projet consistera essentiellement à lire les références, comprendre ces méthodes - qui pourront être importantes dans votre formation d'économètre - et les appliquer.

Données La base de données [3] regroupe $n = 70944$ individus suivis pendant $T = 6$ trimestres, entre 2014 et 2016 (panel cylindré). Pour chaque observation individu $\times$ trimestre, on dispose de variables :

- démographiques et socio-économiques individuelles : âge, sexe, diplôme, CSP des parents, "origine", "immi", une indicatrice d'être descendant d'immigré ("desc"), état de santé déclaré, statut matrimonial, etc.
- individuelles liées à la position sur le marché du travail : statut (actif occupé, chômeur, inactif), salaire, nombre d'heures travaillées, type d'horaires, type de contrat de travail, etc.
- relatives à la zone résidentielle dans laquelle habite l'individu : tranche d'unité urbaine, commune urbaine ou rurale, proportions de populations de différentes origines dans cette zone, département, etc.

1 Impact de la variable "origine" sur le salaire

Dans un premier temps, on néglige l'aspect censuré de la variable salaire en se concentrant sur les individus actifs occupés et on s'intéresse à l'existence de potentielles inégalités salariales dues à l'origine. Le salaire n'est disponible qu'au premier et au dernier trimestre - il s'agit d'un panel avec $T = 2$ pour cette variable. Le concept d'inégalité salariale due à l'origine est similaire à celui d'inégalité salariale homme-femme. Dans la notion même, on voit un lien avec l'approche économétrique puisqu'on parlera ici d'inégalité salariale due à l'origine lorsque que *toutes choses égales par ailleurs* les salaires des individus diffèrent selon leur origine. En effet, dès lors que ces différences ne sont pas justifiées, elles peuvent être interprétées comme des inégalités [4]. La question est donc ce qui est inclus dans le *toutes choses égales par ailleurs*.

Vous avez une marge de liberté quant aux choix des quantités d'intérêt, variables et observations utilisées pour étudier cette question. Pensez notamment à la sélection des observations (population d'intérêt, valeurs manquantes), l'interprétation des paramètres (effet causal ou corrélation), la fiabilité de l'inférence en ce qui concerne le calcul des écarts-types.

Vous répondrez néanmoins aux questions suivantes qui doivent vous guider dans votre réponse à la problématique.

On néglige la dimension panel pour l'instant en ayant des observations à l'échelle des individus et on considère un modèle linéaire avec comme variable dépendante le salaire en se restreignant aux actifs occupés. Vous êtes laissés libre de la spécification précise du modèle et notamment du choix des variables de contrôle [5].

3. Le détail des variables et leur encodage, ainsi que la base de données elle-même, seront transmis ultérieurement aux élèves travaillant sur ce projet. Je reste disponible en cas de questions plus précises concernant la base et les variables disponibles.

4. Ces quelques explications précisant ce que nous allons chercher dans ce projet n'entendent absolument pas épuiser ces questions. On peut en tout cas garder en tête la différence entre égalité et équité et le fait que mettre en évidence des différences à un instant donné ne rend pas compte des dynamiques et de l'évolution des situations individuelles au cours de la vie, qui est un film plus qu'une photo.

5. On pourra également considérer différentes variables explicatives d'intérêt à partir des variables "origine", "immi" et "desc" : interactions, regroupement de plusieurs modalités de la variable "origine", etc. On pourra également discuter de la spécification fonctionnelle : modèle en niveau, en logarithme, etc. ?

(a) Estimer un tel modèle et expliquer vos choix concernant les variables de contrôle. Souhaite-t-on ici avoir des contrôles ? [6] Quel intérêt pourrait avoir l'inclusion de variables de contrôle dans un objectif de politiques publiques visant à réduire d'éventuelles inégalités salariales ? Commenter et interpréter les résultats de votre estimation au regard de la problématique, à savoir la recherche d'éventuelles inégalités salariales liées à l'origine.

Un déterminant important du salaire est le niveau de diplôme et on peut souhaiter l'inclure comme variable de contrôle.

(b) Quel problème pourrait survenir en incluant cette variable explicative dans votre modèle ? Au vu des variables disponibles dans la base, proposer un instrument pour le niveau de diplôme et discuter de sa crédibilité. Estimer le modèle correspondant et interpréter les résultats.

Sous certaines hypothèses, les panels peuvent également résoudre un problème d'endogénéité.

(c) Néanmoins, quelles difficultés se posent ici si l'on souhaite étudier l'effet de la variable "origine" tout en incluant un effet fixe individuel ? Pouvez-vous utiliser la structure panel pour étudier l'effet de la variable "origine" ? [7] Le cas échéant, estimer ce ou ces modèles et interpréter les résultats obtenus.

On dispose également de variables à l'échelle de la zone résidentielle mesurant les proportions pour plusieurs groupes minoritaires définis par la variable "origine".

(d) Est-ce que celles-ci varient au cours du temps ? Estimer un modèle panel en incluant ces variables de contrôle [8]. Interpréter les résultats obtenus : cette estimation peut-elle apporter des éléments de réponses pour notre problématique ?

2 Impact de la variable "origine" sur le fait d'être ou non au chômage

Les inégalités sur le marché du travail dues à l'origine peuvent également survenir sur une marge extensive : avoir ou non un emploi et donc un salaire [9]. On s'intéresse ici aux individus actifs et aux déterminants d'être ou non chômeur, en particulier l'effet de la variable "origine". Le statut sur le marché du travail (actif occupé, chômeur, inactif) est disponible chaque trimestre (donnés de panel avec $T = 6$) mais la variable dépendante est binaire.

6. On pourra notamment se demander si l'on cherche à mesurer une causalité ou une corrélation lorsqu'on s'intéresse aux inégalités salariales.

7. On gardera en tête la structure générale de votre cours sur les panels : *(i)* exogénéité des résidus mais autocorrélation, *(ii)* effets individuels α_i corrélés avec les régresseurs.

8. A nouveau, la spécification précise est à choisir ; il ne faudra pas nécessairement utiliser toutes les variables de type proportion de minoritaire mais utiliser ce type de variables explicatives.

9. Plus largement, on pourrait également s'intéresser aux nombres d'heures travaillées, au type d'horaires et aux conditions de travail hors salaire.

Une première manière d'aborder ce problème est de négliger la limitation de la variable dépendante avec un modèle de probabilité linéaire.

(a) Après avoir discuté de la crédibilité des hypothèses d'exogénéité stricte et faible, estimer un modèle de panel dans ce cadre (i.e. en considérant la variable dépendante comme continue). Quels problèmes pose cette approche ?

On prend maintenant en compte la limitation binaire. Votre cours sur les méthodes de panel comprend trois parties principales selon les hypothèses sur le résidus, composé de l'effet individuel (α_i dans les notations du cours) et d'un choc idiosyncratique (ε_{it}).

(b) Dans le cas "exogénéité des résidus mais autocorrélation", quelles sont les conséquences d'avoir une variable dépendante binaire ? Les hypothèses sur les résidus dans les modèles probit ou logit sont-elles compatibles avec l'existence d'un effet individuel et d'une autocorrélation des erreurs ? Estimer un probit ou un logit et discuter de sa crédibilité. On estimera également un modèle probit ou logit en négligeant la dimension panel avec des observations à l'échelle des individus [10] en guise de comparaison.

(c) Dans les sections "exogénéité stricte" et "exogénéité faible" du cours, vous avez vu deux transformations (*first-difference* et *within*) pour éliminer les effets individuels α_i. Ces transformations sont-elles applicables ici ?

Deux solutions sont proposées dans la littérature dans ce cas de panel binaire. La première consiste à supposer les effets individuels α_i indépendants des covariables et à spécifier une distribution paramétrique pour ces α_i. En particulier, un modèle classique est le "random effect probit model", qui suppose $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$.

(d) Comment s'estime ce modèle ? Peut-on calculer les effets marginaux dans ce modèle ? [11]
Estimer un "random effect probit model" (on pourra utiliser la commande Stata `xtprobit`).
NB : on pourra se référer à la section 15.8.2 du Wooldridge - *Econometrics Analysis of Cross Section and Panel Data*, 2nd edition.

Le "random effect probit model" n'autorise pas de corrélation entre les régresseurs et l'effet individuel. Le "fixed effect logit model" (parfois appelé "conditional fixed effect logit" ou encore "conditional logit estimator") relâche cette hypothèse en utilisant une transformation - similaire dans sa fonction au *first-difference* ou au *within* dans le cadre linéaire - permettant d'éliminer l'effet individuel.

(e) Comment s'estime ce modèle ? Peut-on calculer les effets marginaux dans ce modèle ? [11]
Estimer un "fixed effect logit model" (on pourra utiliser la commande Stata `xtlogit`).

10. On pourra prendre un trimestre de référence arbitraire ou chercher à synthétiser les six trimestres, avec la modalité la plus fréquente par exemple.

11. Pour ces questions théoriques, il n'est pas attendu des preuves ou des calculs. L'idée est de découvrir et comprendre les modèles - en lisant la référence ou autre chose - et de répondre aux questions avec des mots ou au plus quelques équations.

NB : on pourra se référer à la section 15.8.3 du Wooldridge - *Econometrics Analysis of Cross Section and Panel Data*, 2nd edition.

(f) Peut-on comparer les différentes estimations réalisées dans cette partie ? Quelles conclusions en tirez-vous quant à l'effet de la variable "origine" sur le fait d'être au chômage ?

3 Impact de la variable "origine" dans un modèle de sélection ou censure

Dans les deux sections précédentes, on a analysé séparément le salaire pour les actifs occupés et le fait d'être au chômage parmi les actifs. On pense également à des modèles de censure ou de sélection pour étudier la variable salaire parmi les actifs, à la fois actifs occupés et chômeurs - la variable n'étant observée que pour les actifs occupés. On s'intéresse ici à l'effet de la variable "origine" dans cette modélisation.

(a) Quel cadre, entre le modèle de censure (tobit simple) et le modèle de sélection (tobit généralisé ou tobit II), semble le plus approprié ici ? [12] S'intéresse-t-on a l'effet de l'origine sur la variable observée (Y dans les notations du cours) ou sur la variable latente (Y^*) ?

(b) Estimer le modèle choisi à la question précédente et commenter les résultats obtenus.

4 Conclusion

Suite à ces différentes estimations, quelles sont vos conclusions quant à l'effet de la variable "origine" sur la position sur le marché du travail ? Diriez-vous qu'il existe des inégalités sur le marché du travail liées à l'origine des individus ? Les estimations réalisées suggèrent-elles des interventions publiques permettant d'agir sur ces inégalités éventuelles ?

Quelles données supplémentaires ou autres pistes de recherche vous sembleraient être intéressantes pour approfondir votre réponse ? Quelles limites voyez-vous dans les analyses menées ?

12. Indice pour une modélisation possible. Dans l'appréhension du chômage structurel (par opposition à conjoncturel et frictionnel), on peut penser que les chômeurs recherchent un travail - ils sont actifs sur le marché du travail - mais qu'il existe un salaire plancher (salaire minimum légal) en deçà duquel les employeurs ne peuvent les payer et donc les embaucher.

Annexes : descriptif de la base

La base de données contient 425 664 observations et 81 variables. Elle est au format Stata (.dta).

NB : un certain nombre de variables ont déjà un label clair dans la base Stata et ne seront donc pas détaillées ci-dessous lorsque les modalités ne posent pas de difficultés particulières (par exemple il y a dans la base une variable `dep` avec pour label "département" et des modalités allant de 01 à 95). Cela n'interdit pas de les regarder et éventuellement de les utiliser évidemment.

Par ailleurs, les modalités "." pour des variables de type numérique et "" pour des variables de type "*character*" désignent - c'est la convention en Stata - des valeurs manquantes ou sans objet i.e. non définie par construction [13]. On pourra bien sûr prendre en compte ces informations et sélectionner les observations sur lesquelles on mène les analyses en fonction.

La base est large. Comme expliqué dans les questions, vous êtes laissés relativement libre quant aux choix des variables de contrôle et spécifications précises des modèles ; il n'est probablement pas nécessaire d'utiliser toutes ces variables ; au contraire, il s'agit d'identifier et de sélectionner celles qui vous semblent les plus pertinentes pour répondre de manière convaincante à la problématique et de justifier ces choix. On ne s'interdira pas de définir de nouvelles variables, par exemple en regroupant les modalités de certaines variables ou en considérant des interactions, des termes au carré, etc. De même, il sera sûrement intéressant de sélectionner les observations pour ne conserver qu'une population d'intérêt, celle qui nous semble la plus pertinente pour répondre à la question.

Une observation est uniquement identifiée dans la base par un couple individu $\times$ période, avec la variable `id_individu` pour identifiant des individus (i.e. un individu est identifié uniquement par une valeur de `id_individu`) et la variable `t`, allant de 1 à 6 (on suit les individus pendant 6 trimestres). Il s'agit ainsi de données de panel avec $n = 70\ 944$ individus et $T = 6$ périodes de temps. Le panel est cylindré.

On dispose également d'un identifiant des ménages. Il y a 44 154 ménages dans la base. Les individus au sein d'un même ménage partagent la même valeur pour la variable `id_menage`. Au sein d'un ménage, les individus sont numérotés par la variable `num_ind_par_menage`. On pourra utiliser ou non cet autre niveau des données ; en première approche, il est conseillé de ne pas prendre en compte cette structure supplémentaire des données.

On pourra néanmoins remarquer que cette variable `num_ind_par_menage` ne prend que deux valeurs 1 et 2. Cela amène à dire quelques mots sur le travail de sélection et de construction de cette base i.e. comment ont été sélectionnées les observations [14] Les individus dans la base sont la personne de référence du ménage et, le cas échéant, son conjoint. Il s'agit également des individus qui ont pu être interrogés pendant les six trimestres consécutifs de l'enquête : d'une part, ils n'ont pas changé de logements (l'enquête est réalisée en tirant aléatoire des logements dont on interroge, ou du moins cherche à interroger, les occupants), d'autre part ils ont pu être contactés et ont répondu à l'enquête.

13. Par exemple l'indicatrice d'être pacsé lorsque le statut matrimonial légal est d'être marié ; la catégorie socio-professionnel des actifs lorsqu'on est inactif ; etc.

14. Ces informations vous seront principalement utiles pour répondre aux questions de conclusion.

On détaille ci-dessous la liste des variables dans la base, à l'exception des variables dont le label dans la base Stata et les modalités sont suffisamment explicites.

- `rim` : rang d'interrogation du ménage. Les ménages sont interrogés pendant six trimestres consécutifs a priori mais pour certains types d'individus ils ne sont interrogés que deux fois, au premier et au sixième trimestre. ***NB** : on pourra voir quels sont les individus concernés [indice : regarder leur âge] et on pourra tout à fait justifier de ne pas les prendre en compte dans l'analyse.*
- `typmen` : type de ménage : 1 = ménage d'une seule personne, 2 = famille monoparentale, 3 = couple sans enfant, 4 = couple avec enfant(s), 5 = ménage de plusieurs personnes ayant toutes un lien de parenté avec la personne de référence du ménage, ni couple, ni famille monoparentale, 6 = ménage de plusieurs personnes n'ayant pas toutes un lien de parenté avec la personne de référence du ménage, ni couple, ni famille monoparentale, 9 = autres ménages complexes de plus d'un personne.
- `region` : (avant les nouveaux noms pour certaines régions) 1 = Ile de France, 2 = Bretagne, 3 = Aquitaine - Limousin - Poitou-Charente, 4 = PACA, 5 = Auvergne - Rhône-Alpe, 6 = Alsace Lorraine Champagne Ardennes , 7 = Midi Pyrenee Languedoc Roussillon 8 = Nord pas de Calais - Picardien 9 = Normandie, 10 = Pays de la Loire, 11 = Centre - Val de Loire, 12 = Bourgogne Franche-Comté, 13 = Corse, 14 = hors France-métropolitaine (DOM)
- `tuu` : tranche d'unité urbaine du logement (un logement est associé à un unique ménage et aux individus composant ce ménage ; puisque la base est restreinte aux individus n'ayant pas changé de logement comme expliqué ci-dessus) : 0 = Commune rurale, 1 = Unité urbaine de moins de 5 000 habitants, 2 = Unité urbaine de 5 000 à 9 999 habitants, 3 = Unité urbaine de 10 000 à 19 999 habitants, 4 = Unité urbaine de 20 000 à 49 999 habitants, 5 = Unité urbaine de 50 000 à 99 999 habitants, 6 = Unité urbaine de 100 000 à 199 999 habitants, 7 = Unité urbaine de 200 000 à 1 999 999 habitants, 8 = Unité urbaine de Paris
- `tuu_r` : idem avec des modalités regroupées : 1 = Commune rurale, 2 = Unité urbaine de moins de 20 000 habitants, 3 = Unité urbaine de 20 000 à moins de 200 000 habitants, 4 = Unité urbaine de 200 000 habitants ou plus (sauf agglomération parisienne), 5 = Agglomération parisienne
- `typvois` : type d'habitat ou de voisinage du logement : 1 = Maisons dispersées, hors agglomération, 2 = Maison en lotissement, en quartier pavillonnaire ou en ville, 3 = Immeubles en ville (autres que cité ou grand ensemble), 4 = Immeubles en cité ou grand ensemble, 5 = Habitat mixte : à la fois immeubles et maisons
- `occupation_logement` : statut d'occupation dans le logement : 0 = Personne habitant en permanence dans ce logement, 1 = Enfant en garde alternée, 2 = Militaire de carrière logé par ailleurs en caserne ou camp, 3 = Élève interne, étudiant habitant par ailleurs en communauté (cité U ou foyer d'étudiants), 4 = Élève ou étudiant habitant par ailleurs dans un logement indépendant, considéré comme occasionnel (sauf communauté) (Le logement enquêté est celui des parents), 5 = Jeune vivant par ailleurs dans un foyer de jeunes travailleurs, 6 = Élève ou étudiant hébergé dans un logement différent de celui de ses parents, considéré comme occasionnel (sauf communauté) (Le logement enquêté n'est pas celui des parents), 7 = Personne vivant par ailleurs dans un sanatorium, hôpital, établissement de soin, ou en prison, 8 = Personne vivant par ailleurs dans une maison

de retraite, un hospice, 9 = Personne (autre qu'étudiant) habitant par ailleurs dans un logement occasionnel (sauf communauté)

— `matri` : statut matrimonial légal : 1 = Célibataire, 2 = Marié ou remarié, 3 = Veuf, 4 = Divorcé

— `nat` : nationalité (en quatorze modalités) : 10 = Français, 11 = Espagne, 12 = Italie, 13 = Portugal, 14 = Autres pays de l'UE à 28, 15 = Autres pays d'Europe, 21 = Algérie, 22 = Maroc, 23 = Tunisie, 24 = Autres pays d'Afrique, 31 = Turquie, 32 = Cambodge-Laos-Vietnam, 41 = Amérique (continent américain), 51 = Autres pays (d'Asie ou Océanie) ou apatride, 99 = Personne étrangère de nationalité non précise

— `acteu` : statut sur le marché du travail au sens du BIT (nom de la variable pour activité EU - harmonisation de la définition au niveau de l'Union Européenne) : 1 = Actif occupé, 2 = Chômeur, 3 = Inactif

— `acteu_detail` : idem avec des modalités plus détaillées : 1 = Actif occupé, 3 = Chômeur PSERE (Population sans Emploi à la Recherche d'un Emploi), 4 = Autre chômeur BIT, 5 = Etudiant, élève, stagiaire en formation (inactifs), 6 = Autres inactifs (dont retraités)

— `ancchom` et `ancinactif` : variables d'ancienneté du chômage et de l'inactivité en mois (cf. label Stata) ; précision simplement : 99 signifie 99 mois ou plus

— `halochomage` : halo autour du chômage (caractérisation partielle des chômeurs au sens du BIT) : 1 = Recherchent un emploi, mais ne sont pas disponibles, 2 = Disponibles pour prendre un emploi, mais n'en recherchent pas, 3 = Souhaitent un emploi, mais n'en recherchent pas et ne sont pas disponibles

— `csp_actif_detaille` : catégorie socio-professionnelle (CSP) pour les actifs (modalités détaillées) : Vide = Sans objet (ACTEU distinct de 1 et 2), 00 = Non renseigné, 10 = Agriculteurs, 11 = Agriculteurs sur petite exploitation, 12 = Agriculteurs sur moyenne exploitation, 13 = Agriculteurs sur grande exploitation, 21 = Artisans, 22 = Commerçants et assimilés, 23 = Chefs d'entreprise de 10 salariés ou plus, 31 = Professions libérales, 33 = Cadres de la fonction publique, 34 = Professeurs, professions scientifiques, 35 = Professions de l'information, des arts et des spectacles, 37 = Cadres administratifs et commerciaux d'entreprises, 38 = Ingénieurs et cadres techniques d'entreprises, 42 = Instituteurs et assimilés, 43 = Professions intermédiaires de la santé et du travail social, 44 = Clergé, religieux, 45 = Professions intermédiaires administratives de la fonction publique, 46 = Professions intermédiaires administratives et commerciales des entreprises, 47 = Techniciens, 48 = Contremaîtres, agents de maîtrise, 52 = Employés civils et agents de service de la fonction publique, 53 = Policiers et militaires, 54 = Employés administratifs d'entreprise, 55 = Employés de commerce, 56 = Personnels des services directs aux particuliers 62 = Ouvriers qualifiés de type industriel, 63 = Ouvriers qualifiés de type artisanal, 64 = Chauffeurs, 65 = Ouvriers qualifiés de la manutention, du magasinage et du transport 67 = Ouvriers non qualifiés de type industriel, 68 = Ouvriers non qualifiés de type artisanal, 69 = Ouvriers agricoles, 81 = Chômeurs n'ayant jamais travaillé

— `csp_actif_int` : catégorie socio-professionnelle (CSP) pour les actifs (modalités intermédiaires) : Vide = Sans objet (ACTEU distinct de 1 et 2), 00 = Non renseigné, 10 = Agriculteurs exploitants, 21 = Artisans, 22 = Commerçants et assimilés, 23 Chefs d'entreprise de 10 salariés ou plus, 31 = Professions libérales, 32 = Cadres de la fonction publique, professions intellectuelles et artistiques, 36 = Cadres d'entreprise, 41 =

Professions intermédiaires de l'enseignement, de la santé, de la fonction publique et assimilés, 46 = Professions intermédiaires administratives et commerciales des entreprises, 47 = Techniciens, 48 = Contremaîtres, agents de maîtrise, 51 = Employés de la fonction publique, 54 = Employés administratifs d'entreprise, 55 = Employés de commerce, 56 = Personnels des services directs aux particuliers, 61 = Ouvriers qualifiés, 66 = Ouvriers non qualifiés, 69 = Ouvriers agricoles, 81 = Chômeurs n'ayant jamais travaillé

— `csp_actif` : catégorie socio-professionnelle (CSP) pour les actifs (modalités agrégées) : Vide = Sans objet (ACTEU distinct de 1 et 2), 0 = Non renseigné, 1 = Agriculteurs exploitants, 2 = Artisans, commerçants et chefs d'entreprise, 3 = Cadres et professions intellectuelles supérieures, 4 = Professions intermédiaires, 5 = Employés, 6 = Ouvriers, 8 = Chômeurs n'ayant jamais travaillé

— `contrat` : catégorie de contrat de travail : Vide = Sans objet ou non renseigné, 0 = Pas de contrat de travail, 1 = Contrat à durée indéterminée (CDI), 2 = Contrat à durée déterminée (CDD) autre que saisonnier, 3 = Contrat saisonnier, 4 = Contrat d'intérim ou de travail temporaire, 5 = Contrat d'apprentissage

— `heuretra_tranche` : nombre d'heures travaillées en moyenne par semaine dans l'emploi prinicipal (en tranches) : 1 = Moins de 15 heures, 2 = Au moins 15 heures, mais moins de 30 heures, 3 = Au moins 30 heures, mais moins de 35 heures, 4 = Au moins 35 heures, mais moins de 40 heures, 5 = 40 heures ou plus

— `typehoraire` : type d'horaires effectués dans l'emploi principal : Vide = Sans objet (car ACTEU distinct de 1) ou non renseigné, 1 = À peu près semblables d'une semaine sur l'autre, 2 = Alternés : 2x8, 3x8, équipes, ..., 3 = Variables d'une semaine sur l'autre, 4 = Sans objet (a travaillé une seule semaine...)

— `compenheuresup` : existence et le cas échéant nature de la compensation auxquelles donnent droit les heures supplémentaires effectuées : Vide = Sans objet ou non renseigné, 1 Oui, sous forme de rémunération, 2 = Oui, sous forme de repos compensateur (y compris récupération), 3 = Oui, à la fois sous forme de rémunération et de repos compensateur (y compris récupération), 4 = Non, pas de compensation, 9 = Ne sait pas

— `education` : diplôme le plus élevé obtenu (6 modalités) : 1 = Diplôme supérieur à baccalauréat + 2 ans, 3 = Baccalauréat + 2 ans, 4 = Baccalauréat ou brevet professionnel ou autre diplôme de ce niveau, 5 = CAP, BEP ou autre diplôme de ce niveau, 6 = Brevet des collèges, 7 = Aucun diplôme ou certificat d'études primaires

— `education_detail1` : idem (16 modalités) : 10 = Master (recherche ou professionnel), DEA, DESS, Doctorat, 12 = Ecoles niveau licence et au-delà, 22 = Maîtrise (M1), 21 = Licence (L3), 30 = DEUG, 31 = DUT, BTS, 32 = Autre diplôme (niveau bac+2), 33 = Paramédical et social (niveau bac+2), 41 = Baccalauréat général, 42 = Bac technologique, 43 = Bac professionnel, 44 = Brevet de technicien, brevet professionnel, 50 = CAP, BEP, 60 = Brevet des collèges, 70 = Certificat d'études primaires, 71 = Sans diplôme

— `education_detail2` : idem (11 modalités) : 10 = Licence (L3), Maitrise (M1), Master (recherche ou professionnel), DEA, DESS, Doctorat, 11 = Ecoles niveau licence et au-delà, 30 = DEUG, 31 = BTS, DUT ou équivalent, 33 = Paramédical et social (niveau bac+2), 41 = Baccalauréat général, 42 = Baccalauréat technologique, bac professionnel ou équivalents, 50 = CAP, BEP ou équivalents, 60 = Brevet des collèges, 70 = Certificat d'Etudes Primaires, 71 = Sans diplôme

— `csp_mere` et `csp_pere` : CSP de la mère et du père : 00 = Non renseigné, 10 = Agriculteurs, 11 = Agriculteurs sur petite exploitation, 12 = Agriculteurs sur moyenne exploitation, 13 = Agriculteurs sur grande exploitation, 21 = Artisans, 22 = Commerçants et assimilés, 23 = Chefs d'entreprise de 10 salariés ou plus, 31 = Professions libérales, 33 = Cadres de la fonction publique, 34 = Professeurs, professions scientifiques, 35 = Professions de l'information, des arts et des spectacles, 37 = Cadres administratifs et commerciaux d'entreprises, 38 = Ingénieurs et cadres techniques d'entreprises, 42 = Instituteurs et assimilés, 43 = Professions intermédiaires de la santé et du travail social, 44 = Clergé, religieux, 45 = Professions intermédiaires administratives de la fonction publique, 46 = Professions intermédiaires administratives et commerciales des entreprises, 47 = Techniciens, 48 = Contremaîtres, agents de maîtrise, 52 = Employés civils et agents de service de la fonction publique, 53 = Policiers et militaires, 54 = Employés administratifs d'entreprises, 55 = Employés de commerce, 56 = Personnels des services directs aux particuliers, 62 = Ouvriers qualifiés de type industriel, 63 = Ouvriers qualifiés de type artisanal, 64 = Chauffeurs, 65 = Ouvriers qualifiés de la manutention, du magasinage et du transport, 67 = Ouvriers non qualifiés de type industriel, 68 = Ouvriers non qualifiés de type artisanal, 69 = Ouvriers agricoles, 71 = Anciens agriculteurs exploitants, 72 = Anciens artisans, commerçants, chefs d'entreprise, 74 = Anciens cadres, 75 = Anciennes professions intermédiaires, 77 = Anciens employés, 78 = Anciens ouvriers, 81 = Chômeurs n'ayant jamais travaillé, 82 = Autres personnes sans activité professionnelle
— `nat_mere`, `nat_pere`, `origine`. Ces trois variables ont les mêmes modalités :
  — 01 = France,
  — 03 = Europe du Nord,
  — 04 = Europe du Sud,
  — 05 = Europe de l'Est,
  — 06 = Maghreb,
  — 07 = Reste de l'Afrique,
  — 08 = Proche-Orient,
  — 09 = Laos, Vietnam, Cambodge,
  — 10 = Reste du monde
  — 99 = Non renseigné / Pays inconnu

  Les premières sont les nationalités à leur naissance des parents des individus. La variable `origine` est construite de la manière suivante :
  — `origine` = 01 si les deux parents avaient la nationalité française à leur naisse i.e. `nat_mere` = 01 et `nat_pere` = 01
  — Dans le cas contraire, ou bien les deux parents avaient la même nationalité (au sens des modalités précédentes) à leur naissance et alors la variable `origine` est égale à cette modalité : `origine` = x si `nat_mere` = x et `nat_pere` = x pour $x \in \{03, 04, 05, 06, 07, 08, 09, 10, 99\}$
  — ou bien un des parents avait la nationalité française et le deuxième parent avait une autre nationalité, et alors la variable `origine` est égale à la modalité de ce deuxième parent : `origine` = x si `nat_mere` = x et `nat_pere` = 01 ou si `nat_mere` = 01 et `nat_pere` = x, pour $x \in \{03, 04, 05, 06, 07, 08, 09, 10, 99\}$
  — ou bien les deux parents avaient une autre nationalité que la nationalité française

mais différentes ; dans ce derniers cas, la variable `origine` est égale arbitrairement à la variable `nat_mere` (choisir la nationalité du père comme référence ne changerait que marginalement les résultats obtenus ; ce dernier cas étant le plus rare et peu fréquent dans les données) : `origine` $= x$ si `nat_mere` $= x$ et `nat_pere` $= y$ avec $x \neq y$ et $x, y \in \{03, 04, 05, 06, 07, 08, 09, 10, 99\}$

Comme évoqué dans les questions, dans les analyses, on pourra regrouper certaines modalités de la variable `origine`, ou encore considérer des interactions avec les variables `immi` et `descimmi` par exemple, ou encore avec l'âge également.

— `sante_declaree` : appréciation personnelle et déclarée de l'individu sur son état de santé : Vide = Sans objet (ré-interrogation intermédiaire, i.e. autre que `t` = 1 ou 6) ou non renseigné, 1 = Très bon, 2 = Bon, 3 = Assez bon, 4 = Mauvais, 5 = Très mauvais, 8 = Refus, 9 = Ne sait pas

— `nb_ind_for_prop` et les variables `prop_*` et `prop_*_immi`. Dans l'échantillonnage et la réalisation de l'enquête, chaque logement est associé à une zone qui regroupe en moyenne une vingtaine de logements contigus [15]. Les variables `prop_*` et `prop_*_immi` renseignent sur la proportion d'individus ayant certaines caractéristiques pour les variables `origine` et `immi` au sein de la zone dans lequel se trouve le logement de l'individu. La variable `nb_ind_for_prop` indique le nombre d'individus utilisé pour calculer cette proportion. Les caractéristiques en question se réfèrent aux modalités de la variable `origine` (*cf. supra*) et à l'indicatrice d'être ou non immigré (variable `immi`).

15. Noter que cette définition peut regrouper des réalités physiques très différentes selon la densité de peuplement de la zone : plusieurs kilomètres dans des zones rurales entre des lieux-dit, à peu près une rue dans une zone pavillonnaire, un immeuble dans un immeuble de centre-ville, voire un étage d'un grand ensemble.