Projet d'Économétrie 2

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

Alexandre Filiot
Germain Evrard

Encadrant :
Lucas Girard

Préambule

Avant d'entamer ce projet d'économétrie 2, nous souhaitons exposer, dans le cadre de cette première partie, les quelques transformations que nous avons appliquées à la base. Nous avons par ailleurs choisi de détailler certaines de ces transformations (et notamment nos choix) dans le code directement, afin de ne pas alourdir le rapport. En premier lieu, nous avons choisi de supprimer les observations relatives aux individus âgés de plus de 63 ans (strictement). Nous nous sommes pour cela basés sur le nombre total d'interrogations par individu, comme suggéré dans le descriptif de la base. Pour les individus actifs occupés, cette suppression représente 2,90% des individus. **Dans toute la suite, nous nous limiterons donc à des individus âgés de 18 à 64 ans (vous trouverez plus de détails dans le code)**.

Ensuite, nous avons regroupé les régions en 4 catégories selon leur PIB par habitant. On distinguera donc les régions dont le pib par habitant est supérieur à 30 000 euros (catégorie 1, Île-de-France uniquement), compris entre 25 000 et 30 000 euros (catégorie 2), compris entre 20 000 et 25 000 euros (catégorie 3), et enfin inférieur à 20 000 euros (catégorie 4, DOM). Ces chiffres datent de 2015 (année la plus représentée dans la base) et sont basés sur une enquête INSEE [1].

En ce qui concerne les variables ayant trait à l'origine, nous avons également opéré quelques changements. En premier lieu, nous avons regroupé les modalités d'Europe du Nord, de l'Est et du Sud dans une seule modalité (la modalité "02"). Par souci de cohérence, nous avons également créé les variables `prop_02`, `prop_02_immi`, (sommes de proportions). Les variables `origine`, `nat_mere`, `nat_pere`, `immi` et `descimmi` nous ont permis de construire une seule et même variable appelée `cat_origine`. Celle-ci se décompose de la manière suivante (tableau 1) :

Tableau 1 – Modalités de la variable `cat_origine`

Modalités	**Significations**	Poids (%)
`fra2pfra`	Français né français de deux parents français	84,93 %
`fra1pimmi_europe`	Français né d'au moins un parent immigré Européen	5,28%
`immi_Nnatur_Neurope`	Immigré né hors Europe et non-naturalisé	2,61%
`fra1pimmi_Neurope`	Français né d'au moins un parent immigré non-Européen	2,00%
`immi_Nnatur_europe`	Immigré né en Europe (hors France) et non-naturalisé	1,96%
`immi_natur_Neurope`	Immigré né hors Europe et naturalisé français	1,80%
`immi_natur_europ`	Immigré né en Europe (hors France) et naturalisé français	1,25%

Nous avons également modifié les variables `csp_pere` et `csp_mere` afin de regrouper les catégories sociaux professionnelles des parents selon une nomenclature plus générale donnée par l'INSEE : 1 = Agriculteurs exploitants, 2 = Artisants, commerçants et chefs d'entreprise, 3 = Cadres et professions intellectuelles supérieures, 4 = Professions intermédiaires, 5 = Employés, 6 = Ouvriers, 7 = Retraités, 8 = Autres personnes sans activité professionnelle. En fonction de ces regroupements, nous avons créé la variable `csp_parents` égale à :

- `csp_pere` si la mère est inactive ;
- `csp_mere` si le père est inactif ;
- `csp_pere` si les deux parents sont actifs [2] ;
- "7" si le père est retraité et la mère est retraitée ou sans activité professionnelle ;
- "8" si le père est sans activité professionnelle et la mère est retraitée ou sans activité professionnelle.

Ainsi que la variable `parents_act` égale à :

- "99" si `csp_pere` et `csp_mere` ne sont pas renseignés ;
- "0" si les deux parents sont inactifs au sens large (retraités ou sans activité professionnelle) ;
- "1" si au moins l'un des deux parents est actif ;
- "2" si les deux parents sont actifs.

1. Disponible sur `https://statistiques-locales.insee.fr/#c=indicator&i=tcr062.pib_brut&i2=tcr062.pib_hab&s=2015&s2=2015&view=map3`, consulté le 25 Avril 2018.
2. On remarque que, très souvent, l'activité professionnelle des femmes n'est pas renseignée.

Quelques statistiques descriptives

Pour obtenir les statistiques descriptives suivantes, nous avons relevé les informations de la base au moment de la deuxième interrogation, de manière à capter également les individus interrogés deux fois seulement.

Tableau 2 – Statistiques descriptives (1)

	Actifs	**Chômeurs**	**Inactifs**
Sexe			
Homme	51,80%	4,69%	43,51%
Femme	43,81%	4,36%	51,83%
Age			
Moins de 25 ans	32,80%	12,54%	54,66%
Entre 25 et 34 ans	75,61%	9,48%	14,91%
Entre 35 et 49 ans	81,78%	7,33%	10,89%
Entre 50 et 64 ans	58,78%	4,53%	36,69%
Plus de 65 ans	2,58%	0,08%	97,34%
Region			
Ile de France	54,25%	3,94%	41,81%
Province	47,55%	3,87%	48,58%
Outre-mer	39,82%	10,56%	49,62%
Type d'unité urbaine			
Commune rurale	51,68%	2,81%	45,51%
Unité urbaine de moins de 20 000 hab.	45,85%	4,24%	49,91%
Unité urbaine de 20 000 à 200 000 hab.	42,23%	6,30%	51,46%
Unité urbaine de plus de 200 000 hab.	46,85%	5,06%	48,09%
Agglomération parisienne	54,41%	3,93%	41,66%
Type de voisinage			
Maisons dispersées	47,35%	4,07%	48,58%
Lotissement, pavillons (ville)	47,95%	3,22%	48,83%
Immeubles en ville	41,92%	9,90%	48,18%
Immeubles en cité	48,98%	6,91%	44,11%
Zone urbaine sensible	37,17%	9,80%	53,04%
Total sur 70 944 individus	47,47%	4,51%	48,01%

Voici nos constats par rapport à ce premier ensemble de statistiques descriptives :

— La proportion d'hommes actifs est nettement plus importante que celle des femmes actives. En revanche, les taux de chômage sont presque identiques. La compensation vient des proportions hommes/femmes d'inacti.fs.ves, celles-ci étant parfaitement symétriques avec celles d'hommes/femmes acti.fs.ves. On peut y voir une moindre inclination des femmes à se porter sur le marché du travail : femmes au foyer, plus grandes proportions de retraitées étant donné les départs anticipés (avantages liés au nombre d'enfants notamment), etc... Ce dernier point se confirme dans la base ; en effet, les individus âgés de plus de 63 ans et inactifs (dont potentiellement et vraisemblablement retraités) sont à 56.21% des femmes et 43.79% des hommes.
— L'âge semble être positivement corrélé à la probabilité d'être actif. En revanche on anticipe une dépendance négative au carré de l'âge. On remarque également que le taux de chômage est de 12,5% chez les moins de 25 ans. Ceux-ci sont généralement à la recherche d'emploi après leur sortie d'études.
— La donnée de la région marque de grandes différences dans les proportions. On note d'une part que le fait de travailler en Ile-de-France (et *a fortiori* l'agglomération parisienne) influence positivement l'activité. On peut penser que les activités techniques et de gouvernance (sièges sociaux, cadres, etc...) centralisées en Ile-de-France poussent la proportion d'actifs vers le haut. Les taux de chômage entre Province et Ile-de-France sont encore une fois identiques, la différence se fait sur le nombre d'inactifs. On notera enfin le très fort taux de chômage des individus habitant en Outre-mer.
— Le type de voisinage est sujet à diverses interprétations. On note la grande influence d'être habitant d'une

ZUS sur la probabilité d'être actif. Il est étonnant de voir que les habitants d'immeubles en cité présentent le taux d'activité le plus élevé. En particulier, la différence entre immeubles en cité et immeubles en ville est marquante. On aurait pu s'attendre à des proportions inversées.

Tableau 3 – Statistiques descriptives (2)

	Actifs	**Chômeurs**	**Inactifs**
Type de ménage			
Une seule personne	32,97%	4,06%	62,97%
Famille monoparentale	48,34%	11,08%	40,58%
Couple sans enfant	30,39%	2,11%	67,50%
Couple avec enfant	76,34%	5,53%	18,13%
Diplôme			
Diplôme supérieur	70,87%	3,24%	25,89%
Bac +2	69,86%	3,68%	26,46%
Bac ou brevet pro.	58,14%	4,92%	36,94%
CAP, BEP	52,60%	5,22%	42,19%
Brevet des collèges	37,25%	3,62%	59,14%
Aucun diplôme	22,18%	4,73%	73,09%
Origine			
`fra2pfra`	48,09%	4,01%	47,89%
`fra1pimmi_europe`	42,35%	3,53%	54,13%
`fra1pimmi_Neurope`	57,96%	9,15%	32,89%
`immi_natur_europ`	31,57%	2,71%	65,73%
`immi_natur_Neurope`	53,80%	10,04%	36,16%
`immi_Nnatur_europe`	43,52%	4,32%	52,16%
`immi_Nnatur_Neurope`	35,67%	16,03%	48,30%
Catégorie socioprofessionnelle			
Agriculteurs	99,90%	0,10%	-%
Artisants, commerçants	95,69%	4,31%	-%
Cadres	96,48%	3,52%	-%
Professions intermédiaires	94,92%	5,08%	-%
Employés	90,33%	9,67%	-%
Ouvriers	86,46%	13,54%	-%
Jamais travaillé	0	323 ind.	-%
Santé déclarée			
Très bon	66,21%	5,39%	28,40%
Bon	52,85%	4,50%	42,64%
Assez bon	32,19%	4,09%	63,72%
Mauvais	17,69%	3,87%	78,43%
Très mauvais	10,23%	2,11%	87,66%
Proportions moyennes de minorités			
`prop_n01_immi`	7,17%	11,0%	7,48%
`prop_678910_immi`	4,26%	8,30%	4,42%
Total sur 70 944 individus	47,47%	4,51%	48,01%

Ce second tableau permet d'avoir de plus amples observations sur l'influence de l'origine et des proportions de minorités sur les actifs. Les proportions observées au niveau des autres variables (type de ménage, diplôme, catégorie socioprofessionnelle, santé déclarée) s'interprétent de manière assez intuitive. Voici nos constats :

— Les individus d'origine «non-européenne», qu'ils soient français d'au moins un parent non-européen, français naturalisé ou immigré non-européen (non-naturalisé), présente des taux de chômage des plus élevés, en comparaison à leur équivalent «européen». En particulier, les immigrés d'origines extérieures à l'Europe sont particulièrement touchés par le chômage.

— Les individus d'origine «européenne» sont davantage inactifs en comparaison avec les modalités équivalentes «non-européennes». Ceci n'est cependant pas vérifié pour les immigrés non-naturalisés. Ils présentent également des taux de chômage très faibles. Nous pensons alors à l'expatriation d'Européens du Nord ou de l'Est vers la France dont les compétences sont mieux valorisées que les personnes originaires de pays d'Afrique ou d'Asie.
— Enfin, la proportion moyenne de minorités semble être positivement corrélée au fait d'être chômage ou inactif. Ceci est un résultat important.

Tableau 4 – Statistiques descriptives (3)

Type de voisinage	`prop_n01_immi`	`prop_678910_immi`
Maisons dispersées	4,44%	2,29%
Lotissement, pavillons (ville)	5,23%	2,48%
Immeubles en ville	11,82%	7,51%
Immeubles en cité	20,78%	17,10%
Zone urbaine sensible	25,18%	21,31%

Le tableau 4 rend compte de la corrélation positive entre précarité du logement et proportions moyennes de minorités. Il s'agit d'un deuxième résultat important et qui permet de faire le lien avec l'observation précédente. En combinant ces chiffres, nous aurions tendance à dire que le taux de chômage augmente avec la précarité du logement. Cette précarité étant elle-même positivement corrélée aux proportions de minorités. En résumé, il semblerait que les individus d'origine étrangère sont plus enclin à vivre dans des logements plus précaires. Ces derniers sont également plus victimes du chômage ou de l'inactivité. Un individu quelconque au chômage aura donc plus de risque de se retrouver dans un habitat précaire avec un voisinage d'origines diverses.

Tableau 5 – Statistiques descriptives (4)

	Moyennes de proportions				
Origine	`prop_n01_immi`	`prop_678910`	`prop_678910_immi`	`prop_02`	`prop_02_immi`
`fra1pimmi_Neurope`	18%	19%	15%	9%	3%
`fra1pimmi_europe`	9%	7.6%	4%	17%	4%
`fra2pfra`	6%	6%	3%	8%	3%
`immi_Nnatur_Neurope`	32%	44%	29%	7%	3%
`immi_Nnatur_europe`	17%	11%	6%	19%	11%
`immi_natur_Neurope`	24%	33%	21%	9%	4%
`immi_natur_europ`	14%	9%	5%	19%	9%

Le tableau 5 illustre enfin la corrélation entre nationalité d'origine et nationalité dominante de l'environnement social. Systématiquement, nous notons que les individus d'origine non-européenne, au sens large, habitent dans une zone résidentielle où la minorité non-européenne est la plus représentée parmi les autres minorités, au travers des variables `prop_678910` et `prop_678910_immi`. Ce raisonnement s'applique pour les individus européens (mais non français) vis-à-vis de la communauté "européenne". Nous remarquons également que les français nés en France de parents français, très majoritaires dans la base, sont en moyenne peu brassés avec les minorités, si ce n'est la minorité non-immigrée européenne. Enfin, si nous prenons uniquement la colonne `prop_678910` et les individus (en ligne) d'origine non-européenne, nous nous apercevons que plus ces derniers sont intégrés au territoire français au sens administratif (soit français, soit naturalisé, soit immigré non naturalisé), plus la proportion moyenne de personnes non-européennes présente dans leurs voisinages est faible. Nous y reviendrons.

1. Impact de la variable `origine` sur le salaire

(a) Estimer un tel modèle et expliquer vos choix concernant les variables de contrôle. Souhaite-t-on ici avoir des contrôles ? 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.

L'analyse de la discrimination salariale suscite de nombreuses divergences quant aux méthodes statistiques employées. Prenons l'exemple des discriminations de salaire entre hommes et femmes. Imaginons que nous disposions d'un échantillon aléatoire et représentatif d'individus dont nous connaissons certaines caractéristiques (sexe, localisation, conditions de travail, éducation, santé, etc...) et que voulions étudier l'impact du sexe sur le salaire horaire perçu. Une première démarche serait de régresser simplement le salaire horaire sur l'indicatrice d'être un homme. Seulement, serions-nous en mesure de distinguer l'effet « pur »du sexe sur le salaire ? Sûrement pas. Il est même fort probable que nous obtenions des résultats biaisés positivement du type : « sur le marché français du travail, les hommes gagnent en moyenne 30% de plus que les femmes », alors qu'en réalité ce chiffre tablerait sur une différence de 20%. Notre analyse souffrirait donc d'un très fort biais de variables omises.

Pour étudier l'effet réel des inégalités salariales, l'analyse *toutes choses égales par ailleurs* s'avère être la méthodologie retenue par les économistes et économètres afin de mesurer des effets « purs ». Si les femmes sont moins bien payées [1], c'est en partie parce qu'elles occupent des emplois peu qualifiés, parfois à temps partiel, dans les secteurs à faible niveau de salaires. Autrement dit, il y peut donc y avoir d'autres caractéristiques, distinctes du sexe, capables d'expliquer l'écart de salaire observé. L'écart de 30% évoqué à titre d'exemple n'est donc pas uniquement et intrinsèquement lié au fait d'être un homme ou une femme, mais plutôt résulte de l'agrégation de facteurs divers que nous devons prendre en compte dans nos régressions. Ceci s'inscrit dans la différence au combien cruciale entre corrélation et causalité. Si le fait d'être une femme entraîne un écart de salaire de l'ordre de 30%, il est clair que le fait d'être une femme est négativement corrélé au salaire. Seulement, cette corrélation n'implique pas une causalité complète dans la mesure où celle-ci masque d'autres effets sous-jacents eux-mêmes corrélés au fait d'être une femme et *a forciori* corrélés au salaire. Tout l'intérêt de la méthode *toutes choses égales par ailleurs* réside dans l'exploitation plus ou moins exhaustive de variables de contrôle afin d'isoler et de quantifier précisément l'effet causal du sexe sur l'écart de salaire. Cette méthode vise donc à réduire au maximum le biais de variables omises et contrôler l'endogénéité des variables ajoutées.

Pour estimer l'ampleur de la discrimination salariale sur un groupe d'individus, une stratégie classique consiste à regrouper d'un côté les caractéristiques des individus et de leurs travails ; de l'autre, celles ayant trait à l'appartenance au groupe. Dans ce cas, la part expliquée des variables d'appartenance au groupe est interprétée comme une conséquence directe d'une discrimination salariale, dans la mesure où les caractéristiques des individus et de leurs travails décrivent parfaitement les autres déterminants - non discriminatoires, du salaire. Seulement, cette situation est idéaliste dans la mesure où nous n'observons pas certaines variables aux effets significatifs sur le salaire (motivation, qualités intrinsèques de l'individu, machisme de la part de l'employeur, etc...) soit en majorité des effets fixes.

La problématique de ce sujet s'inscrit dans la mesure de l'effet potentiel de l'origine d'un individu sur sa position sur le marché du travail, et *a fortiori* son salaire. Les remarques soulevées au paragraphe précédent peuvent s'y généraliser. Cependant, intéressons-nous plus spécifiquement à l'intérêt d'inclure des variables de contrôle dans notre modèle de régression. Avant de basculer complètement dans notre problématique, considérons un dernier exemple : l'augmentation du plafond de remboursement des frais de santé. Supposons qu'un gouvernement s'intéresse aux dépenses de soins de ces citoyens et veuille les comparer à leur état de santé effectif. Celui-ci observera en premier lieu une causalité inverse : une santé dégradée accroît les dépenses de santé car il s'agit avant tout des individus en moins bonne santé qui dépensent de plus. Seulement, il est réaliste de constater que les classes sociales supérieures consenteront une part plus importante de leur revenu aux dépenses de soins, sans pour autant en avoir réellement besoin étant donné qu'elles exercent en moyenne des activités moins pénibles. Dans ce cas

1. En l'occurence, l'étude numéro 1436 publiée le 8 Mars 2013 par Thomas Morin et Nathan Remila conclut qu' « en 2010, dans le secteur privé, les femmes [avaient] un revenu salarial inférieur de 28 % à celui des hommes. »

particulier, l'augmentation des dépenses de santé n'est pas nécessairement corrélée à une dégradation de l'état de santé, mais plutôt du niveau de richesse. La politique visant à ré-hausser le plafond de remboursement des frais de santé profiterait donc, sans réel fondement, à une partie de la population. D'où l'importance d'adopter une analyse *toutes choses égales par ailleurs* fondée sur de nombreuses variables de contrôle, afin d'extraire le biais de corrélation et interpréter correctement l'effet causal des dépenses de santé.

Un décisionnaire gouvernemental sera donc bien plus intéressé par le pourquoi que le comment. Ceci s'applique dans le cadre des disciminations salariales liées à l'origine : inclure des variables de contrôle permet d'affiner la relation de cause à effet de l'origine sur le salaire horaire moyen. En ce sens, l'instauration d'aides publiques en faveur de la réduction des discriminations liées à l'origine sera davatange ciblée et donc plus efficace.

Nous pourrions néanmoins nous poser la question : pourquoi ne pas utiliser un modèle de régressions extrêmement riche afin de capter tous les effets non-discriminatoires expliquant les différences de salaire entre immigré et non-immigré ? Tout d'abord pour des raisons de significativité : certaines variables sous-représentées ne sont pas signficatives au sens de l'inférence statistique ce qui fragilise voire invalide leur interprétation. D'autre part, ajouter un trop grand nombre de variables peut causer des redondances (en terme d'échelle de regroupement) et complexifier outre mesure l'interprétation des résultats. Mais surtout, à cause d'un problème d'endogénéité. En effet, à vouloir ajouter trop de variables, nous augmentons le risque d'ajouter de l'endogénéité au modèle ce qui joue également sur le biais de corrélation.

Cette première question vise à construire un modèle simple qui sera ajusté dans les questions suivantes afin d'interpréter à première vue l'effet de la variable `origine` sur le salaire. Le tableau suivant résume l'ensemble des variables de contrôle que nous avons choisi d'inclure dans notre modèle. Celui-ci résulte d'une réflexion *a priori* des variables les plus susceptibles d'impacter le salaire. À ce stade de la lecture, la significativité des différentes variables de la base n'a pas été étudiée. Nous y reviendrons au besoin ultérieurement afin de supprimer certaines d'entre elles.

Tableau 6 – Premier modèle de régression du logarithme du salaire horaire

Catégories	**Variables**
Individuelles démographiques et socio-économiques	`cat_origine`, `age`, `age`^{2}, `taille_menage`, `typmen`, `matri`, `homme` `csp_parents`, `parents_act`, `sante_declaree`, `csp_actif`
Individuelles professionnelles	`encadre`, `plract`, `contrat`, `typhoraire`, `heuretra_tranche`, `expe_specifique` `expe_specifique`^{2}
Zone résidentielle	`pibhab_cat`, `tuu_r`, `typvois`, `zus`

Au total, il s'agira donc de régresser le logarithme du salaire mensuel `logsalhoraire` sur 19 variables. Nous justifions tout d'abord l'utilisation de variables géographiques car ces caractéristiques aggrégées influent différement sur les caractéristiques de nos individus. Il semble en effet réaliste de penser que les salaires d'individus seront corrélés les uns aux autres selon l'échelle géographique dans laquelle nous les considérons. Les caractéristiques des habitants d'un même village seront davantage dépendantes et leurs salaires respectifs seront *a fortiori* plus corrélés. Ne pas prendre en compte qu'un individu habite dans un pays d'Outre-Mer, zone particulièrement pauvre par rapport aux régions métropolitaines, entraînerait un biais de variables omises. Nous devons donc prendre en compte ces différentes échelles car celles-ci imposent des contraintes plus ou moins grandes sur le salaire des individus qui y vivent (nous aurions pu d'ailleurs considérer les taux de chômage régionaux moyens).

En ce qui concerne les données individuelles liées à la position sur le marché du travail, nous considérons que les variable `contrat` et `typehoraire` sont particulièrement intéressantes dans la mesure où nous pouvons imaginer que les individus immigrés ou descendants d'immigrés auront davantage tendance à occuper des postes moins qualifiés aux horaires variables, probablement en CDD ou en intérim. De surcroît, il est d'autant plus plausible de supposer que ces derniers seront plus enclin à occuper plusieurs activités simultanément afin de compléter leurs revenus. C'est par exemple le cas dans le domaine du bâtiment où la majorité des ouvriers sont embauchés pour la durée du chantier et sont souvent payés en-dessous du salaire minimum (on peut penser aux travailleurs détachés).

Les variables `encadre` et `expe_specifique` nous seront également utiles puisqu'il sera intéressant de savoir si, à expérience spécifique et niveau d'encadrement égaux, les salaires touchés par un descendant d'immigré sont inférieurs. Nous choisissons l'expérience spécifique car celle-ci est *a priori* moins endogène que l'expérience réelle (nombre de mois travaillés en France depuis la fin des études). En effet, si nous assimilons le terme des résidus (nous y reviendrons) à la productivité et aux qualités intrinsèques de l'individu, un individu faiblement productif aura davantage tendance à écourter sa carrière, traverser des périodes de chômage ou d'inactivité et ne pourra pas prétendre à un salaire élevé. Contrairement à l'expérience spécifique qui elle ne tient en compte que de l'expérience au poste actuel.

Enfin, en ce qui concerne les variables individuelles démographiques et socio-économiques, nous justifions l'utilisation de la variable `homme` par le fait que les dynamiques des marchés du travail pour les hommes et les femmes sont différentes. En ce sens, hommes et femmes présentent des préférences et des aspirations différentes face au travail. Comme nous l'avons décrit, le sexe est également un facteur particulièrement discriminatoire du salaire. Nous aurions pu envisager systématiquement deux modèles pour comparer les influences de la variable `cat_origine` sur le logarithme du salaire horaire. Ensuite, les variables `parents_act` et `csp_parents` restent un bon contrôle du milieu social d'origine. Enfin, l'utilisation de la variable `csp_actif` est délicate. D'un côté, il est certain que les différences de salaire sont expliquées par le secteur professionnel. Cependant, il est également probable que cette variable soit corrélée au terme d'erreur. Cette idée rejoint la notion d'"assortative matching" : chaque secteur d'emploi attire une catégorie spécifique de travailleurs déterminée par leur productivité. Ainsi, un individu peu productif se dirigera vers certaines filières plutôt que d'autres, et ne touchera pas le même salaire au final. Nous pouvons appliquer ce raisonnement à de nombreuses variables caractérisant le travail et le salaire des individus. Le rôle de la structure de panel est indispensable dans la réduction des phénomènes d'endogénéité.

Suite à notre première régression, nous avons décidé d'enlever les variables suivantes :

- `taille_menage` ($p - valeur =$ 0,912) ;
- `expe_specifique`^{2} ($p - valeur =$ 0,022, $\beta = -3,14 \times 10^{-7}$) et telle que l'expérience spécifique d'inversion de l'effet de l'expérience spécifique sur le salaire est de 100 ans... ;
- `parents_act` ($p - valeur =$ 0,033, $\beta = 7,35 \times 10^{-3}$) ;
- `matri` ($p - valeur =$ -0,008, $\beta = -5,71 \times 10^{-3}$: l'effet du régime matrimonial est capté par la variable `type_menage`.).

De plus, la statistique de test de Breusch-Pagan / Cook-Weisberg est de 193.05, ce qui rejette très fortement l'hypothèse d'homoscédasticité $\left(\forall\, i,\ \ \mathbb{V}(\epsilon_i|X_i) = \sigma^2\right)$. Nous utiliserons donc l'instruction `robust` (si c'est possible). Vous trouverez finalement les résultats de notre première régression à la page 9, tableau 7.

La grande majorité de nos variables de contrôle sont significatives, ceci vient appuyer la pertinence et la véracité de nos interprétations. Les signes des coefficients des variables sont globalement tous en accord avec nos statistiques descriptives et/ou nos intuitions, nous ne commenterons donc pas les coefficients des variables `pibhab_cat`, `tuu_r`, `typ_vois`, `zus`, `typmen`, `csp_parents`, `csp_actif`, `contrat` et `sante_declaree`. En revanche, il est important de commenter l'effet de l'âge sur le salaire horaire. Comme attendu, le salaire dépend quadratiquement de l'âge. En faisant un simple calcul de dérivée, nous constatons que l'âge auquel l'effet de cette variable sur le salaire s'inverse est de 57 ans, ce qui correspond presque à l'âge moyen de départ en retraite en 2015 en France, *id est* 63 ans[1]. Seulement, les individus retraités ne perçoivent plus de salaires (valeurs manquantes). Ceci veut dire qu'à partir de 57 ans, l'âge influence négativement le salaire des actifs occupés. On peut penser aux ré-aménagements d'horaires, aux reticences des employeurs à payer pleinement des employés âgés nouvellement arrivés (frais de formation, départs proches, etc...).

Ensuite, toutes choses égales par ailleurs, le fait d'être un homme plutôt qu'une femme augmente en moyenne le salaire horaire de 9,3%. L'expérience spécifique est telle qu'un an d'ancienneté au poste correspondant augmente en moyenne le salaire horaire de 0,68%, et donc que 10 ans d'ancienneté participent à la hausse de ce même salaire de l'ordre de 6,8%. La varible `plract` a un impact négatif sur le salaire. Ceci est logique dans la mesure où les individus ayant plusieurs travails auront tendance à être moins payés dans leur activité principale (moins disponibles, activité principale sous-remunérée, horaires flexibles). Le fait d'encadrer des salariés augmente le salaire horaire

1. Source : Cnav - SNSP

en moyenne de 5,5% (fonctions de manager liées à l'ancienneté, chef de projet, chef de groupe, etc...). Enfin, des horaires de travail stables offrent en moyenne des salaires moins élevés que pour des horaires alternés ou variables.

Concentrons-nous maintenant sur la variable `cat_origine` que nous avons créée. Le signe des coefficients rejoint notre analyse préliminaire. En effet, nous remarquons que le fait d'être immigré naturalisé ou non d'origine non-européenne contribue à une baisse moyenne de salaire de l'ordre de -3,9% et -7,5% par rapport à la situation où l'individu est français né de deux parents français. Étonnemment, les immigrés d'origine européenne ne souffrent pas de cet écart et sont en moyenne mieux payés que les individus de la catégorie *fra2pfra* ; ceci avec des coefficients significatifs. Enfin, le fait d'être français et d'avoir au moins un parent immigré (européen ou non) joue peu sur les différences de salaire.

En résumé, toutes choses égales par ailleurs, fait d'être immigré ou non caractérise de manière significative des différences de salaire aux signes différents. Bien que ces coefficients se soient pas très grands : entre -4% et 8% en moyenne par rapport à la catégorie de référence, ce qui correspond respectivement à une baisse de 73 euros sur le salaire mensuel moyen de la base (1 827 euros) et une hausse 146 euros ; les écarts existent. Ceci laisse présager d'inégalités salariales entre immigrés et français, dont le sens varie selon le fait d'avoir un profil européen ou non. Ceci se croise avec vos propos "on parlera d'inégalité salariale due à l'origine lorsque que toutes choses égales par ailleurs les salaires des individus diffèrent selon leur origine." La question est de savoir si l'utilisation du « *toutes choses égales par ailleurs* »est pertinente. En ce sens, ne pas intégrer le niveau de diplôme peut, par exemple, être source d'endogénéité de la variable `cat_origine`. Nous ne pouvons dès lors pas parler d'effet causal.

De plus, et c'est là l'inconvénient majeur de notre modélisation : certaines variables sont endogènes et rajoutent de l'endogénéité à notre modèle. La volonté d'intégrer des variables de contrôle se heurte à ce problème majeur. Nous l'avons dit, la variable `csp_actif` est vraisemblablement endogène. Le type de contrat l'est également. Il faut être prudent quant à l'utilisation de cette expression et nos interprétations. Ce problème d'endogénéité des variables renforce le besoin de prendre en compte la dimension panel de l'étude dans la mesure où celle-ci permet d'éliminer les effets individuels. Elle permet aussi de rendre compte d'une certaine dynamique. Nous pouvons effectivement pensé qu'un immigré aura, au cours de la période, pu mieux s'intégrer, su mieux valoriser ses compétences, mieux maîtriser le français, etc... et donc *a fortiori* prétendre à des postes aux salaires plus élevés. S'ajoute à ce problème d'endogénéité le fait que nous ne prenons en compte ici que les individus les plus productifs puisque le salaire est systématiquement observé. Il existe donc un biais supplémentaire (autre que le biais de variables omises) dû a un effet de sélection : nous mesurons mal les inégalités car nous n'observons que les personnes les plus productives. Le modèle de sélection généralisée permet de contourner ce problème majeur qui fragilise nos interprétations.

En dernier lieu, nous aurions pu effectuer une régression *backward stepwise* de manière à ne garder que les variables utiles au modèle. Cependant, ceci n'est pas possible avec des variables catégorielles sur Stata. De plus, l'intérêt de cette méthode est mineur et peut être substitué par une réflexion en amont, comme nous l'avons fait. En ce sens, nous aurions également pu calculer les AIC et BIC correspondants de manière à arbitrer correctement entre perte de précision des paramètres et gain du pouvoir explicatif. Nous n'avons pas jugé cela utile dans la mesure où ce projet ne se réduit pas à un problème de performance mais plutôt d'interprétabilité du modèle. Comme nous l'avons dit, l'idée ici dans le cadre de politiques publiques est d'ajouter le maximum de contrôle tout en s'interrogeant sur l'endogénéité des variables ajoutées.

Tableau 7 – Régression du logarithme du salaire sur la population d'actifs

	(1) logsalhoraire	(s.e)
pibhab_cat		
.1	ref.	(.)
.2	-0.063***	(0.011)
.3	-0.102***	(0.011)
tuu_r		
- *Commune rurale*	ref.	(.)
- *Uu < 20 000 habitants*	-0.011*	(0.005)
- *Uu < 200 000 habitants*	0.018***	(0.005)
- *Uu > 200 000 habitants*	0.028***	(0.005)
- *Agglo. parisienne*	0.021	(0.011)
typvois		
- *Hors agglo.*	ref.	(.)
- *Lotissement, pavillons*	-0.002	(0.004)
- *Immeubles en ville*	-0.027***	(0.006)
- *Immeuble ou cité*	-0.050***	(0.008)
- *Habitat mixte*	-0.012	(0.009)
zus	-0.045***	(0.009)
typmen		
- *1 personne*	-0.022***	(0.005)
- *Famille monop.*	-0.024***	(0.006)
- *Couple sans enfant*	-0.027***	(0.004)
- *Couple avec enfant(s)*	ref.	(.)
- *Ménages (av. parenté)*	-0.023	(0.026)
- *Ménages (ss. parenté)*	-0.071**	(0.025)
- *Autres*	-0.108***	(0.013)
age	0.019***	(0.000)
age^2	$-2,084 \times 10^{-4}$***	(0.000)
homme	0.093***	(0.004)
csp_parents		
- *Non renseigné*	-0.005	(0.008)
- *Agriculteurs expl.*	-0.000	(0.006)
- *Artisans, com., c.e.*	0.038***	(0.005)
- *Cadres*	0.081***	(0.006)
- *Professions int.*	0.032***	(0.005)
- *Employés*	0.021***	(0.005)
- *Ouvriers*	ref.	(.)

	(1) logsalhoraire	(s.e)
cat_origine		
- *fra1pimmi_Neurope*	0.016	(0.010)
- *fra1pimmi_europe*	0.017**	(0.007)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.073***	(0.012)
- *immi_Nnatur_europe*	0.076***	(0.011)
- *immi_natur_Neurope*	-0.043***	(0.011)
- *immi_natur_europe*	0.065***	(0.016)
csp_actif		
- *Non renseigné*	ref.	(-)
- *Cadres*	0.320***	(0.052)
- *Professions int.*	0.034	(0.052)
- *Employés*	-0.19***	(0.052)
- *Ouvriers*	-0.16**	(0.052)
contrat		
- *Pas de contrat*	0.047***	(0.005)
- *CDI*	ref.	(.)
- *CDD*	-0.110***	(0.007)
- *Contrat saisonnier*	-0.110***	(0.025)
- *Interim*	0.035	(0.035)
- *Apprentissage*	-0.36***	(0.027)
plract	-0.090***	(0.008)
encadre	0.069***	(0.004)
expe_specifique	0.00063***	(0.000)
typehoraire		
- *Stable*	ref.	(.)
- *Alterné*	0.088***	(0.005)
- *Variable*	0.008	(0.004)
- *Sans objet*	0.0031	(0.057)
heuretra_tranche		
- *Moins de 15h*	ref.	(.)
- *Entre 15 et 30h*	-0.120***	(0.018)
- *Entre 30 et 35h*	-0.110***	(0.018)
- *Entre 35 et 40h*	-0.155***	(0.017)
- *40h et plus*	-0.233***	(0.018)
sante_declaree	-0.023***	(0.002)
Nombre d'observations	40 403	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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

La variable `education` caractérisant le niveau de diplôme est endogène, et cette assertion rejoint la théorie du capital humain. Il est raisonnable d'affirmer que le salaire d'un individu est positivement corrélé à sa productivité. Cette productivité est elle-même intrinsèquement liée à l'aptitude individuelle (capacité d'apprentissage et de mémorisation, efficacité dans l'organisation, facultés cognitives innées, etc...). Piketty (2001) la définit comme étant « très générale, puisqu'elle inclut les qualifications proprement dites (diplômes, etc...), l'expérience et, plus généralement, toutes les caractéristiques individuelles qui ont un impact sur la capacité à s'intégrer au processus de production». Nous pouvons donc intuiter que les individus les plus productifs toucheront en moyenne des salaires plus élevés. Ces mêmes individus étant par ailleurs les plus susceptibles d'entreprendre des études longues et techniques. En conclusion, le niveau de diplôme, au travers de la variable `education`, est corrélé aux facultés intrinsèques de l'individu ($\equiv$ résidus) qui déterminent, par l'intermédiaire de la productivité, le salaire.

Du fait de l'endogénéité de la variable `education`, il est nécessaire de trouver un ou plusieurs instruments vérifiant les hypothèses de pertinence et d'exclusion. Étant donné que l'hypothèse d'exclusion n'est pas testable pour un unique instrument, nous nous sommes intéressés à la recherche de deux instruments afin de réaliser le test de sur-identification de Sargan. Au vu des variables disponibles dans la base, nos intuitions nous ont amené à considérer les variables `tuu_r` et `csp_parents`. Premièrement, nous justifions le choix de la variable instrumentale `tuu_r` par l'hypothèse qu'un individu aura tendance à habiter dans un type d'unité urbaine semblable à celle dans laquelle il a vécu étant jeune. Le fait d'être isolé (géographiquement parlant) freine l'accès à des formations plus techniques ou plus renommées. Les individus vivant dans un petit village mal desservi seront moins enclin à faire des études longues ou ne seront pas poussés à le faire, contrairement à un citadin n'étant qu'à quelques minutes d'une faculté ou évoluant dans un milieu social plus aisé et plus porteur. De ce fait, la localisation actuelle, puisque déterminée par la localisation passée (hypothèse pas toujours vérifiée), influe sur le niveau de diplôme. Seulement, il n'est pas certain qu'un individu soit moins productif s'il vient d'un village : la corrélation est obscure. Pour cela, nous devons revenir à la définition même des qualités intrinsèques de l'individu : sont-elles innées ou exprimées durant l'enfance ? A priori les deux. Nous pouvons supposer qu'en présence d'un plein accès aux études supérieures, un lycéen pourra prétendre à n'importe quel type d'étude (nous entrons dans des considérations quasi politiques...). Cette affirmation est bien sûr à nuancer dans la mesure où le milieu social d'origine va influer sur l'individu et par ce biais développer des aptitudes plus ou moins compatibles avec le monde du travail (un jeune issu d'un environnement très rural aura peu de chance d'être incité à faire de l'informatique, donc une école d'informatique).

Le second instrument que nous proposons est la variable `csp_parents`. Celle-ci réfère à l'exemple classique des variables instrumentales de l'éducation appelées communément `fathereduc` et `mothereduc`. Le domaine professionnel des parents influence le niveau de diplôme de l'enfant dans la mesure où le contexte familial lui permet d'affirmer ses aspirations professionnelles, de s'orienter vers telles ou telles études. Un individu ayant des parents cadres supérieurs sera d'autant plus poussé et enclin à vouloir faire des études longues et techniques. Seulement, la productivité intrinsèque de l'individu n'est pas entièrement déterminée par le niveau d'études des parents (et *a fortiori* les métiers des parents). Encore une fois, cette affirmation est discutable.

Afin d'étudier la validité de nos instruments, nous avons procédé aux tests des hypothèses de pertinence et de compatibilité des instruments. La p-valeur du test de Sargan (régression de `logsalhoraire` sur nos précédentes variables et la variable `education` instrumentée par `csp_parents` et `tuu_r`) est égale à 0,0012. On rejette donc l'hypothèse de sur-identification au seuil de 1%, mais pas au seuil de 0,1%. La statistique de Fisher de nullité jointe des instruments dans la régression de la variable `education` est quant à elle égale à 224, on rejette donc fortement la nullité jointe des instruments, ceux-ci sont donc pertinents. Étant donné que les instruments de peuvent pas figurer dans la régression du logarithme du salaire, nous n'utiliserons qu'un seul instrument, `csp_parents` afin de ne pas perdre le niveau "unité urbaine" de la régression. Vous trouverez en page 11 la sortie de la régression instrumentale du logarithme du salaire. En page 12, nous superposons les coefficients associés aux modalités de la variable `cat_origine` selon que la régression inclut ou non la variable `education`, et si elle inclut la variable `education` mais pas la variable `csp_actif`.

Tableau 8 – Régression instrumentale du logarithme du salaire sur la population d'actifs

	(1) logsalhoraire	(s.e)
pibhab_cat		
.1	ref.	(.)
.2	-0.069***	(0.011)
.3	-0.11***	(0.011)
tuu_r		
- *Commune rurale*	ref.	(.)
- *Uu < 20 000 habitants*	-0.009*	(0.004)
- *Uu < 200 000 habitants*	0.011*	(0.005)
- *Uu > 200 000 habitants*	0.015**	(0.006)
- *Agglo. parisienne*	0.019	(0.012)
typvois		
- *Hors agglo.*	ref.	(.)
- *Lotissement, pavillons*	0.002	(0.004)
- *Immeubles en ville*	-0.025***	(0.006)
- *Immeuble ou cité*	-0.029***	(0.008)
- *Habitat mixte*	-0.009	(0.008)
zus	-0.023*	(0.010)
typmen		
- *1 personne*	-0.025***	(0.005)
- *Famille monop.*	-0.012	(0.006)
- *Couple sans enfant*	-0.020***	(0.004)
- *Couple avec enfant(s)*	ref.	(.)
- *Ménages (av. parenté)*	-0.010	(0.025)
- *Ménages (ss. parenté)*	-0.054*	(0.023)
- *Autres*	-0.084***	(0.016)
age	0.004***	(0.001)
age^2	-1.8×10^{-4}***	(0.000)
homme	0.14***	(0.005)
education	-0.087 ***	(0.015)
cat_origine		
- *fra1pimmi_Neurope*	0.011	(0.010)
- *fra1pimmi_europe*	0.014**	(0.007)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.042**	(0.014)
- *immi_Nnatur_europe*	0.101***	(0.014)
- *immi_natur_Neurope*	-0.050***	(0.011)
- *immi_natur_europe*	0.051**	(0.018)
csp_actif		
- *Non renseigné*	ref.	(.)
- *Cadres*	0.20***	(0.054)
- *Professions int.*	-0.013	(0.049)
- *Employés*	-0.157***	(0.048)
- *Ouvriers*	-0.092	(0.049)
contrat		
- *Pas de contrat*	0.016*	(0.007)
- *CDI*	ref.	(.)
- *CDD*	-0.103***	(0.007)
- *Contrat saisonnier*	-0.087**	(0.027)
- *Interim*	0.054	(0.037)
- *Apprentissage*	-0.35***	(0.025)
plract	-0.089***	(0.008)
encadre	0.077***	(0.004)
expe_specifique	0.00068***	(0.000)
typehoraire		
- *Stable*	ref.	(.)
- *Alterné*	0.098***	(0.005)
- *Variable*	0.010*	(0.004)
- *Sans objet*	-0.001	(0.054)
heuretra_tranche		
- *Moins de 15h*	ref.	(.)
- *Entre 15 et 30h*	-0.154***	(0.019)
- *Entre 30 et 35h*	-0.157***	(0.020)
- *Entre 35 et 40h*	-0.2***	(0.019)
- *40h et plus*	-0.284***	(0.020)
sante_declaree	-0.016***	(0.002)
Nombre d'observations	40 403	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tableau 9 – Comparaison des coefficients de la variable `cat_origine` en fonction de l'instrumentation de la variable `education`

	(MCO sans `education`) logsalhoraire	(s.e)
cat_origine		
- *fra1pimmi_Neurope*	0.016	(0.010)
- *fra1pimmi_europe*	0.017**	(0.007)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.073***	(0.012)
- *immi_Nnatur_europe*	0.076***	(0.011)
- *immi_natur_Neurope*	-0.043***	(0.011)
- *immi_natur_europe*	0.065***	(0.016)
Nombre d'observations	40 403	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(IV `education`) logsalhoraire	(s.e)
cat_origine		
- *fra1pimmi_Neurope*	0.011	(0.010)
- *fra1pimmi_europe*	0.014**	(0.007)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.042**	(0.014)
- *immi_Nnatur_europe*	0.101***	(0.014)
- *immi_natur_Neurope*	-0.050***	(0.011)
- *immi_natur_europe*	0.051**	(0.018)
Nombre d'observations	40 403	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(MCO sans `csp_actif`) logsalhoraire	(s.e)
cat_origine		
- *fra1pimmi_Neurope*	.027**	(0.011)
- *fra1pimmi_europe*	0.025***	(0.008)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.131***	(0.013)
- *immi_Nnatur_europe*	0.048***	(0.012)
- *immi_natur_Neurope*	-0.081***	(0.012)
- *immi_natur_europe*	0.073***	(0.018)
Nombre d'observations	40 403	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Quelles conclusions pouvons-nous donc tirer de ces 3 tableaux de régressions au regard de notre problématique ? Tout d'abord, nous observons que le biais de variable omise (en l'occurrence la variable omise ici est `education` dans la régression "MCO sans `education`", tableau 9) varie selon les modalités de la variable `cat_origine`. En effet, les coefficients des modalités *immi_Nnatur_Neurope*, *immi_Nnatur_europe* sont inférieurs en l'absence de la variable de diplôme (biais négatif) tandis que les coefficients des autres modalités sont supérieurs en l'absence de la variable de diplôme (biais positif). L'interprétation du biais est donc assez floue. Néanmoins, les tendances restent les mêmes : être immigré naturalisé ou non et d'origine extérieure à l'Europe implique un écart moyen négatif de salaire par rapport à la situation d'être français né en France de parents français. Le fait que ces modalités aient des coefficients toujours négatifs et ce malgré notre prise en compte de la variable de diplôme, nous montre que les diplômes étrangers sont moins bien valorisés que ceux des français. En effet, toutes choses égales par ailleurs, c'est-à-dire à niveaux de diplômes équivalents, un immigré non-naturalisé (ou naturalisé) touchera en moyenne moins qu'un français. D'ailleurs, cette affirmation semble être confirmée par le signe positif du coefficient *fra1pimmi_Neurope* : un individu ayant grandi en France et fait ses études en France, bien qu'ayant des origines non-européennes, touche en moyenne un salaire plus élevé qu'un individu immigré ayant les mêmes origines. De fait, il semblerait bel et bien que les diplômes nationaux soient mieux valorisés. Cependant, nous pouvons également penser que les individus immigrés ne sont pas tous des immigrés politiques ou immigrés de guerre (susceptibles d'avoir un niveau de diplôme élevé) mais bien des immigrés fuyant la pauvreté, la famine, les conditions météorologiques extrêmes et n'ayant de surcroît pas un bagage suffisant (voire inexistant) pour prétendre à des postes français aux salaires plus élevés.

De plus, nous remarquons que la significativité des modalités de la variable `cat_origine` diminue (globalement) avec l'introduction de la variable instrumentée d'éducation. Ceci semble donc bien signifier que l'effet de l'origine est capté par l'éducation en elle-même, ce qui motive l'introduction de cette variable dans l'optique de capter des effets « purs »de l'immigration sur le salaire, si nous l'instrumentons !

Enfin, prendre en compte l'éducation mais pas la variable `csp_actif` conduit également à de forts biais, positifs ou négatifs, sur les modalités de la variable `cat_origine`. De plus, ne pas inclure cette variable rend la variable d'origine plus significative ce qui montre encore une fois que l'origine est partiellement captée par le secteur professionnel de travail. Ne pas inclure la variable `csp_actif` conduirait donc également à un biais de variable omise. Cependant, l'inclure est également source d'endogénéité.

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

L'introduction d'un effet fixe individuel α_i a pour but de capter tous les facteurs inobservés constants dans le temps. Sans se préoccuper pour l'instant des hypothèses sur les résidus, la première difficulté réside dans le fait qu'il sera difficile de séparer l'effet des variables X_{it} constantes dans le temps et de α_i. Or, la variable `origine` qui caractérise l'origine des parents **à leur naissance**, est, par définition, constante dans le temps...

On ne peut donc pas utiliser les données de panel pour étudier l'impact de la variable origine sur le salaire. En effet, on ne peut réaliser de régression sur une variable de contrôle ne variant pas dans le temps.Une possibilité alternative aurait pu être de réaliser cette régression panel sur la variable `nat`. Une telle régression expliquerait l'impact de la volonté d'intégration sur la productivité, caractère acquis d'un individu au contact de son environnement. Mais elle n'expliquerait pas l'impact intrinsèque de l'origine géographique sur la productivité.

(d) On dispose également de variables à l'échelle de la zone résidentielle mesurant les proportions pour plusieurs groupes minoritaires définis par la variable "origine". Est-ce que celles-ci [`prop_*_immi` et `prop_*`] varient au cours du temps ? Estimer un modèle panel en incluant ces variables de contrôle. Interpréter les résultats obtenus : cette estimation peut-elle apporter des éléments de réponses pour notre problématique ?

Comme suggéré dans l'énoncé, nous n'avons pas pris en considération toutes les variables de proportions des groupes minoritaires. Nous n'avons en effet considéré que les variables `prop_n01`, `prop_n01_immi` (immigré au sens large), `prop_678910`, `prop_678910_immi` (immigration depuis l'extérieur de l'Europe). Nous avons également ajouté nos deux variables `prop_02` et `prop_02_immi` (immigration depuis l'intérieur de l'Europe), ainsi qu'une dernière variable `prop_01`, désignant la proportion de français (collinéaire aux précédentes) nés en France de parents français.

Ces variables évoluent bien au cours du temps. D'une part, cette affirmation est motivée par nos observations sur la base (confère la partie stata). À titre d'exemple, nous nous sommes intéressés à l'évolution de plusieurs proportions au cours des 6 moments d'interrogation, ceci pour les actifs occupés. Pour cela, nous avons choisi de représenter la variation moyenne des proportions entre chaque période (en pourcentage), en omettant les individus pour lesquels ces proportions étaient constamment égales à 0. Sans prendre en compte cet aspect, les variations moyennes seraient très proches de 0%, étant donné le nombre non-négligeable d'individus concernés. Le tableau 10 recense ces résultats. On observe en moyenne des pourcentages d'évolutions non-nuls de l'ordre de 2%, ce qui est peu si l'on raisonne en terme de nombre de personnes nouvelles / parties (étant donné que la moyenne de la variable `nb_ind_for_prop` est égale à 34, ces valeurs varient de 0 à 4 en moyenne). Il est intéressant de constater que, dès lors que l'on considère dans leur globalité uniquement des individus ayant toujours eu des voisins immigrés (i.e. à toutes dates d'interrogation), les tendances sont à la hausse, traduisant un regroupement croissant des minorités entre elles. D'autre part il semble réaliste que, si pour chaque individu, les variables `nb_ind_for_prop` varient à chaque date d'interrogation, les valeurs des proportions ne peuvent être strictement les mêmes.

Tableau 10 – Variabilité temporelle des proportions

Sur 28 714 observations [1]		**Pourcentage moyen d'évolution**				
Proportions	**Nulles** [2]	**rim1 → rim2**	**rim2 → rim3**	**rim3 → rim4**	**rim4 → rim5**	**rim5 → rim6**
`prop_678910`	12 006	3,43%	3,63%	2,41%	2,90%	2,56%
`prop_02`	6 124	3,82%	3,32%	2,24%	1,55%	1,53%
`prop_n01`	3 389	3,00%	2,51%	1,76%	1,75%	2,81%
`prop_678910_immi`	16 132	3,74%	2,43%	2,42%	2,32%	6,40%
`prop_02_immi`	14 564	3,63%	2,58%	1,93%	2,57%	3,08%
`prop_n01_immi`	9 142	3,59%	3,13%	1,14%	3,36%	4,35%
`prop_01`	0	0,27%	0,27%	0,17%	0,18%	0,21%

[1] 28 714 actifs occupés aux 6 dates dates d'interrogation, 6 exclusivement.

[2] Nombre d'individus n'ayant jamais été voisin de populations minoritaires (donc quelque soit le rand d'interrogation).

L'estimation d'un modèle de panel suppose de se placer dans un cadre d'hypothèses. En l'occurence, trois cas sont à distinguer (comme évoqué en partie 2) : "exogénéité des résidus mais autocorrélation", "exogénéité stricte" et "exogénéité faible". Nous pouvons d'ores et déjà [1] nous interroger sur la validité de ces hypothèses. Premièrement, l'hypothèse d'exogénéité

$$\mathbb{E}(X_{it}\alpha_i) = 0 \ \forall t\,, i$$

semble trop forte. En effet, cela reviendrait à supposer que les facultés intrinsèques des individus sont décorrélées de la mixité sociale dans laquelle ils vivent. Or, nous avons vu au travers de nos statistiques descriptives que la part de chômeurs actifs et inactifs (et $afortiori$ les faibles salaires) est plus importante dans les immeubles et cités, zone résidentielle dans lesquelles les proportions de minorité sont les plus fortes. En d'autres termes, un individu peu productif, vivant en ville (français ou non), du fait de son faible salaire, aura tendance à habiter dans un immeuble (de cité ou pas) et de fait sera davantage entouré de minorités. Pour un individu peu productif vivant en milieu rural, cette remarque est moins pertinente dans la mesure où ce milieu regroupe peu d'habitats de type grands ensembles et cités. Néanmoins cette hypothèse demeure restrictive. Il est plus réaliste de supposer la corrélation entre X_{it} et α_i. À la question du type d'exogénéité, il s'avère que les deux hypothèses d'exogénéité "stricte" ou "faible" sont questionnables. Dans la partie 2 (question a), nous expliquons pourquoi nous supposons vraie l'hypothèse d'exogénéité forte. Ici, dans le cas du salaire, il s'avère que le test d'exogénéité stricte n'est pas concluant. En effet, le test de nullité jointe des variables associées au coefficient β_1 dans la régression :

$$\Delta(logsalhoraire)_2 = \Delta X_2'\beta_0 + X_2'\beta_1 + \Delta\epsilon_2$$

avec $X =$ (`age typmen education csp_actif contrat plract encadre expe_specifique typehoraire sante_declaree prop_678910_immi prop_02_immi prop_678910 prop_02`) ; donne une valeur

$$F(14, 17137) = 4.42$$

soit une p-valeur très faible de l'ordre de 10^{-5}. Nous rejetons donc la nullité du coefficient β_1 et donc l'hypothèse d'exogénéité stricte. Seulement, supposer l'hypothèse d'exogénéité faible conduit à la faillite des estimateurs *within* et *first-differences*. Il est possible alors de procéder à l'instrumentation de ΔX_2 par la variable X_1 dans l'équation de différences premières. Sous cette hypothèse, il est possible d'obtenir un estimateur convergent de β_0 dans la régression en différences premières par 2MC. De plus, lorsque nous intégrons uniquement dans la variable X_2 (équation de différences premières) les variables de proportions, nous trouvons cette fois-ci un F-test égal à :

$$F(4, 17137) = 1.15$$

soit une p-valeur de 0,33. De fait, les variables de proportions de minorités sont strictement exogènes, il n'y a pas besoin de les instrumenter. En revanche, dès que nous ajoutions une variable dans le test de Fisher (autre que proportions) le test était rejeté. Les résultats sont disponibles en page 15.

1. Cette réponse sera davantage détaillée dans la partie 2, question a).

Graphique 1 – Régression instrumentée de $\Delta(logsalhoraire)_2$ par méthode 2MC sous hypothèse d'exogénéité faible

```
Instrumental variables (2SLS) regression        Number of obs     =     17,088
                                                F(13, 17074)      =       0.86
                                                Prob > F          =     0.5922
                                                R-squared         =          .
                                                Root MSE          =     .71836
```

d_logsalhoraire	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
d_age	1.789739	1.84008	0.97	0.331	-1.817008	5.396486
d_typmen	-.1233956	.3220043	-0.38	0.702	-.7545572	.507766
d_csp_actif	-.72907	.5745956	-1.27	0.205	-1.855336	.3971965
d_contrat	-.100886	.0500396	-2.02	0.044	-.1989688	-.0028032
d_plract	-.1139385	.1417865	-0.80	0.422	-.3918546	.1639776
d_encadre	1.773853	1.443181	1.23	0.219	-1.054931	4.602636
d_expe_specifique	.0005656	.0093689	0.06	0.952	-.0177985	.0189297
d_typehoraire	-.0586943	.0546767	-1.07	0.283	-.1658662	.0484776
d_sante_declaree	.0088536	.0214513	0.41	0.680	-.0331931	.0509003
prop_678910_immi	-.0866084	.2218842	-0.39	0.696	-.5215242	.3483075
prop_02_immi	-.3563171	.3641586	-0.98	0.328	-1.070106	.3574713
prop_678910	.0918317	.129609	0.71	0.479	-.1622152	.3458787
prop_02	.1529813	.1497045	1.02	0.307	-.1404549	.4464175
_cons	-2.034289	2.035597	-1.00	0.318	-6.024268	1.955691

```
Instrumented:  d_age d_typmen d_csp_actif d_contrat d_plract d_encadre
               d_expe_specifique d_typehoraire d_sante_declaree
Instruments:   prop_678910_immi prop_02_immi prop_678910 prop_02 lag_age
               lag_typmen lag_csp_actif lag_contrat lag_plract lag_encadre
               lag_expe_specifique lag_typehoraire lag_sante_declaree
```

Le tableau ci-dessus est difficilement exploitable car aucune variable n'est significative à 5% excepté le type de contrat. Nous ne pouvons donc pas rejeter l'absence d'effet des variables proportions de minorités sur le logarithme du salaire horaire. Ceci étant, nous constatons qu'une hausse entre les deux périodes de 10% de la proportion d'immigrés Européens fait augmenter en moyenne la différence de salaire horaire de 1,5% ce qui est considérable. À l'inverse, une hausse entre les deux périodes de 10% de la proportion d'Européens fait diminuer en moyenne la différence de salaire horaire de 3,5%. Ces chiffres passent, dans les mêmes conditions d'augmentations, à -0,87% et +0,92% selon qu'il s'agit des proportions de non-européens immigrés ou non. Le fait qu'il s'agisse de proportions d'immigrés conduit, dans les deux cas (européen ou non), à des coefficients négatifs. Les ordres de grandeur mis en jeu sont faibles et les variables non significatives ; à ce stade, il est donc difficile de pouvoir conclure vis-à-vis de notre problématique. De plus, certains écarts-types sont énormes puisque les variations entre $t = 1$ et $t = 2$ de ces variables sont souvent nulles, c'est le cas par exemple de la variable `encadr`. Essayons de voir si la prise en compte, même grossière, de l'hypothèse d'exogénéité stricte, donne des résultats plus significatifs.

Sous l'hypothèse d'exogénéité stricte, il est possible d'obtenir un estimateur convergent de β_0 via les transformations *within* et *first-difference*. En l'occurence, les estimateurs $\hat{\beta}_0^W$ $\hat{\beta}_0^{FD}$ coïncident dans le cas du salaire puisque celui-ci n'est relevé que deux fois ("T"$= 2$). Nous avons voulu voir comment se comporte le modèle sous cette hypothèse, bien que rejetée statistiquement. Dans ce cadre, nous supposerons **dans la suite de ce projet** également l'autocorrélation des erreurs (option `cluster` sur Stata) :

$$\mathbb{E}(\epsilon_{it}\epsilon_{jt}) \neq 0 \ \forall i \neq j$$

En effet, étant donné que l'étude interroge des individus dans des logements [1], la base regroupe donc des "paquets" de voisins. Il est alors naturel de dire que les chocs de salaire entre individus peuvent être corrélés. Ceci est d'autant plus vrai lorsque l'on considère des immeubles évidemment. Si une entreprise embauche un grand nombre

1. "L'enquête est réalisée en tirant aléatoire des logements dont on interroge, ou du moins cherche à interroger, les occupants".

d'inidividus de l'immeuble, ou si l'activité économique d'une zone entourant l'immeuble ralentit, les individus le ressentiront sur leur salaire, et ce de manière *a priori* corrélée. De manière générale, un choc macroéconomique peut avoir des répercussions sur les salaires d'individus voisins. La régression suivante est une régression *within* avec hypothèse de corrélation faible des erreurs. Si nous avions suspecté une corrélation forte des termes $\Delta\epsilon_s$ et $\Delta\epsilon_t$, nous aurions privilégié une méthode *first-differences*.

Tableau 11 – Régression panel du logarithme du salaire sur la population d'actifs

	(1) logsalhoraire	(s.e)
typmen		
- *1 personne*	ref.	()
- *Famille monop.*	0.014	(0.022)
- *Couple sans enfant*	0.012	(0.020)
- *Couple avec enfant(s)*	0.015	(0.021)
- *Ménages (av. parenté)*	-0.023	(0.071)
- *Ménages (ss. parenté)*	-0.021	(0.047)
- *Autres*	-0.010	(0.032)
age	0.065***	(0.009)
age^2	-0.00051***	(0.000)
contrat		
- *Pas de contrat de travail*	0.004	(0.038)
- *CDI*	ref.	(-)
- *CDD*	-0.030	(0.018)
- *Contrat saisonnier*	0.048	(0.066)
- *Interim*	0.118	(0.059)
- *Apprentissage*	-0.234**	(0.056)
plract	-0.005	(0.018)
encadre	0.042	(0.025)
expe_specifique	0.00006	(0.000)
typehoraire		
- *Stable*	ref.	(-)
- *Alterné*	0.0021	(0.020)
- *Variable*	-0.015	(0.012)
- *Non renseigné*	-0.153	(0.11)
heuretra_tranche		
- *Moins de 15h*	ref.	(.)
- *Entre 15 et 30h*	-0.343***	(0.046)
- *Entre 30 et 35h*	-0.493***	(0.052)
- *Entre 35 et 40h*	-0.500***	(0.051)
- *40h et plus*	-0.633***	(0.051)
sante_declaree	-0.003	(0.052)
Proportions minorités		
- *prop_678910_immi*	-0.305**	(0.0113)
- *prop_02_immi*	0.099	(0.122)
- *prop_678910*	0.170*	(0.082)
- *prop_02*	-0.064	(0.084)
Nombre d'observations	40 403	
corr(α_i, X_b)	-0.4803	
ρ	0.835	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Le premier constat marquant de cette régression est que la prise en compte de la dimension panel entraîne une baisse considérable de la significativité de l'ensemble de nos variables. Néanmoins, nous rejetons toujours largement l'hypothèse selon laquelle l'évolution du temps de travail n'a pas d'effet pas le salaire. Nous rejetons également fortement l'hypothèse selon laquelle le vieillissement n'a pas d'effet le salaire. Intéressons nous aux variables non significatives. Contrairement aux attentes, nous ne parvenons pas ici à rejeter l'hypothèse selon laquelle la composition du ménage (`typmen`) a un impact sur le log du salaire. Le divorce, l'arrivée d'un nouvel enfant, le départ d'un enfant devenu adulte ou encore le début d'un concubinage n'auraient pas d'impact significatif sur le salaire, du moins à court terme (3 ans). Nous pourrions effectivement nous demander ici si ce n'est pas un choc positif du salaire dans le passé qui aurait un impact sur la forme du ménage (i.e. promotion au travail donc je peux avoir un autre enfant / je peux divorcer car financièrement ce sera moins difficile). Le changement de situation du ménage interviendrait la plupart du temps lorsque la situation du couple/de la famille est stable financièrement.

Par exemple, quelqu'un qui a un nouvel enfant ou qui se sépare, va avoir tendance à sécuriser son salaire.

Parmi les variables significatives ($**$ ou $*$), on remarque que les proportions d'immigrés dans l'environnement résidentiel `prop_678910_immi` et `prop_678910` affectent différemment (en moyenne) le salaire des individus comparé à l'origine même de l'individu. La présence d'immigrés non-européens présente un impact négatif sur le salaire, mais cette fois-ci au travers des proportions des minorités. Les ordres de grandeur des coefficients ne sont cependant pas significatifs. En effet, si la proportion d'immigrés non-européens (`prop_678910_immi`) augmente de 10%, alors le salaire de l'individu concerné sera en moyenne 3% plus faible...

En conclusion, cette régression panel n'est pas exploitable, en terme d'ordres de grandeur et de valeurs des écarts-types, mais surtout en terme de significativité statistique. Rappelons également que nous avons supposé l'hypothèse d'exogénéité stricte en dépit de son rejet par notre test statistique. La prise en compte de l'exogénéité faible via un modèle 2MC et l'instrumentation de ΔX_2 conduit à des résultats non significatifs et les variables de proportions semblent peu affecter le salaire. Nous aurions pu considérer les données empilées pour étudier le biais de variables omises.

La partie suivante vise non plus à considérer le salaire (et *a fortiori* la population d'actifs) comme variable expliquée mais plutôt l'indicatrice d'être actif (et *a fortiori* à la fois les chômeurs et les actifs). Nous espérons à ce stade que les régressions panel seront plus pertinentes dans ce cadre.

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

Préambule

Comme nous l'avons dit au début de ce projet, l'étude des inégalités entre deux groupes d'individus nous amène à considérer des modèles de régression linéaire prenant en compte de nombreuses variables de contrôle. Pour par exemple mesurer les écarts de rémunération entre hommes et femmes, il est courant d'ajouter des variables de contrôle : le niveau de diplôme, la quantité de travail, le secteur d'activité, l'éducation des parents, etc... Ici nos variables principales sont les proportions de minorités présentes dans le voisinage des individus. Comme nous l'avons vu, il est souvent rare de parvenir à sélectionner les bonnes variables de contrôle du premier coup. C'est pourquoi certaines de celles utilisées à la partie d'avant ne pourraient plus figurer dans les modèles de cette deuxième partie. En effet, il est toujours pertinent d'effectuer des ajustements des variables de contrôle afin de faciliter l'interprétation et la lisibilité des résultats. Néanmoins, dans l'optique d'un raisonnement « *toutes choses égales par ailleurs* », le but est d'en ajouter un "maximum" afin de réduire les biais de variables omises tout en réduisant l'endogénéité du modèle, et *a fortiori* d'éviter toutes interprétations erronées. Une fois encore, les coefficients obtenus sur les variables de proportions peuvent varier sensiblement selon l'utilisation des variables de contrôle.

Ainsi, l'expression « *toutes choses égales par ailleurs* »est d'autant plus censée que l'on ajoute des variables pertinentes au modèle. Dans cette optique, nous avons souhaité incorporer aux modèles suivants de panel des variables catégorielles variables dans le temps [1]. Nous avons rapidement constaté que certaines variables ne sont disponibles que pour une catégorie des individus (à savoir, actif occupé ou chômeur). Par exemple, les variables `ancchom` et `halochomage` ne sont disponibles que pour les chômeurs. Inversement, les variables `contrat` ou `typehoraire` ne sont disponibles que pour les actifs occupés. Ces variables présentent donc des valeurs manquantes. Non pas car elles n'ont pas été collectées, mais parce que celles-ci n'ont pas de sens pour certaines catégories d'actifs sur le marché du travail (typiquement, l'ancienneté du chômage pour un un actif occupé). Nous aurions pu abandonner l'idée de ne pas les intégrer dans les modèles. Cependant, il paraît naturel d'établir une causalité plus ou moins forte entre le fait de retrouver un emploi et l'ancienneté du chômage. De même, certains types de contrat de travail favorisent le passage du statut d'actif occupé à chômeur. L'idée à laquelle nous avons pensée [2] est de procéder à une modélisation prédictive de ces variables. Seulement, cette approche n'a pas donné de résultats probants. Nous utilisions des instructions du type :

```
mi set mlong
mi register imputed contrat
mi impute pmm contrat region zus taille_menage age homme prop_03 prop_03_immi prop_04
prop_04_immi prop_05 prop_05_immi prop_06 prop_06_immi prop_07 prop_07_immi prop_08
prop_08_immi prop_09 prop_09_immi prop_10 prop_10_immi typmen dep tuu tuu_r typuu typvois
occupation_logement matri nat acteu_detail csp_actif_detaille education_detail2 cat_origine
csp_pere csp_mere, add(6) knn(3) rseed(2235)
```

ou

```
mi set mlong
mi register imputed contrat
mi impute mlogit contrat region zus taille_menage age homme prop_03 prop_03_immi prop_04
prop_04_immi prop_05 prop_05_immi prop_06 prop_06_immi prop_07 prop_07_immi prop_08
prop_08_immi prop_09 prop_09_immi prop_10 prop_10_immi typmen dep tuu tuu_r typuu typvois
occupation_logement matri nat acteu_detail csp_actif_detaille education_detail2 cat_origine
csp_pere csp_mere, add(6) rseed(2235)
```

1. **Pour la modélisation, nous avons supprimé les individus inactifs. Nous justifions ce choix fort dans le préambule de la partie 3.** De plus, et nous le verrons dans la troisième partie, dans la mesure où nous cherchons à déceler des inégalités liées à l'origine sur le marché du travail, il fait plus de sens de ne pas considérer les inactifs qui n'ont pas cherché à se porter sur le marché du travail.
2. Et que vous avez confirmée par e-mail.

La première instruction permet d'entraîner un algorithme des plus proches voisins sur les variables inscrites afin de compléter les valeurs manquantes de la variable `contrat` de telle sorte à obtenir des résultats du type « cet individu chômeur aurait été le plus susceptible d'avoir un contrat de travail "interim" s'il avait été actif ». Même chose avec la méthode de régression logistique multinomiale. Nous avons essayé d'optimiser les extraparamètres des modèles par validation croisée mais les taux d'erreurs restaient de l'ordre de 40%... Au-delà du manque de résultats probants, il aurait été plus judicieux de réaliser des régressions séparées entre individus actifs occupés et population de chômeurs afin d'exhiber les effets de ces variables en particulier. Nous avons également tenté de transformer certaines variables manquantes. Par exemple, attribuer aux actifs occupés une ancienneté du chômage égale à 0. Cependant, une telle démarche était erronée dans la mesure où certains chômeurs n'étaient au chômage que depuis moins d'un mois strictement. De même pour le type de contrat ou le type d'horaire. Le fait est que nous avons finalement choisi de ne pas inclure ces variables dans nos modèles de panel et d'utiliser celles valables à notre disposition.

L'utilisation de modèles de panel "naïfs" - comme c'est le cas dans le début de cette partie, implique la suppression d'un grand nombre de variables constantes dans le temps. Nous avons du également procéder à l'enlèvement de certaines variables. Nous vous proposons une nouvelle version réduite du tableau 6 (tableau 12).

Tableau 12 – Modèle de régression panel sur l'indicatrice d'être actif occupé

Catégories	**Variables**
Individuelles démographiques et socio-économiques	~~`cat_origine`~~, `age`, `age`^{2}, `typmen`, ~~`homme`~~ ~~`csp_parents`~~, `sante_declaree`, ~~`csp_actif`~~
Individuelles professionnelles	~~`encadre`~~, ~~`plract`~~, ~~`contrat`~~, ~~`typhoraire`~~, ~~`heuretra_tranche`~~, ~~`expe_specifique`~~ ~~`expe_specifique`~~^{2}
Zone résidentielle	~~`pibhab_cat`~~, ~~`tuu_r`~~, ~~`typvois`~~, ~~`zus`~~[1]

[1] Les individus ne déménagent pas au cours de la période.

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

Pour un modèle de base :

$$Y_{it} = X'_{it}\beta_0 + \nu_{it} \quad \text{pour } i = 1...n \text{ et } t = 1...T$$

avec

$$\nu_{it} = \alpha_i + \epsilon_{it} \ ,$$

l'hypothèse d'exogénéité stricte s'écrit :

$$\mathbb{E}(X_{it}\alpha_i) \neq 0 \text{ mais } \mathbb{E}(X_{it}\epsilon_{it'}) = 0 \ \ \forall (t, t') \ ,$$

et l'hypothèse d'exogénéité faible :

$$\mathbb{E}(X_{it}\alpha_i) \neq 0 \text{ mais } \mathbb{E}(X_{it}\epsilon_{it'}) = 0 \ \ \forall t' \geq t$$

Dans notre cas, X_{it} caractérise principalement les proportions de minorités présentes autour de l'individu concerné. ϵ_{it} lui caractérise les chocs inobservables d'employabilité variables au cours de la période d'interrogation (de $t = 1$ à $t = 6$) d'employabilité. Discutons tout d'abord l'hypothèse d'exogénéité stricte. Celle-ci signifie que les proportions de minorités (et autres variables de contrôle dépendantes du temps et des individus) ne sont pas corrélées aux chocs passés et futurs d'employabilité. De manière générale, on peut supposer que le fait de devenir chômeur entraîne plus de précarité sociale : risque de déménagement, risque de divorce, problèmes financiers, etc... Cette précarité se retrouve davantage dans les zones urbaines sensibles, banlieues, cités. Il n'est pas politiquement

incorrect de constater que les banlieues péri-urbaines (ou immeubles en ville !) concentrent davantage de minorités (d'où le problème du communautarisme) et de chômeurs. Ainsi, un individu nouvellement chômeur aura tendance à déménager dans des zones socialement plus mixtes.

On pourrait donc supposer que le choc passé d'employabilité a une conséquence sur les proportions de minorités, dans le sens où celles-ci augmentent. Seulement, la période d'étude est de six trimestres. De plus, une période de chômage, courte ou potentiellement longue, n'implique pas immédiatement le basculement dans la précarité sociale. S'il y a corrélation, alors cette corrélation est faible. Qu'en est-il de la corrélation entre les proportions de minorités et les chocs futurs ? Si la France venait à entrer dans une vague de récession ou de forte croissance sur le court terme, cela représenterait un choc futur d'employabilité. Dans le cas d'une forte croissance annoncée, l'on peut imaginer que l'individu chômeur bénéficiera de la conjoncture pour trouver un emploi. De plus, cette situation attira davantage d'étrangers à venir travailler en France, ce qui potentiellement pourrait augmenter les proportions de minorités. Seulement, ce type de chocs n'est pas ou peu anticipable par l'individu. Là encore, il est vraisemblable de supposer la non-corrélation du terme idiosynchratique futur avec les variables de contrôle de type proportions de minorités.

Qu'en est il des autres régresseurs `typmen, sante_declaree` ? Nous soupçonnons que ces régresseurs ne soient pas strictement exogènes. En effet, un choc d'employabilité influence la situation future du ménage, tout comme le fait d'être chômeur peut-être influencé par une dégradation passée de l'état de santé. L'âge est *a priori* corrélé aux chocs d'employabilité dans la mesure où un âge avancé peut conduire à des licenciements (parfois abusifs).

En conclusion, l'hypothèse d'exogénéité stricte semble pertinente pour les proportions de minorités uniquement. Après une étape d'inférence statistique, tout comme à la question 1.d), cette hypothèse est rejetée pour l'ensemble des régresseurs. En effet, le test de nullité jointe de toutes nos variables de contrôle associées aux coefficients β_1 à β_5 dans la régression :

$$\Delta(actop)_2 = \Delta X_2'\beta_0 + X_2'\beta_1 + X_3'\beta_2 + X_4'\beta_3 + X_5'\beta_4 + X_6'\beta_5 + \Delta\epsilon_2$$

est à égal à :

$$F(39, 28710) = 2.31$$

soit une p-valeur très faible. Si l'on prend donc en compte tous les régresseurs dans X_2',...,X_6' (où l'indice fait référence au temps d'interrogation t), nous rejetons l'hypothèse d'exogénéité stricte à 5%. Cependant, si ces variables ne contiennent que les proportions de minorités, `age` et `age2`, alors le F-test devient :

$$F(26, 28710) = 1.42$$

soit une p-valeur égale à 0,0774. Nous ne pouvons dans ces conditions pas rejeter la condition d'exogénéité stricte des variables de proportions de minorités et d'âge à 5% ![1] Les variables de proportions de minorités et d'âge sont donc strictement exogènes ce qui est cohérent avec nos suppositions du premier paragraphe.

Le problème des modèles linéaires est qu'ils sont mal adaptés à l'étude de variables binaires comme le fait d'être actif ou chômeur. En outre, étant donné que

$$\mathbb{E}(Y|X) = \mathbb{P}(Y = 1|X) \in [0, 1] \ ,$$

le modèle linéaire n'offre aucune garantie que :

$$\mathbb{E}(Y|X) = X'\beta_0 \in [0, 1]$$

Malgré son intérêt dans le cadre des modèles de panel (fait disparaître l'effet fixe), le modèle linéaire n'offre pas d'interprétation claire. Puisqu'il n'introduit pas de probabilité d'être actif ou chômeur, comment interpréter l'effet des variables de contrôle sur la variable `actop` ?

Une autre limite réside dans le fait que les résidus d'un modèle linéaire sont hétéroscédastiques.

1. Dans le cours de panel, slide "Un test simple de la condition d'exogénéité stricte", il est dit que si $T > 2$, on inclut $X_2, ..., X_T$ (**ou un sous-ensemble de ces régresseurs**) dans l'équation en différences premières et on teste leur significativité. C'est ce que nous avons fait avec les variables de proportions.

Voici le résultat de notre régression de panel de l'indicatrice d'être actif occupé dans le cadre d'un modèle linéaire. Pour que l'hypothèse d'exogénéité stricte soit vérifiée, nous n'avons pas pris les variables `sante_declaree` et `typmen`. Nous aurions pu les instrumenter par leur lag respectifs afin de se placer sous l'hypothèse d'exogénéité faible.

Tableau 13 – Régression panel de l'indicatrice d'être actif occupé dans le cadre d'un modèle linéaire

	(1)	
	actop	(s.e)
age	0.012**	(0.004)
age^2	-0.0009	(0.000)
Proportions minorités		
- *prop_ 678910_ immi*	0.035	(0.041)
- *prop_ 02_ immi*	-0.032	(0.041)
- *prop_ 678910*	-0.031	(0.030)
- *prop_ 02*	-0.026	(0.026)
Nombre d'observations	57 514	
corr(α_i, X_b)	-0.17	
ρ	0.64	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Nos estimations sont difficilement exploitables dans la mesure où, à l'exception de l'âge, aucune variable n'est significative à 5%. De plus, les coefficients observés pour les proportions de minorités sont très faibles et leur interprétation, comme nous l'avons dit précédemment, n'est pas fiable dans le cadre d'un modèle linéaire.

(b) 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}). 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 en guise de comparaison.

Si nous supposons l'exogénéité des résidus mais autocorrélation, alors, pour le modèle suivant :

$$Y_{it} = X'_{it}\beta_0 + \nu_{it} \quad \text{pour } i = 1...n \text{ et } t = 1...T$$

avec

$$\nu_{it} = \alpha_i + \epsilon_{it} \ ,$$

on a :

$$\mathbb{E}(X_{it}\nu_{it}) = 0 \ \forall\,(i,t) \ , \text{ et } \operatorname{cov}(\nu_{it}, \nu_{is}) \neq 0 \ \forall\,(t,s)\, t \neq s$$

Dans ce cas, on retombe dans les travers du modèle linéaire. En effet, si l'on suit la partie du cours dont il est question ici, pour une variable Y_{it} binaire, nous avons que :

$$\mathbb{E}(Y_{it}|X_{it} = x) = X'_{it}\beta_0 + \mathbb{E}(\nu_{it}|X_{it} = x) = X'_{it}\beta_0 = \mathbb{P}(Y_{it} = 1|X_{it} = x)$$

De fait, il n'y a aucune raison que la quantité $X'_{it}\beta_0$ appartienne à l'intervalle [0,1]. Pire, dans le cas d'un modèle probit ou logit où l'estimation se fait par maximum de vraisemblance, cet aspect empêcherait d'obtenir un estimateur convergent de β_0 !

D'autre part, les modèles logit et probit supposent l'indépendance des résidus ν_{it}. Pour le modèle logit, on a (avec nos notations du modèle de panel) :

$$\nu_{it} \overset{iid}{\sim} \Lambda(0,1) \ \forall\, i \text{ à } t \text{ fixé}$$

et pour le modèle probit :

$$\nu_{it} \overset{iid}{\sim} \mathcal{N}(0,1) \ \forall\, i \text{ à } t \text{ fixé}$$

De fait ceci n'est pas compatible avec l'hypothèse de non-corrélation car dans le cas iid nous avons bien que :

$$\operatorname{cov}(\nu_{it}, \nu_{is}) = 0 \ \forall\,(t,s)\, t \neq s$$

De plus, nous avons vu que pour qu'un modèle logit ou probit soit identifié, il est nécessaire de fixer le seuil et la variance des résidus (dans le cadre d'une interprétation en terme de variables latentes). Ainsi, imposer une décomposition des résidus en un terme individuel et un terme idiosynchratique pose problème dans la mesure où déjà $\mathbb{E}(\alpha_i) \neq 0$ en général et les résidus ϵ_{it} ne suivent pas nécessairement une loi normale de variance égale à 1 et d'espérance également nulle.

Tableau 14 – Effets marginaux sur l'indicatrice d'être au chômage d'un modèle logit avec ou sans panel

	(Avec panel) actop	(s.e)
typmen		
- *1 personne*	ref.	(.)
- *Famille monop.*	-0.012*	(0.005)
- *Couple sans enfant*	0.018***	(0.004)
- *Couple avec enfant(s)*	0.021***	(0.003)
- *Ménages (av. parenté)*	0.004	(0.016)
- *Ménages (ss. parenté)*	0.008	(0.017)
- *Autres*	-0.011	(0.010)
age	0.006***	(0.001)
age^2	-0.0001***	(0.000)
sante_declaree	-0.009***	(0.001)
Proportions minorités		
- *prop_678910_immi*	-0.008	(0.025)
- *prop_02_immi*	-0.022	(0.029)
- *prop_678910*	-0.066***	(0.016)
- *prop_02*	0.002	(0.017)
Nombre d'observations	57 514	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(Sans panel) actop	(s.e)
typmen		
- *1 personne*	ref.	(.)
- *Famille monop.*	-0.008	(0.006)
- *Couple sans enfant*	0.016***	(0.004)
- *Couple avec enfant(s)*	0.021***	(0.004)
- *Ménages (av. parenté)*	0.021	(0.015)
- *Ménages (ss. parenté)*	-0.001	(0.021)
- *Autres*	-0.011	(0.012)
age	0.005***	(0.001)
age^2	-0.0001***	(0.000)
sante_declaree	-0.009***	(0.001)
Proportions minorités		
- *prop_678910_immi*	0.002	(0.026)
- *prop_02_immi*	-0.016	(0.031)
- *prop_678910*	-0.068***	(0.017)
- *prop_02*	-0.016	(0.018)
Nombre d'observations	28 757	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tout d'abord, rappelons que nous nous plaçons sous l'hypothèse d'exogénéité des résidus et autocorrélation. En effet, supposer l'hypothèse d'exogénéité stricte ou faible n'est pas pertinent puisque les transformations *within* et *first-differences* ne sont pas applicables : elles n'éliminent pas l'effet fixe individuel (question (c)). Nous l'avons dit, cette hypothèse est critiquable et, surtout, pas compatible avec l'existence d'un effet fixe et d'un terme idiosynchratique. Nous nous attendons donc à avoir des résultats peu satisfaisants.

Intéressons-nous d'abord au résultat du modèle de panel. Si l'on suppose que les variables concernant la proportion d'immigrés dans l'environnement géographique ne sont pas biaisées, alors l'augmentation de 10% de la population d'origine non-européenne diminue en moyenne de $100 \times 0,1 \times 0,066 = 0,66\%$ les chances d'être actif occupé. Une augmentation similaire d'immigrés d'origine européenne diminuerait de $0,22\%$ la probabilité d'être actif occupé. Ces résultats sont très faibles. Contrairement au tableau 11, où nous avions régressé le logarithme du salaire horaire et non l'indicatrice d'être actif occupé, la proportion de minorités impacte moins le statut de l'emploi que le niveau de salaire. Enfin l'apport de la structure panel en terme d'interprétations (qualitative et quantitative) n'est pas conséquent. De fait, l'intérêt des modèles de panel est de pouvoir contrôler la dépendance d'état et de la distinguer de l'hétérogénéité inobservée. Or ici, les résultats sont sensiblement les mêmes ce qui ne laisse pas présager de biais de variables omises. En ce sens, l'approche panel n'est pas crédible.

L'absence de contrôle est problématique. Notamment l'éducation, considérée fixe dans le temps (elle ne peut être intégrée au modèle sans que Stata renvoie une erreur de type "aucune variation au cours de la période"). Nous aimerions pouvoir utiliser des transformations *within* ou *first differences* pour pouvoir intégrer plus de contrôles et potentiellement obtenir des résultats plus satisfaisants en terme d'ordres de grandeur et de significativité. À ce stade, supposer l'exogénéité stricte et l'autocorrélation des erreurs sans appliquer de telle transformation n'est pas pertinent. Comme nous allons le voir, les transformations usuelles étudiées dans le cours sur les données de panel ne sont pas applicables dans le cadre d'une variable dépendante binaire. En effet, il n'est pas possible de supprimer l'effet fixe individuel. Des modèles plus élaborés permettront d'y parvenir.

(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 ?

Pour rappel, l'estimateur *within* est basé sur la transformation :

$$U_{it} \to \tilde{U}_{it} = U_{it} - \frac{1}{T}\sum_{t=1}^{T=6} U_{it}$$

et l'estimateur des *first-difference* sur :

$$U_{it} \to \Delta U_{it} = U_{it} - U_{it-1}$$

En adoptant les mêmes notations que précédemment, il vient que, dans le cadre de l'exogénéité stricte et faible, l'effet individuel α_i est corrélé aux régresseurs X_{it}. L'intérêt d'utiliser un modèle linéaire plutôt qu'un modèle logit ou probit dans le cadre d'un panel est de pouvoir éliminer l'effet individuel via les transformations précédentes. Par exemple, nous avons sous l'hypothèse d'exogénéité stricte :

$$\mathbb{E}(X_{it}\alpha_i) \neq 0 \text{ mais } \mathbb{E}(X_{it}\epsilon_{it'}) = 0 \ \ \forall\,(t,t') \ ,$$

que

$$\mathbb{E}(Y_{it}|X_{it},\alpha_i) = X'_{it}\beta_0 + \alpha_i$$

et donc

$$\mathbb{E}(Y_{it} - Y_{it-1}|X_{it}, X_{it-1}) = (X_{it} - X_{it-1})'\beta_0$$

ce qui permet bien d'éliminer l'effet fixe individuel.
Dans le cas des modèles non-linéaires où Y_{it} est binaire (actif occupé ou chômeur), alors

$$\mathbb{E}(Y_{it}|X_{it},\alpha_i) = F\left(X'_{it}\beta_0 + \alpha_i\right)$$

et donc

$$\mathbb{E}(Y_{it}|X_{it},\alpha_i) = F\left(X'_{it}\beta_0 + \alpha_i\right) - F\left(X'_{it-1}\beta_0 + \alpha_i\right)$$

et alors l'effet fixe inobservé n'est pas éliminé. De même, la transformation *within* mène à :

$$\begin{aligned}
\mathbb{E}(\tilde{Y}_{it}|\tilde{X}_{it},\alpha_i) &= \mathbb{E}(Y_{it} - \bar{Y}_i|\tilde{X}_{it},\alpha_i) \\
&= \mathbb{E}(Y_{it}|\tilde{X}_{it},\alpha_i) - \mathbb{E}(\bar{Y}_i|\tilde{X}_{it},\alpha_i) \\
&= F(X'_{it}\beta_0 + \alpha_i) - \frac{1}{T}\sum_{t=1}^{T} F(X'_{it}\beta_0 + \alpha_i) \\
&= -\frac{1}{T}\sum_{t'\neq t}^{T} F(X'_{it'}\beta_0 + \alpha_i)
\end{aligned}$$

ce qui ne permet pas non plus d'éliminer l'effet fixe individuel α_i.

(d) 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 $\alpha_i \sim \mathcal{N}(0, \sigma^2)$. Comment s'estime ce modèle ? Peut-on calculer les effets marginaux dans ce modèle ? Estimer un "random effect probit model" (on pourra utiliser la commande Stata xtprobit).

Modélisons notre problème de la manière suivante :

$$Y_{it} = \mathbb{1}_{Y_{it}^* \geq 0}$$

avec

$$Y_{it}^* = X_{it}'\beta_0 + \alpha_i + \epsilon_{it}$$

Les modèles "random effect probit model" et "fixed effect logit model" tirent leur utilité du problème des paramètres incidents. Comme nous l'avons vu à la question précédente, utiliser un modèle non-linéaire ne permet pas d'éliminer l'effet fixe individuel α_i. L'alternative consistant à estimer par maximum de vraisemblance les paramètres $(\beta, \alpha_1, ..., \alpha_N)$ - avec N le nombre d'individus - ne conduit malheureusement pas à des estimateurs convergents des α_i. Pour un grand nombre d'individus et un petit nombre de périodes de temps, ce qui est notre notre cas ici, l'estimation des α_i n'est pas consistante. De ce fait, l'estimation de β_0 n'est pas consistante non plus (problème des paramètres incidents : si certains paramètres ne sont estimés de manière convergente, alors aucun paramètre ne sera estimé de manière convergente). Autrement dit, augmenter le nombre N d'individus à T fixé ne réduira pas le biais d'estimation sur β_0.

Les hypothèses fondamentales (certaines peuvent être relâchées) du "random effect probit model" sont les suivantes :

- α_i et X_{it} sont indépendants ;
- X_{it} est strictement exogène : $\mathbb{E}(X_{it}\alpha_i) = 0$;
- $Y_{i1}, ..., Y_{iT}$ sont indépendants conditionnellement à (X_i, α_i) ;
- $\alpha_i | X_i \sim \mathcal{N}\left(0, \sigma_\alpha^2\right)$

Étant donné qu'α_i est aléatoire, on ne peut pas écrire :

$$\begin{aligned} L_i(y_i|x_i;\beta_0) &= \prod_{t=1}^{T} L_{it}(y_{it}|x_i;\beta_0) \\ &= \prod_{t=1}^{T} \left[\phi(x_{it}'\beta_0 + \alpha_i)\right]^{y_{it}} \left[1 - \phi(x_{it}'\beta_0 + \alpha_i)\right]^{(1-y_{it})} \end{aligned}$$

avec

$$\begin{aligned} L_{it}(y_{it}|x_i;\beta_0,) &= \mathbb{P}(Y_{it} = y_{it}|x_i;\beta_0) \\ &= \mathbb{P}(Y_{it} = y_{it}|x_{it};\beta_0) \end{aligned}$$

De fait on utilise la formule de Bayes pour ré-écrire la vraisemblance :

$$\begin{aligned} L_{it}(y_{it}|x_i;\beta_0,\sigma_\alpha^2) &= \int_{-\infty}^{+\infty} L_{it}(y_{it},\alpha_i|x_i;\beta_0,\sigma_\alpha^2)d\alpha_i \\ &= \int_{-\infty}^{+\infty} L_{it}(y_{it}|x_i,\alpha_i;\beta_0,\sigma_\alpha^2)(1/\sigma_\alpha^2)\phi(\alpha_i/\sigma_\alpha^2)d\alpha_i \end{aligned}$$

avec cette fois-ci la possibilité d'écrire

$$L_{it}(y_{it}|x_i,\alpha_i;\beta_0,\sigma_\alpha^2) = \left[\phi(x_{it}'\beta_0 + \alpha_i)\right]^{y_{it}} \left[1 - \phi(x_{it}'\beta_0 + \alpha_i)\right]^{(1-y_{it})}$$

Et finalement on obtient une fonction de vraisemblance du type :

$$L_i(y_i|x_i;\beta_0,\sigma_\alpha^2) = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T} L_{it}(y_{it}|x_i,\alpha_i;\beta_0,\sigma_\alpha^2)\right] (1/\sigma_\alpha^2)\phi(\alpha_i/\sigma_\alpha^2)d\alpha_i$$

Et la vraisemblance totale :

$$l(\beta_0,\sigma_\alpha^2) := \sum_{i=1}^{N} L_i(y_i|x_i;\beta_0,\sigma_\alpha^2)$$

peut alors être maximisée en β et σ_α^2 de telle sorte à obtenir des estimateurs convergents $\hat{\beta}$ et $\hat{\sigma}_\alpha^2$. En résumé, la formule de Bayes permet de contourner le problème de l'effet fixe aléatoire. L'avantage de cette méthode est qu'elle ne nécessite pas l'estimation du paramètre α_i. On peut donc estimer l'effet marginal par :

$$\frac{\partial \widehat{\mathbb{P}(Y_{it}=1|X=x)}}{\partial x_{itk}} = \hat{\beta}_k \int_{-\infty}^{+\infty} \phi(x_{it}'\beta_0 + \hat{\sigma}_\alpha^2\alpha_i)\phi(\alpha_i)d\alpha_i \tag{1}$$

qui est du même signe que $\hat{\beta}_k$. Wooldridge parle en particulier de calculer l'effet marginal à $\alpha = 0$. Ceci réfère à l'estimation de l'effet marginal à la moyenne (car $\alpha = 0$ est la valeur moyenne de α dans la population d'après l'hypothèse de normalité). L'espérance du terme (1) permet d'estimer l'effet marginal moyen. Vous trouverez le résultat de notre régression question (f).

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

Le "fixed effect logit model" permet d'éliminer l'effet fixe individuel α_i au travers d'une transformation se basant sur l'hypothèse d'exogénéité stricte, comme dans le cas de l'estimateur *within* ou de *first-difference*. Celui-ci ne requiert pas de condition sur le terme α_i. Cependant, $Y_{i1},...,Y_{iT}$ sont supposés indépendants conditionnellement à (X_i,α_i). Et la distribution de $Y_{i1},...,Y_{iT}$ conditionnellement à $(X_i,\alpha_i,\, n_i = \sum_{t=1}^{T} Y_{it})$ ne dépend pas de α_i. Sans rentrer dans les détails, l'idée de cette méthode est de considérer les probabilités :

$$\mathbb{P}\left(y_{i1}=0,...,y_{it}=1,...,y_{iT}=0 \,\middle|\, x_{i1},...,x_{iT},\alpha_i,\sum_{t'=1}^{T} y_{it'}=1\right) \;\forall\; t\in[1,...,T]$$

Sachant que

$$\mathbb{P}(y_{it}=1|x_{it},\alpha_i) = \Lambda\left(x_{it}\beta_0+\alpha_i\right)$$

Conditionnellement au nombre de "succès", l'effet individuel disparaît de la vraisemblance fonctionnelle. Par exemple, dans le cas T=2, on trouve :

$$\mathbb{P}(y_{i1}=0,y_{i2}=1|x_{i1},x_{i2},y_{i1}+y_{i2}=1) = \frac{\exp(\Delta x_{i2}\beta_0)}{1+\exp(\Delta x_{i2}\beta_0)}$$

et

$$\mathbb{P}(y_{i1}=1,y_{i2}=0|x_{i1},x_{i2},y_{i1}+y_{i2}=1) = \frac{1}{1+\exp(\Delta x_{i2}\beta_0)}$$

de telle sorte à obtenir des probabilités indépendantes de α_i.

Toujours dans le cas $T = 2$, la vraisemblance conditionnelle s'écrit :

$$l(\beta_0) := \sum_{i=1}^{N} \left[k_{01i} \ln \left(\frac{\exp(\Delta x_{i2}\beta_0)}{1 + \exp(\Delta x_{i2}\beta_0)} \right) \right] + \left[k_{10i} \ln \left(\frac{1}{1 + \exp(\Delta x_{i2}\beta_0)} \right) \right]$$

Finalement, β_0 peut être estimé de manière convergente mais les effets marginaux dépendent cette fois de α_i. On ne peut donc pas les calculer. En revanche, l'on peut montrer que :

$$\frac{\partial \mathbb{P}(\widehat{Y_{it} = 1|X = x})}{\partial x_{itk}} = h(x, \hat{\beta}_0, \alpha_i) \times \hat{\beta}_0$$

avec h une fonction positive. De fait, le signe de β_0 donne le signe de l'effet marginal.

L'avantage de ce modèle, contrairement au "random effect probit model" est qu'il n'impose pas la condition de non-corrélation entre les variables de contrôle X_{it} et l'effet fixe individuel α_i ni de condition de distribution. Vous trouverez le résultat de notre régression question (f). Nous y avons reporté les effets marginaux sous l'hypothèse (que fait Stata par défaut) que l'effet fixe individuel est nul. Sinon, il n'est pas possible d'interpréter le modèle par ce biais. Il faut alors considérer les odd-ratios.

(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 ?

Dans un premier temps, nous avons appliqué les modèles précédents en intégrant les proportions de minorités et non la variable `cat_origine`. Comme nous allons le voir, ces modèles sophistiqués n'apportent pas plus de pertinence à nos résultats dans ce cadre. Ceux-ci sont disponibles en page 27. **Les variables `education` et `csp_actifs` ont été intégrées au modèle "random effect logit" mais ne figurent pas dans le résultat de la régression en haut à gauche du tableau 15 pour des raisons de place**. Déjà, nous observons que les effets marginaux estimés par le modèle "fixed effect logit" sont aberrants et très proches de 0. En effet, nous l'avons vu, ce modèle ne peut estimer les effets marginaux sans prendre une hypothèse très forte, celle où $\alpha = 0$. En effet, la formule des effets marginaux dépend de l'effet fixe individuel inconnu et non estimé. Cette régression est complètement inexploitable en terme d'effets marginaux. Nous nous sommes intéressés aux odd-ratio mais cela n'a pas donné de résultats concluants.
Les résultats du modèle "random effect probit" sont eux plus intéressants. Seulement, les coefficients estimés, ainsi que leur signe, ne diffèrent presque pas de l'approche naïve et critiquable du modèle de panel de la question 2b). De plus, la comparaison des modèles n'est pas aisée au vu des différentes hypothèses sous-jacentes. Nous pouvons néanmoins dire, au vu de la significativité du coefficient, que si la proportion de voisins d'origine non-européenne (pouvant être françaises) augmente de 10% alors la probabilité que l'individu soit chômeur augmentera en conséauence de 0,64%...

Il est important de rappeler l'utilité du modèle "random effect logit" pour notre problématique. Alors que les variables de type proportions de minorité semblent ne pas être suffisamment influentes sur les chances d'être actif, l'intérêt de ce modèle plus sophistiqué est de pouvoir réintégrer la variable `cat_origine` dans nos régressions. Les hypothèses sur lesquelles se basent ce modèle permettent d'éliminer l'effet fixe individuel de nos régressions. En particulier, nous n'avons plus rencontré d'erreurs, sous Stata, à intégrer des variables variant peu dans le temps dans le modèle, comme `cat_origine`. En faisant l'hypothèse forte que X_{it} est strictement exogène mais que α_i n'est pas nul, le modèle permet d'intégrer les variables `education` et `csp_actif` tout en résolvant le problème de leur endogénéité. Cependant, l'hypothèse de normalité de l'effet fixe est forte. Cette hypothèse sera d'autant plus crédible que ce terme résiduel d'effet individuel captera peu d'effets, en le sens où nous devons ajouter des contrôles dans nos régressions (ce qui est possible maintenant) pour que le modèle soit applicable et interprétable. À ce propos, pour n'avoir dans un premier temps pas intégrer les variables `education` et `csp_actif`, nous nous sommes rendus compte que les coefficients des modalités de la variable d'origine étaient biaisés vers le bas.

Tableau 15 – Effets marginaux sur l'indicatrice d'être actif occupé des modèles "random effect probit" et "fixed effect logit" et comparaison avec les modèles précédents - Utilisation des proportions de minorités

	(Random effect probit)	
	actop	(s.e)
typmen		
- *1 personne*	ref.	(.)
- *Famille monop.*	-0.012*	(0.005)
- *Couple sans enfant*	0.017***	(0.004)
- *Couple avec enfant(s)*	0.020***	(0.003)
- *Ménages (av. parenté)*	-0.003	(0.018)
- *Ménages (ss. parenté)*	0.004	(0.016)
- *Autres*	-0.006	(0.009)
age	0.006***	(0.001)
age^2	-0.0001***	(0.000)
`sante_declaree`	-0.007***	(0.001)
Proportions minorités		
- *prop_ 678910_ immi*	-0.014	(0.025)
- *prop_ 02_ immi*	-0.028	(0.027)
- *prop_ 678910*	-0.064***	(0.016)
- *prop_ 02*	-0.0005	(0.016)
Nombre d'observations	57 409	

	(Fixed effect logit)	
	actop	(s.e)
age	0.005***	(0.001)
age^2	-0.0001***	(0.000)
`sante_declaree`	-0.009***	(0.001)
Proportions minorités		
- *prop_ 678910_ immi*	0.002	(0.009)
- *prop_ 02_ immi*	-0.001	(0.006)
- *prop_ 678910*	-0.0003	(0.003)
- *prop_ 02*	-0.005	(0.019)
Nombre d'observations	2 438	

	(Panel cluster)	
	actop	(s.e)
typmen		
- *1 personne*	ref.	(.)
- *Famille monop.*	-0.012*	(0.005)
- *Couple sans enfant*	0.018***	(0.004)
- *Couple avec enfant(s)*	0.021***	(0.003)
- *Ménages (av. parenté)*	0.004	(0.016)
- *Ménages (ss. parenté)*	0.008	(0.017)
- *Autres*	-0.011	(0.010)
age	0.006***	(0.001)
age^2	-0.0001***	(0.000)
sante_declaree	-0.009***	(0.001)
Proportions minorités		
- *prop_ 678910_ immi*	-0.008	(0.025)
- *prop_ 02_ immi*	-0.022	(0.029)
- *prop_ 678910*	-0.066***	(0.016)
- *prop_ 02*	0.002	(0.017)
Nombre d'observations	57 514	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(Sans panel)	
	actop	(s.e)
typmen		
- *1 personne*	ref.	(.)
- *Famille monop.*	-0.008	(0.006)
- *Couple sans enfant*	0.016***	(0.004)
- *Couple avec enfant(s)*	0.021***	(0.004)
- *Ménages (av. parenté)*	0.021	(0.015)
- *Ménages (ss. parenté)*	-0.001	(0.021)
- *Autres*	-0.011	(0.012)
age	0.005***	(0.001)
age^2	-0.0001***	(0.000)
sante_declaree	-0.009***	(0.001)
Proportions minorités		
- *prop_ 678910_ immi*	0.002	(0.026)
- *prop_ 02_ immi*	-0.016	(0.031)
- *prop_ 678910*	-0.068***	(0.017)
- *prop_ 02*	-0.016	(0.018)
Nombre d'observations	28 757	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Le tableau 16 ci-dessous rend compte de la régression de la variable `actop` en prenant compte la variable d'origine. Les résultats de ce modèle sont les plus satisfaisants de cette partie 2. En effet, nous observons déjà la grande significativité de nos modalités d'origine. De fait, être immigré non-européen diminue la probabilité d'être actif d'en moyenne 7,1%. Un immigré naturalisé et de la même origine aura en moyenne 4,8% moins de chance d'être actif comparé à un français né en France de parents français. Ce chiffre diminue encore, mais reste négatif, pour un français toujours d'origine non-européenne. Il est intéressant de remarquer que, pour une fois, les coefficients associés à l'immigration européenne sont négatifs, mais rencontre toujours un problème de significativité. Nous passons à de plus amples interprétations et à une "comparaison" des modèles en page 29.

Tableau 16 – Effets marginaux sur l'indicatrice d'être actif occupé du modèle "random effect probit" - Utilisation de l'origine

	(Random effect probit)	
	actop	(s.e)
typmen		
- *1 personne*	ref.	(.)
- *Famille monop.*	-0.006	(0.005)
- *Couple sans enfant*	0.019***	(0.004)
- *Couple avec enfant(s)*	0.024***	(0.003)
- *Ménages (av. parenté)*	0.008	(0.016)
- *Ménages (ss. parenté)*	0.003	(0.018)
- *Autres*	0.008	(0.008)
age	0.004***	(0.001)
age^2	-0.0001***	(0.000)
`sante_declaree`	-0.007***	(0.001)
education		
- *Dipl. supérieur*	ref.	(.)
- *Bac +2*	-0.010***	(0.003)
- *Bac ou brevet pro.*	-0.012***	(0.003)
- *CAP, BEP*	-0.018***	(0.003)
- *Brevet des collèges*	-0.015**	(0.005)
- *Pas de diplôme*	-0.038***	(0.005)
csp_actif		
- *Non renseigné*	ref.	(.)
- *Cadres*	0.081	(0.061)
- *Professions int.*	0.082	(0.061)
- *Employés*	0.076	(0.061)
- *Ouvriers*	0.061	(0.061)
cat_origine		
- *fra1pimmi_Neurope*	-0.023***	(0.007)
- *fra1pimmi_europe*	-0.005	(0.004)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.071***	(0.011)
- *immi_Nnatur_europe*	-0.024**	(0.008)
- *immi_natur_Neurope*	-0.048***	(0.009)
- *immi_natur_europe*	-0.013	(0.012)
Nombre d'observations	57 409	

L'utilisation des variables de proportions de minorités a été motivée dès la partie 1 par le fait que la variable d'origine, constante dans le temps, se confond avec l'effet fixe individuel et ne peut tout simplement pas être incorporée aux premiers modèles de panel (questions 1d), 2a), 2b)) sous risque de causer une erreur sur Stata. À la fin de la première partie de ce projet, nous remarquions que les proportions de minorités présentes dans le voisinage de l'individu interrogé n'étaient pas influentes sur le salaire, ni significatives au seuil de 5%. En l'occurrence, l'utilisation d'une variable dépendante binaire, l'indicatrice d'être actif, n'a pas non plus amené à des résultats probants en terme d'ordre de grandeurs. Les premiers modèles de cette partie 2 n'étaient pas crédibles dans la mesure où les hypothèses entre modélisation probit/logit et structure de panel n'étaient pas compatibles. Plus particulièrement, la structure de panel de notre modèle logit n'ajouterait que pas ou peu de significativité à nos variables par rapport à la structure empilée ($t = 6$). Notre volonté d'ajouter plus de contrôles tout en limitant les effets d'endogénéité s'est heurtée à l'impossibilité d'appliquer, sur les modèles précédents, des transformations de type *within* ou *first-difference* qui ne pouvaient faire disparaître l'effet fixe individuel. Deux nouveaux modèles, plus sophistiqués, ont laissé présager de meilleurs résultats quant à l'influence des variables de proportions sur la probabilité d'être actif. Cependant, là encore, les résultats sont décevants. Nous pouvons penser qu'il n'y pas équivalence entre le fait d'être entouré de beaucoup d'immigrés et le fait d'être immigré soi-même (cela est en fait du bon sens). Il existe, d'après nos statistiques descriptives, une corrélation claire entre le fait d'être immigré et celui d'être entouré d'immigrés dans son voisinage. Cependant, corrélation n'est pas causalité. En particulier, nous ne pouvons pas affirmer, au regard de nos estimations sur les proportions de minorités, qu'être entouré d'immigrés non-européens est particulièrement négatif sur le salaire perçu (en moyenne). Comme nous l'avons dit, la précarité va de paire avec la mixité sociale. De fait, si un individu touche un salaire inférieur, ou est chômeur, celui-ci sera plus susceptible d'habiter un immeuble et d'être entouré de personnes d'origines étrangères. De surcroît, la situation dans laquelle se trouve cet individu n'est pas à attribuer aux immigrés qui l'entourent. Les proportions de minorités ne permettent pas de conclure quant à d'éventuelles discriminations. Cela serait le cas si, systématiquement, un individu d'origine lambda habitait volontairement avec des personnes de la même origine, ce qui est observé (communautarisme) mais pas toujours vérifié. De fait, ces variables servirait davantage à mesure la volonté d'intégration des individus plutôt qu'une éventuelle discrimination sur le marché du travail. Nous pourrions par exemple régresser la catégorie d'origine par rapport aux densités de minorités et contrôler cette fois par l'indicatrice d'être actif voire le salaire.

Il s'avère que l'utilisation de la variable `cat_origine` dans le modèle "random effect probit" offre des interprétations pertinentes. En ce sens, comme nous l'avons dit, le fait d'être d'origine non-européenne baissent, entre 2,5 et 7,1% les probabilités d'être actif occupé et sont croissantes en l'appartenance "juridique" de l'individu au territoire français. Nous pouvons affirmer que ces résultats témoignent d'une discrimination à l'embauche des individus originaires de pays extérieurs à l'Europe. Cette tendance est la plus marquée pour les individus originaires des pays du Maghreb et ceux de l'Afrique Noire. La variable d'origine a été contrôlée par le niveau d'éducation et le domaine professionnel d'activité. De fait, nous sommes plus en mesure de parler de discriminations puisque l'approche « toutes choses égales par ailleurs »est légitime. Contrairement aux estimations précédentes (partie 1 comprise), nous observons des coefficients négatifs sur les modalités européennes, laissant supposer que la probabilité d'être actif occupé diminue selon que l'individu est immigré, même européen. Dans la mesure où la significativité statistique de ces modalités est faible voire très faible, nous ne pouvons donc pas apporter d'interprétations ayant vocation à orienter des politiques publiques futures. Le modèle de sélection de la partie 3 va nous permettre de conclure quant à l'effet de l'immigration européenne sur le salaire et la probabilité d'être chômeur.

S'il fallait enfin parler de la comparaison des modèles de cette partie 2, le cheminement logique de cette partie nous montre comment la comparaison des estimations que nous avons faites est difficile. Déjà, car certains modèles sont corrompus du point de vue des hypothèses utilisées. Ensuite, ces mêmes hypothèses, conditions nécessaires de validité des modèles, sont différentes. Si l'on prend l'exemple du modèle "random effect probit", celui-ci fait une hypothèse très forte de normalité du terme d'effet fixe individuel. Le modèle "fixed effect logit" ne peut quant à lui être interprété en terme d'effets marginaux.

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

Préambule

Dans la partie précédente, nous n'avons pas considéré les individus inactifs pour nous concentrer uniquement sur les individus actifs occupés et actifs chômeurs. Nous avons dit, je cite : « De plus [...] dans la mesure où nous cherchons à déceler des inégalités liées à l'origine sur le marché du travail, il fait plus de sens de ne pas considérer les inactifs qui n'ont pas cherché à se porter sur le marché du travail. »En l'occurence nous avons mal interprété l'énoncé dans le sens où il est écrit : « 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** ». Nous nous étions alors focalisé uniquement sur les déterminants d'être ou non chômeur. Cependant, lorsque nous regardons la distribution des modalités de la variables `acteu` croisées avec celles de la variable `cat_origine`, nous nous apercevons que les proportions d'individus inactifs sont plus grandes que celles des individus au chômage, en particulier pour les immigrés. Nos statistiques descriptives le confirmait d'ailleurs.

Tableau 17 – Proportions d'inactifs et de chômeurs selon la catégorie d'origine

Catégorie d'origine	Actif	Chômeur	Inactif
- *fra1pimmi_ Neurope*	67.93%	**11.86**%	20.20%
- *fra1pimmi_ europe*	73.23%	5.54%	21.24%
- *fra2pfra*	74.76%	4.68%	20.56%
- *immi_ Nnatur_ Neurope*	48.53%	**13.24**%	38.24%
- *immi_ Nnatur_ europe*	68.22%	7.98%	23.80%
- *immi_ natur_ Neurope*	67.25 %	**10.01**%	22.74%
- *immi_ natur_ europe*	68.03 %	5.46 %	26.50 %

En particulier, les individus issus de pays extérieurs à l'Europe sont très touchés par l'inactivité. En fait, nous n'avons pas pris en compte qu'un individu immigré nouvellement (ou pas) arrivé en France n'est pas chômeur mais inactif s'il ne fait pas les démarches adéquates. Ceci est néanmoins à nuancer. En effet, tous les individus interrogés possèdent un logement (plus ou moins précaire). Ceci témoigne d'une installation plus ou moins permanente sur le territoire national et donc sous-entend que les individus concernés sont plus susceptibles d'avoir fait les démarches de déclaration du statut de demandeur d'emploi. S'inscrire comme demandeur d'emploi ne signifie pas toucher les allocations chômage, car celles-ci sont calculées à partir du salaire, et qu'un nouvel arrivant n'a jamais travaillé en France. Autrement dit, la question de prendre en compte ou non les inactifs est délicate. Dans tous les cas, nos analyses de la partie 2, même si elles se concentrent uniquement sur les actifs au sens du BIT, n'en demeurent pas moins vraies. D'ailleurs, il est courant de lire des études qui établissent des liens de corrélations entre immigration et le fait être actif occupé ou chômeur (uniquement). En particulier, il est d'actualité de comparer les taux de chômage des immigrés ou français d'origine étrangère avec ceux des français d'origine. Nous aurions pu dans la partie 2 considérer une variable `acteu2` égale à 0 si l'individu est chômeur ou inactif, 1 si actif occupé. Dans cette partie nous vous proposons de tester ces deux approches. Abordons déjà la question (a).

(a) Quel cadre, entre le modèle de censure (tobit simple) et le modèle de sélection (tobit généralisé), semble le plus approprié ici ? 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 ($\mathbf{Y}^*$) ?

Nous avons adopté dans cette partie un modèle de sélection générale (ou tobit II). En référence avec le cours, nous nous apercevons que le problème de considérer les individus inactifs est toujours présent. En effet, dans le cours, les inactifs sont pris en compte. De manière générale, le tobit généralisé permet de prendre en compte de potentielles discriminations à l'embauche dont peuvent souffrir à la fois les chômeurs et les inactifs. Là est tout l'avantage de ce type de modélisation par rapport à la partie 1 qui elle ne se focalise que les discriminations de salaire à proprement parler. Ici nous considérons donc un modèle du type :

$$Y^* = X'\beta_0 + \epsilon \text{ avec } D = \mathbb{1}\{Z'\gamma_0 + \eta\}$$

où Z contient au moins une composante qui est exclue de X et Y^* est le logarithme du salaire horaire **réel**. Ainsi Y^* est soumis à un phénomène de sélection dans la mesure où les individus chômeurs ou inactifs de touchent pas de salaire. On observe donc uniquement :

$$Y = Y^* D.$$

De plus, η est le résidu qui explique le fait d'avoir ou non un revenu (participation) et ϵ est le résidu du salaire, autrement dit un terme de productivité inobservée. Il est alors intuitif de supposer que ϵ et η sont corrélés. Un individu à faible productivité touchera un faible salaire et sera moins incité à se porter sur le marché du travail. Nous avons ensuite choisi Z comme étant la variable `NBENF` comme instrument corrélé à la participation (joue sur la probabilité d'être actif ou non) mais n'affectant pas directement Y^*. Cet instrument est discutable dans la mesure où l'on fait l'hypothèse que le nombre d'enfants (à charge) est indépendant du salaire. En effet, les évolutions de salaire ne sont pas les mêmes si on a des enfants à charge, en particulier en bas âge. Nous avions également pensé à prendre la variable `taille_menage` mais celle-ci avait moins d'effet sur la participation.

Aussi, nous supposons également que :
- (ϵ, η) indépendants de (X, Z) ;
- $\eta \sim \mathcal{N}(0,1)$;
- $\mathbb{E}(\epsilon, \eta) = \delta_0 \eta$,

de manière à pouvoir estimer de manière convergente les paramètres du modèle.

Quel rapport avec le chômage frictionnel et structurel ? Nous pouvons imaginer qu'un individu possède un salaire horaire de réserve W_r tel que si $W < W_r$, l'individu ne se portera pas sur le marché du travail. Dans le cas de travails peu rémunérés au-dessus du smic, un individu aura comme salaire de réserve le smic dans la mesure où son salaire potentiel n'est pas tellement plus élevé. Si cet individu se présente auprès d'un employeur mais que celui-ci ne dispose pas d'assez de compétences pour le poste visé, l'employeur pourrait décider de l'embaucher mais à un salaire plus faible que le smic (salaire de réserve), ce qui bien sûr n'est pas possible légalement. Ceci cause un chômage dit structurel. À cette inadéquation des qualifications des offreurs de travail aux besoins des entreprises s'ajoutent l'immobilité des salariés ou encore la rigidité des législations sociales. Une fois encore, la question de considérer les inactifs se pose. Dans la mesure où les inactifs (dans la base) ne sont pas en recherche d'emploi (confère l'instruction sur Stata : `tab acteu_details acteu`), nous pourrions penser que notre raisonnement ne s'applique qu'aux individus chômeurs. Ceci étant, un inactif peut très bien n'être pas déclaré chômeur mais rechercher quand même du travail ! Ceci nous permet de justifier l'ajout des inactifs dans notre modèle. D'un part, la sélection (D) peut venir d'une incompatibilité entre compétences effectives et salaire minimum, qui touche aussi bien chômeurs qu'inactifs. D'autre part, on observe aussi que les salaires des individus qui se sont volontairement mis sur le marché du travail. Le phénomène de sélection affecte donc tout autant les inactifs qui ne se sont pas portés sur le marché du travail pour diverses raisons. Pour reprendre l'idée du salaire de réserve, nous pouvons, à titre d'exemple, reprendre les notations du cours :

$$\ln W = X'\beta_0 + \epsilon \text{ avec } D = \mathbb{1}\{\ln W - \ln W_r \geq 0\} \equiv \mathbb{1}\{Z'\gamma_0 + \eta \geq 0\}$$

Enfin, nous cherchons ici à estimer l'effet de la variable latente Y^* plutôt que celui de Y. En effet nous cherchons à mesurer la dynamique sous-jacente du salaire dans l'ensemble de la population et mesurer l'impact de nos variables de contrôle sur le vrai mécanisme du salaire. Pour cela nous utiliserons un modèle heckit étant

donné que les hypothèses sur les résidus sont moins restrictives que l'estimation par maximum de vraisemblance. Avec nos notations précédentes, nous estimerons donc :

$$\mathbb{E}(Y|X, Z, D = 1) = X'\beta_0 + \delta_0\lambda(Z'\gamma_0)$$

avec λ le ratio de Mills et D l'indicatrice d'être actif occupé.

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

Les résultats de notre régression dans le cadre d'un modèle de sélection généralisée **lorsque nous ne considérons que les actifs occupés et les chômeurs** sont disponibles tableaux 19 et 20. Ceux-ci sont divisés en deux parties mais correspondent bien à une seule et même estimation. Vous trouverez en annexes 1 et 2 les régressions complètes de l'indicatrice de participation selon que nous prenons en compte ou pas les individus inactifs.

Concentrons-nous en premier lieu sur la régression du logarithme du salaire horaire (tableau 20). En premier lieu nous remarquons que les signes observés pour les variables explicatives du modèle sont tous les même que l'on considère ou non les inactifs. Les coefficients des variables `pibhab_cat`, `tuu_r`, `typvois`, `zus`, `typmen`, `age`, `age`^{2} et `homme` sont sensiblement les mêmes avec ou sans inactifs. De plus, l'ajout des inactifs n'apporte pas de gains considérables en terme de significativité des variables. On remarque que les coefficients des modalités de la variable `csp_actif` sont plus élevés si l'on en prend en compte les inactifs. En effet, en prenant en compte cette catégorie d'individus, le phénomène de sélection est d'autant plus marqué en terme de proportions. De fait, cette variable capte davantage les écarts de salaire chez les actifs occupés que dans le cas où les inactifs sont exclus. Ceci a des conséquences directes sur la variable `cat_origine` : les coefficients des modalités `cat_origine` sont tous moins élevés. La significativité reste la même. Les résultats observés sur cette variable d'origine sont plus forts et nettement plus "tranchés" que ceux observés précédemment, surtout vis-à-vis des immigrés (au sens large) non-européens. Si l'on considère la première configuration (heckit sans inactifs), nous voyons qu'être immigré non-naturalisé d'origine extérieure à l'Europe fait diminuer le salaire de l'individu actif d'en moyenne 15% par rapport à s'il était français né en France de parents français. Un immigré naturalisé mais d'origine Maghrébine, Africaine ou Asiatique (on ne peut pas dire directement), touchera en moyenne environ 10%. De même, un individu français moins significatif en terme d'ordre de grandeur. L'absence de distinction entre immigré Africain, Maghrébin, Asiatique n'est pas un problème ici et il serait facile de décomposer la variable `immi_Nnatur_Neurope` de manière à obtenir précisément les coefficients associés aux différentes nationalités. Pour l'avoir vérifié directement sur Stata, nous pouvons affirmer avec certitude que les individus d'origine Maghrébine sont ceux pour lesquels les coefficients estimés sont, en valeur absolue, les plus importants, pour une significativité à 5%. De manière générale, ce modèle heckit offre des résultats très intéressants pour notre étude puisque les ordres de grandeurs observés sont les plus importants au regard de nos précédentes estimations. Les individus non-européens qui se sont portés sur le marché du travail semblent souffrir d'une même discrimination de salaire liée très certainement à la couleur de peau ou au niveau de français. Les résultats sont également cohérents dans la mesure où un immigré non-naturalisé touchera moins qu'un immigré naturalisé qui lui-même touchera moins qu'un français ayant des origines non-européennes. L'obtention de la nationalité française semble donc jouer positivement sur le salaire. Seulement, l'apport de celle-ci ne compense l'origine de l'individu.

Que penser du reste du modèle ? Déjà, la variable nombre d'enfants est significative à 0,1% pour le modèle heckit avec inactifs et est négative. Lorsque les femmes ont des enfants (ceci est d'autant plus vrai lorsqu'ils sont en bas âge), celles-ci ont tendance à ne pas travailler et leur salaire est donc inobservé. Ceci est naturel dans la mesure où la raison de l'inactivité des femmes vient davantage du fait que celles-ci sont mères au foyer ou n'ont pas occupé de poste depuis leur congé maternité. Les femmes au chômage, parce qu'elles recherchent un emploi, sont moins affectées par le nombre d'enfants qu'elles ont à charge. En ce sens, ces femmes sont tout autant disposées à travailler, puisqu'elles sont en situation de demandeuses d'emploi malgré les enfants qu'elles peuvent avoir à charge(nous parlons ici d'enfants de tout âge, pas forcément en bas âge, ce qui aboutit à des résultats certainement différents). Encore une fois, cet instrument est critiquable dans la mesure où les évolutions de salaire ne sont pas les mêmes si on a des enfants. De plus, nous ne considérons ici que les femmes, et pas les hommes qui eux sont moins affectés par le nombre d'enfants en général. En ce qui concerne la lecture des coefficients, nous ne pouvons pas ici proposer d'interprétation quantitative des coefficients, seulement qualitatives. En ce qui concerne la variable `cat_origine`, nous remarquons que toutes les modalités ont des coefficients négatifs. De fait, le fait d'être immigré naturalisé, non-naturalisé ou simplement français mais d'origine étrangère joue négativement sur

le fait de participer au marché du travail, et donc d'être chômeur. C'est un deuxième résultat fort, d'autant plus que les coefficients associés sont tous (ou presque) significatifs à 0,1%.

Enfin, les valeurs du coefficient lambda référent à la valeur de :

$$\widehat{\delta_0} = \frac{\widehat{\text{cov}(\epsilon, \eta)}}{\widehat{\mathbb{V}}(\eta)}$$

(avec nos notations utilisées en question (a)) sont tous les deux positifs dans les cas avec ou sans inactifs. Cela signifie que le résidu η expliquant le fait d'avoir un revenu ou pas et le résidu ϵ du salaire (i.e un résidu de productivité) sont positivement corrélés. Ceci vient du fait que ceux qui vont avoir un revenu (de manière générale) sont ceux qui auront un revenu plus important. Ceux qui ne travaillent pas sont ceux qui, à l'inverse, anticipent un faible salaire. La corrélation $\text{corr}(\epsilon, \eta)$ (coefficient rho) est nettement plus forte et d'ailleurs égale à 1 pour le modèle sans inactifs. Cela semble confirmer nos propos sur le chômage structurel : les individus au chômage, donc dont les salaires sont inobservés, sont au chômage car ils n'ont pas les qualifications requises pour toucher le salaire minimum légal. En somme, si l'individu est chômeur (résidu η), ceci vient du fait qu'il est peu productif (résidu ϵ). La valeur de cette corrélation pour le cas avec inactifs est 3 fois plus faibles néanmoins. En effet, nous pouvons penser que les individus inactifs ne sont pas foncièrement moins productifs et pourraient certainement trouver un emploi rémunéré au-dessus du smic s'ils étaient demandeurs. De fait, ces individus sont inactifs pour d'autres raisons que pour des raisons de productivité. Nous pensons dès lors aux discriminations dont souffrent les personnes d'origine étrangère. Celles-ci ne sont pas moins productives qu'un "français de souche", mais sont victimes de discrimination à l'embauche ou ont été victimes de discrimination lors d'une longue recherche passée et infructueuse d'emploi. On voit alors l'intérêt d'avoir ajouté les individus inactifs au modèle et d'avoir considérer un modèle de sélection. Cette observation est confirmée par le rejet de l'hypothèse de sélection exogène à 5% et 1%. En effet, nous rejetons l'hypothèse de nullité du coefficient δ_0. Il ya donc bien un effet de sélection. Nous vous proposons une comparaison des deux modèles heckit, du premier modèle de MCO et de la régression instrumentale de la partie 1.

Tableau 18 – Comparaison avec le modèle instrumental

	(heckit 1)		(heckit 2)		(IV)		(MCO)	
	logsalhoraire	(s.e)	logsalhoraire	(s.e)	logsalhoraire	(s.e)	logsalhoraire	(s.e)
cat_origine								
- *fra1pimmi_Neurope*	-0.015	(0.021)	-0.004	(0.016)	0.011	(0.010)	0.016	(0.010)
- *fra1pimmi_europe*	0.013	(0.013)	0.008	(0.010)	0.014**	(0.007)	0.017**	(0.007)
- *fra2pfra*	ref.	(.)	ref.	(.)	ref.	(.)	ref.	(.)
- *immi_Nnatur_Neurope*	-0.141***	(0.028)	-0.130***	(0.022)	-0.042**	(0.014)	-0.073***	(0.012)
- *immi_Nnatur_europe*	-0.011	(0.024)	0.017	(0.017)	0.101***	(0.014)	0.076***	(0.011)
- *immi_natur_Neurope*	-0.111***	(0.025)	-0.093***	(0.018)	-0.050***	(0.011)	-0.043***	(0.011)
- *immi_natur_europe*	0.023	(0.033)	0.015	(0.023)	0.051**	(0.018)	0.065***	(0.016)

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Heckit 1 : sans iactifs

Heckit 2 : avec iactifs

La comparaison des différents modèles confirme la présence de biais de sélection. Tous les coefficients des modalités de la variable `cat_origine` de la régression heckit 1 sont plus faibles que ceux de la régression instrumentale. On a donc un biais de sélection de sélection positif : nous sous-estimions, dans la partie 1, l'effet de l'origine sur le salaire perçu. Les coefficients des deux modalités *immi_natur_Neurope* et *immi_Nnatur_Neurope* sont 2 à 3 plus significatifs (en ordre de grandeur) ! Le biais de sélection est en revanche plus faible pour les modalités européennes.

En conclusion, bien que l'interprétation quantitative de la régression sur l'indicatrice d'être actif occupé ne soit possible ici, ce modèle heckit confirme nos propos de la fin de la partie précédente (modèle "random effect probit"). C'est-à-dire (nous y reviendrons plus en détails dans la conclusion) :

- Le fait d'avoir des origines étrangères, et ce au sens large, augmente la probabilité d'être chômeur ou inactif. Le fait de ne pas être "français de souche" a donc des conséquences négatives sur l'employabilité des individus de notre base.
- Seulement, lorsque nous comparons les actifs occupés au travers du logarithme du salaire horaire, nous nous apercevons que seuls les individus d'origine (toujours au sens large) non-européenne présentent en moyenne des salaires plus faibles toutes choses égales par ailleurs. Ceci signifierait que les personnes originaires d'Europe sont plus employables, toutes choses égales par ailleurs.
- Enfin, les salaires augmentent en moyenne avec l'appartenance administrative au territoire français. La stabilité administrative semble donc jouer sur l'employabilité mais également sur le salaire.

Tableau 19 – Régression heckit de l'indicatrice de participation sur la population d'actifs et chômeurs avec ou sans prise en compte des inactifs

	(heckit sans inactifs)			(heckit avec inactifs)	
	actop	(s.e)		acteu	(s.e)
Nombre d'enfants	-0.046*	(0.023)	**Nombre d'enfants**	-0.069***	(0.018)
age	0.063***	(0.012)	**age**	0.065***	(0.010)
age^2	-0.001***	(0.000)	**age**^{2}	-0.001***	(0.000)
homme	-0.030	(0.034)	**homme**	0.059*	(0.028)
cat_origine			**cat_origine**		
- *fra1pimmi_Neurope*	-0.300***	(0.076)	- *fra1pimmi_Neurope*	-0.311***	(0.061)
- *fra1pimmi_europe*	-0.044	(0.070)	- *fra1pimmi_europe*	-0.080	(0.054)
- *fra2pfra*	ref.	(.)	- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.394***	(0.085)	- *immi_Nnatur_Neurope*	-0.444***	(0.066)
- *immi_Nnatur_europe*	-0.310***	(0.092)	- *immi_Nnatur_europe*	-0.282***	(0.076)
- *immi_natur_Neurope*	-0.432***	(0.082)	- *immi_natur_Neurope*	-0.364***	(0.050)
- *immi_natur_europe*	-0.068	(0.171)	- *immi_natur_europe*	-0.142	(0.069)
Nombre d'obervations	25 282		Nombre d'obervations	35 869	
Nombre de censures	1 065		Nombre de censures	10 543	
Mills			**Mills**		
lambda	0.48**		lambda	0.22**	
rho	1		rho	0.64	
sigma	0.48		sigma	0.34	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tableau 20 – Régression heckit du logarithme du salaire sur la population d'actifs et chômeurs avec ou sans prise en compte des inactifs

	(heckit sans inactifs) logsalhoraire	(s.e)	(heckit avec inactifs) logsalhoraire	(s.e)
pibhab_cat				
.1	ref.	(.)	ref.	(.)
.2	-0.072*** [1]	(0.016)	-0.068***	(0.015)
.3	-0.108***	(0.015)	-0.103***	(0.014)
tuu_r				
- *Commune rurale*	ref.	(.)	ref.	(.)
- *Uu < 20 000 habitants*	-0.014*	(0.007)	-0.010	(0.007)
- *Uu < 200 000 habitants*	0.020**	(0.007)	0.023**	(0.007)
- *Uu > 200 000 habitants*	0.023**	(0.007)	0.029***	(0.007)
- *Agglo. parisienne*	0.018	(0.017)	0.024	(0.016)
typvois				
- *Hors agglo.*	ref.	(.)	ref.	(.)
- *Lotissement, pavillons*	-0.001	(0.006)	0.001	(0.006)
- *Immeubles en ville*	-0.034***	(0.010)	-0.033***	(0.010)
- *Immeuble ou cité*	-0.054***	(0.013)	-0.049***	(0.012)
- *Habitat mixte*	-0.022	(0.013)	-0.025*	(0.012)
zus	-0.080***	(0.014)	-0.072***	(0.013)
typmen				
- *1 personne*	-0.033***	(0.008)	-0.029***	(0.008)
- *Famille monop.*	-0.041***	(0.010)	-0.041***	(0.011)
- *Couple sans enfant*	-0.030***	(0.006)	-0.025***	(0.006)
- *Couple avec enfant(s)*	ref.	(.)	ref.	(.)
- *Ménages (av. parenté)*	-0.105**	(0.038)	-0.069	(0.036)
- *Ménages (ss. parenté)*	-0.101*	(0.040)	-0.101**	(0.039)
- *Autres*	-0.140***	(0.020)	-0.134***	(0.019)
age	0.028***	(0.003)	0.030***	(0.002)
age^2	-0.000***	(0.000)	-0.000***	(0.000)
homme	0.093***	(0.005)	0.098***	(0.005)
cat_origine				
- *fra1pimmi_Neurope*	-0.015	(0.021)	-0.004	(0.016)
- *fra1pimmi_europe*	0.013	(0.013)	0.008	(0.010)
- *fra2pfra*	ref.	(.)	ref.	(.)
- *immi_Nnatur_Neurope*	-0.141***	(0.028)	-0.130***	(0.022)
- *immi_Nnatur_europe*	-0.011	(0.024)	0.017	(0.017)
- *immi_natur_Neurope*	-0.111***	(0.025)	-0.093***	(0.018)
- *immi_natur_europe*	0.023	(0.033)	0.015	(0.023)
csp_actif				
- *Non renseigné*	ref.	(.)	ref.	(.)
- *Cadres*	0.544***	(0.127)	0.607***	(0.127)
- *Professions int.*	0.328*	(0.126)	0.397***	(0.126)
- *Employés*	0.119	(0.124)	0.120	(0.124)
- *Ouvriers*	0.161	(0.122)	0.237	(0.122)

Nous contrôlons également l'éducation.

Nous contrôlons également l'éducation.

3bis. Modèle de durée

Cette partie facultative a initialement été motivée par notre volonté d'ajouter des variables de contrôle catégorielles uniquement disponibles que pour certaines catégories d'actifs (`contrat`, `typehoraire`, `heuretra_tranche`). En effet, nous pensions initialement qu'en considérant un modèle de durée, les valeurs manquantes de ces variables ne poseraient pas de problème comme précédemment. En effet, nous avons adopté le même cadre d'observation que dans le cours (partie "problème de censure"). Nous nous sommes placés dans la configuration où l'on observe en $t = 6$:

— La durée du dernier épisode de chômage pour ceux ayant retrouvé un emploi en $t = 6$.

— L'ancienneté du chômage pour les individus chômeurs en $t = 6$.

Nous avons ensuite censuré les observations des individus toujours chômeurs en $t = 6$ de telle sorte que, avec les notations du cours, l'on ait :

$$C = t_6 - E \text{ et } D = \mathbb{1}\{\texttt{ancchom} + E < t_6\}$$

avec t_6, E, D respectivement la dernière date d'interrogation, la date effective d'entrée au chômage $\in [\text{"}-\infty\text{"}, t_6[$, et l'indicatrice de non-censure. On observe donc que les individus qui ont retrouvé du travail en $t = 6$, les autres sont censurés. Enfin, nous supposons que C est indépendant de $\texttt{ancchom}|X$ même si cette hypothèse est questionnable. En effet, celle-ci suppose que l'ancienneté du chômage est indépendante de la date d'entrée conditionnellement à X. Un individu rentré au chômage en été aura plus de mal (ou moins) à retrouver un travail que s'il était entré au chômage en octobre. Comme le souligne le cours, cette hypothèse est « potentiellement correcte sur le court terme »ce qui est notre cas ici puisque l'interrogation se fait sur 6 trimestres.

Après avoir obtenu les résultats de notre modèle de durée, nous nous sommes aperçus qu'ajouter les variables de type `contrat`, `typehoraire`, `heuretra_tranche` `expe_specifique` était bel est bien possible mais les individus non-censurés, c'est à dire ceux toujours chômeurs en $t = 6$ avaient été enlevé de la modélisation par Stata. En effet, comme notre modèle final ne tient compte que de la date $t = 6$, les individus toujours chômeurs en $t = 6$ ($D = 0$ ou `failed` $= 0$) ne possédaient pas ce genre d'informations par construction. De fait, le nombre d'observations était égal au nombre d'individus pour lesquels $D = 1$ (actifs en $t = 6$). Stata retournait donc "No. of subjects = 976" et"No. of failures = 976". Le même raisonnement s'appliquait si nous ajoutions la variable `halochomage`, mais dans le sens inverse. Seuls les individus toujours chômeurs en $t = 6$ étaient restants. Logiquement, l'ajout des variables `contrat`, `typehoraire`, `heuretra_tranche` et `informel` provoquait une erreur de type "No observations". De fait, nous avons pris une moyenne de ces informations sur l'ensemble de la période et les avons synthétisés grâce à la création de nouvelles variables non manquantes en $t = 6$ (`contrat2`, `typehoraire2`, `heuretra_tranche2`, `informel2`). Comment interpréter le fait que le ratio de hasard associé à la variable `expe_specifique` est plus grand que 1 ? Cela signifie que plus l'ancienneté au travail occupé avant d'être chômeur était élevée, plus le risque de retrouver un emploi augmente, *id est* plus la durée de chômage diminue. Passons donc à l'interprétation des résultats (tableau 21).

Le premier constat malheureux est qu'aucune des variables ou presque n'est significative à 5%... Ce qui contraint sérieusement la légitimité de nos interprétations. Néanmoins les résultats sont très intéressants et nous vous proposons de passer en revue certains d'entre eux. Tout d'abord, le taux de hasard proportionnel estimé est de 1.318 lorsque l'on prend en compte l'hétérogénéité inobservée, contre 1.097 dans le cas contraire. Ceci témoigne du fait qu'une fois que les différences individuelles sont prises en compte, plus la durée de chômage augmente et plus les individus auront de la chance de retrouver un emploi. Sans cette approche, il est difficile de séparer l'hétérogénéité inobservée de la dépendance temporelle. En effet, les individus très qualifiés mettent en moyenne très peu de temps à retrouver un emploi (chômage frictionnel). Il ne reste donc que les chômeurs de longue durée très peu qualifiés (chômage structurel) , ce qui biaise les résultats. La différence entre nos deux taux de hasard confirme ce biais. Néanmoins, nous rejetons fortement l'hypothèse d'absence d'hétérogénéité inobservée.

— Taux de hasard : 1.318
— LR test de "$\theta = 0$" : 42.17
— Test de nullité conjointe : 124.09

Tableau 21 – Modèle de durée sur la durée du chômage

	(1) ancchom	(p-value)
pibhab_cat		
.1	1.000	(.)
.2	1.068	(0.843)
.3	0.897	(0.732)
tuu_r		
- *Commune rurale*	1.000	(.)
- *Uu < 20 000 habitants*	1.087	(0.527)
- *Uu < 200 000 habitants*	1.090	(0.535)
- *Uu > 200 000 habitants*	1.090	(0.544)
- *Agglo. parisienne*	1.408	(0.338)
typvois		
- *Hors agglo.*	1.000	(.)
- *Lotissement, pavillons*	0.818	(0.116)
- *Immeubles en ville*	0.763	(0.106)
- *Immeuble ou cité*	0.722	(0.090)
- *Habitat mixte*	0.973	(0.895)
zus	0.987	(0.942)
typmen		
- *1 personne*	0.887	(0.350)
- *Famille monop.*	0.663**	(0.003)
- *Couple sans enfant*	0.911	(0.442)
- *Couple avec enfant(s)*	1.000	(.)
- *Ménages (av. parenté)*	0.732	(0.473)
- *Ménages (ss. parenté)*	1.280	(0.600)
- *Autres*	0.449*	(0.041)
csp_parents		
- *Non renseigné*	0.664	(0.063)
- *Agriculteurs expl.*	1.273	(0.267)
- *Artisans, com., c.e.*	1.128	(0.390)
- *Cadres*	0.971	(0.838)
- *Professions int.*	0.944	(0.688)
- *Employés*	1.108	(0.441)
- *Ouvriers*	1.000	(.)
contrat		
- *Pas de contrat*	1.000	(.)
- *CDI*	1.426	(0.327)
- *CDD*	1.647	(0.166)
- *Contrat saisonnier*	1.457	(0.353)
- *Interim*	2.636*	(0.012)
- *Apprentissage*	5.324**	(0.001)
plract	1.337	(0.083)
encadre	1.025	(0.887)
expe_specifique	1.002	(0.063)
informel	0.460	(0.181)

	(1) ancchom	(p-value)
typehoraire		
- *Stable*	1.000	(.)
- *Alterné*	1.139	(0.566)
- *Variable*	1.383**	(0.007)
- *Sans objet*	0.827	(0.589)
heuretra_tranche		
- *Moins de 15h*	1.000	(.)
- *Entre 15 et 30h*	1.606	(0.077)
- *Entre 30 et 35h*	1.617	(0.149)
- *Entre 35 et 40h*	2.108**	(0.005)
- *40h et plus*	3.419***	(0.000)
age	0.980	(0.535)
age^2	1.000	(0.982)
homme	0.956	(0.641)
Diplôme		
Diplôme supérieur	1.000	(.)
Bac +2	0.790	(0.158)
Bac ou brevet pro.	0.820	(0.222)
CAP, BEP	0.809	(0.192)
Brevet des collèges	1.158	(0.520)
Aucun diplôme	0.726	(0.073)
cat_origine		
- *fra1pimmi_Neurope*	0.970	(0.892)
- *fra1pimmi_europe*	0.966	(0.852)
- *fra2pfra*	1.000	(.)
- *immi_Nnatur_Neurope*	0.954	(0.813)
- *immi_Nnatur_europe*	1.239	(0.446)
- *immi_natur_Neurope*	0.534*	(0.012)
- *immi_natur_europe*	0.510	(0.200)
Proportions minorités		
- *prop_678910_immi*	1.220	(0.869)
- *prop_02_immi*	0.469	(0.578)
- *prop_678910*	0.923	(0.922)
- *prop_02*	1.259	(0.761)
csp_actif		
- *Non renseigné*	1.000	(.)
- *Cadres*	32.299**	(0.003)
- *Professions int.*	14.102*	(0.022)
- *Employés*	16.574*	(0.014)
- *Ouvriers*	15.922*	(0.015)
Nombre d'observations	1 290	
Nombre d'obs. non censurées	840	

Pour le reste, nous observons que :

- Le risque instantané de retrouver un emploi en agglomération parisienne est 1,4 fois plus élevé qu'en commune rurale.
- Habiter en immeuble (cité ou ville) diminue le risque instantané de retrouver un emploi d'en moyenne 25%. La durée de chômage est donc plus longue en comparaison.
- Le fait d'exercer plusieurs activité **avant d'être chômage** diminue la durée moyenne de chômage. Ceci est intuitif dans la mesure où si l'individu exerçait plusieurs activités, celui-ci est plus enclin à proposer des compétences diversifiées aux employeurs, et l'on peut naturellement penser que ce dernier est plus productif.
- L'expérience spécifique et le fait d'encadrer une équipe au moment de l'activité jouent très peu sur la diminution ou l'augmentation de la durée du chômage. Ceci semble montrer que la durée du chômage ne dépend pas du nombre de mois d'expérience dans l'entreprise précédente. Cette idée est contre-intuitive, on pourrait penser qu'un individu ayant eu une longue ancienneté dans son entreprise passée peut faire valoir cette expérience auprès des employeurs. Néanmoins, une longue expérience est corrélée avec un âge «avancé», ce qui peut aussi signifier que les recruteurs sont réticents à l'idée d'engager une personne plus âgée.
- Le fait d'exercer une activité non professionnelle augmente considérablement la durée du chômage, ce qui est normal dans la mesure où l'individu ne souhaitera pas retravailler si son activité non professionnelle lui apporte satisfaction voire revenus.
- Plus l'individu avait des horaires soutenus dans son travail précédent (avant d'être chômeur), plus sa chance de retrouver un emploi rapidement est élevée. D'ailleurs, ces modalités sont significatives. Il s'agirait ici davantage de chômage frictionnel.
- L'âge et le fait d'être un homme joue peu sur le risque instantané de retrouver du travail.
- Être diplômé du supérieur offre plus de chance de retrouver un emploi rapidement que tous les autres types de diplômes.
- De même pour le fait d'être un cadre plutôt qu'employé, ouvrier ou avoir une profession intermédiaire.
- Les observations sur la variable `cat_origine` ne sont pas fiables ni interprétables de manière cohérente.
- En revanche, et pour finir, les proportions d'immigrés jouent beaucoup mais l'interprétation des résultats est assez floue par rapport aux tendances dégagées dans les parties précédentes. En effet, la présence de minorités non-européennes et immigrées réduit en moyenne la durée de chômage, mais ceci n'est pas vrai pour les non-européens non immigrés... Et inversement pour les minorités européennes.

En conclusion, malgré le potentiel du modèle, force est de constater que l'interprétation de nos résultats n'est pas fiable et parfois contradictoire ou inexploitable dans un but de politiques publiques. Le faible nombre d'observations (1 290) permet entre autres de l'expliquer.

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 ?

Le but de ce projet a été d'étudier les potentielles discriminations liées à l'origine sur le marché du travail via l'interrogation de 70 944 individus. Pour cela, nous nous sommes intéressés à la fois à la marge extensive - actif occupé, chômeur, inactif - et à la marge intensive - le salaire. L'avantage fondamental de notre base de donnée résidait dans le suivi, sur environ 2 ans, des individus interrogés à 6 périodes successives. L'approche extensive se complète à celle intensive dans la mesure où les discriminations, si elles ont lieu, sévissent dès l'accès à l'embauche et peuvent se prolonger sur le salaire. L'effet de l'origine a été étudié successivement au travers de deux variables distinctes : l'origine individuelle de l'individu (`cat_origine`) et les densités de populations minoritaires étrangères dans son voisinage (`prop_*` et `prop_*_immi`).

Nos diverses estimations n'ont pas montré d'impact significatif de ces proportions sur la marge extensive. Plus précisément, la mixité sociale dans laquelle se trouve l'individu interrogé n'affecte pas ou peu sa probabilité d'être actif (ou chômeur). Pour rappel, notre modèle "random effect probit" - le plus adéquat, rendait compte du fait que l'augmentation de 10% des proportions de personnes d'origine non-européenne et non-immigrées augmentait en moyenne la probabilité des individus sondés d'être chômeurs de 0,64%, toutes choses égales par ailleurs, en particulier à niveau d'éducation et de domaine professionnel d'activité équivalents. Ce chiffre passait à 0,28%, 0,14% et 0,05% pour les proportions respectives d'immigrés européens, d'immigrés non-européens et de personnes originaires d'Europe. De plus, ces trois derniers coefficients ne sont pas significatifs à 5%. L'environnement social (au sens où nous l'entendons dans notre étude) ne semble pas influer sur les chances de retrouver un emploi. Cette assertion peut sembler contre-intuitive dans la mesure où, nous l'avons vu en préambule de ce projet, le tableau 5 montrait une tendance marquée d'un regroupement des individus entre mêmes origines. Dans la mesure où (nous y reviendrons juste après) la variable `cat_origine` influence le salaire et la probabilité d'être actif, pourquoi ne pas observer de tels résultats chez les variables de proportions au regard de la corrélation entre les deux variables ? C'est là la grande différence entre corrélation et causalité. Du fait de nos nombreux contrôles, nous pouvons affirmer que malgré cette corrélation, toutes choses égales par ailleurs, les proportions de minorités ne jouent pas sur la probabilité d'être actif ou chômeur d'un point de vue causal (même si nous ne pourrons jamais affirmer capter un effet causal "pur"). Nous pouvons penser que, déjà, un individu français né en France de parents français peut très bien habiter dans une zone résidentielle très brassée dans la mesure où précarité sociale et mixité sociale vont de paire. Mais si l'individu est chômage, ce n'est certainement pas à cause des minorités qui l'entourent, mais avant tout de son manque de productivité ou d'aptitudes professionnelles qui l'ont peut-être poussées à habiter dans un immeuble en ville ou en cité. De fait, l'effet des proportions dont nous parlons est davantage capté par la qualité du milieu résidentiel.

L'analyse de la variable `cat_origine` apporte des résultats plus pertinents à la fois dans la marge extensive et intensive. Cependant, il faut bien distinguer deux tendances. D'une part, le fait d'avoir des origines étrangères, immigré ou non, augmente la probabilité d'être chômeur ou inactif. C'est ce que révèle les modèle de sélection et "random effect probit" (tableau 16). Ce dernier modèle statue que, par rapport à la situation d'être français né en France de parents français, posséder des origines étrangères, augmente en moyenne la probabilité d'être au chômage d'entre 0,5% et 7,1%. En particulier, les individus originaires de pays extérieurs à l'Europe présentent des coefficients plus élevés : 7,1%, 2,4% et 2,3%. Les individus d'Afrique Noire et du Maghreb sont les plus sévèrement touchés par ce phénomène. Nous n'avons pas considéré leurs modalités explicitement dans les régressions pour gagner en significativité et il ne serait pas difficile d'étudier plus précisément leurs effets respectifs, la démarche de modélisation restant la même. Ici, nous nous focalisons surtout sur le fait d'être Européen ou non. En conclusion, ne pas être "français de souche" a donc des conséquences négatives sur l'employabilité des individus de notre base, et ce quelque soit l'origine mais avant tout, et de manière significative, sur les non-européens.
Nous l'avons dit, il faut bien distinguer marge extensive et intensive ici. En effet, à la fois notre premier modèle instrumenté et notre modèle de sélection témoignent d'écarts de salaire inégaux, en terme de signes, selon l'origine. De fait, d'après notre modèle de sélection, un individu immigré non-européen touchera en moyenne 14,1% moins

en salaire horaire qu'un français de référence, toutes choses égales par ailleurs. Ce chiffre passe à 11,1% pour un immigré non-européen naturalisé et enfin à 1,5% pour un français ayant au moins un parent non-européen. Les individus originaires d'Europe, eux, gagnent en moyenne plus qu'un français de référence. Bien que non significatives statistiquement, ces hausses vont de 1,3 à 2,3%, mais les immigrés européens non naturalisés touchent en moyenne 1,1% moins.

Si nous confrontons ces deux aspects, nous pouvons conclure qu'avoir des origines étrangères joue négativement sur l'employabilité des individus mais que les écarts de salaires dépendent du fait d'être européen ou non. Nous pouvons alors penser que l'origine est plus discriminatoire au moment de l'embauche qu'elle ne l'est une fois l'individu embauché. D'ailleurs, cette affirmation semble faire sens. Les individus non-européens souffrent davantage des discriminations liées à la couleur de peau et aux préjugés raciaux que ceux européens. Néanmoins, être immigré ou juste ne pas être un français de "référence" joue négativement sur l'employabilité ; ce qui semble témoigner de réticences de la part des employeurs à engager une personne dont le statut administratif est différent, ceci à niveau de diplôme équivalent ! Quant au salaire, nous pouvons penser que les Européens d'origine sont plus susceptibles d'être des expatriés. S'ils tardent peut-être à trouver un emploi du fait de la régularisation de leur situation, les salaires touchés sont en moyenne plus élevés puisque leurs compétences et leur diplôme sont mieux valorisés que les individus non-européens. De fait, nos estimations sur les modalités européennes captent sûrement plusieurs effets dans la mesure où la non-significativité statistique ne vient pas d'un problème de manque d'observations. En effet, la modalité `fra1pimmi_europe` est la plus représentée, derrière bien sûr celle `fra2pfra` puisqu'elle représente 5,3% des observations. Les expatriés doivent donc contrebalancer un phénomène d'exode d'immigrés très pauvres Européens de l'Est fuyant la guerre ou la pauvreté.

En conclusion, nous concluons à la présence d'inégalités liées à l'origine des individus. Le chiffre clé de notre étude est qu'un individu immigré non-européen touchera en moyenne 14,1% moins qu'un français de référence et aura 7,1% plus de chance d'être au chômage du fait de son origine. Ceci toutes choses égales par ailleurs, c'est-à-dire en contrôlant l'éducation et le domaine d'activité, et en contrôlant l'endogénéité grâce à notre modèle "random effect probit". Le manque de significativité statistique concernant les individus d'origine européenne ne permet pas de conclure de manière sûre sur l'impact de cette origine sur le salaire. En revanche, il est certain que les inégalités sont nettement plus marquées chez les non-européens, et nettement plus significatives. Nous suspectons des discriminations liées à la mauvaise reconnaissance des diplômes étrangers, la couleur de peau, et le statut administratif. En effet, l'appartenance au territoire français (à savoir, immigré, immigré naturalisé, français) joue sur l'ampleur des discriminations et montre l'apport bénéfique de la naturalisation sur les salaires et l'employabilité. Grâce à nos nombreux contrôles, nous pouvons statuer sur d'éventuelles actions de politique publique pour pallier ce phénomène inégalitaire. Déjà, renforcer et accélérer les procédures de naturalisation ou simplement de régularisation de manière à rassurer les employeurs. Nous pourrions également songer à imposer des quotas de personnes issues de nationalités étrangères, en particulier pour les non-européens, à l'instar des seuils de parité hommes/femmes réglementaires ou de personnes en situation de handicap. De plus, l'équivalence des diplômes doit être systématisée ou moins notifiée officiellement afin de donner plus de crédit à la formation d'origine. Tout ceci n'est possible que si l'individu souhaite s'intégrer. En ce sens, nous remarquons une tendance systématique de rapprochement des individus de mêmes origines entre eux. Malgré l'absence de résultats probants relatifs aux variables de proportions de minorités, nous pouvons penser que le fait d'être entouré de personnes de mêmes origines, immigrés ou non, impacte la situation des individus sur le marché du travail. Cette assertion est néanmoins à prendre avec délicatesse et il est tout aussi plausible que les individus immigrés, généralement plus pauvres, soient dirigés vers des logements moins chers, souvent précaires, plutôt que d'être dispersés dans des zones résidentielles différentes et espacées. Les immeubles en ville ou en cité concentrent davantage de pauvreté et de minorités sans pour autant que la présence de ces minorités soit la cause d'un salaire ou d'une employabilité plus faibles. Le problème de communautarisme n'est pas nouveau et désenclaver les cités est une chose très difficile. Nous pourrions néanmoins imaginer que, selon l'origine, le niveau de pauvreté, le statut marital et familial, des bourses d'installation soient données aux personnes étrangères pour faciliter leur intégration. Dans tous les cas, nous ne pouvons affirmer que l'environnement social est la cause des discriminations observées plutôt que l'origine en elle-même.

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

Un premier point noir est l'absence de significativité statistique des modalités "européennes". Nous l'avons dit, nous suspectons deux tendances opposées chez les immigrés européens : d'un côté l'expatriation, synonyme de salaires plus élevés et de compétences mieux reconnues, de l'autre la fuite de l'extrême pauvreté. Il aurait été intéressant de connaître le salaire des individus dans leurs pays d'origine quitte à appliquer une conversion avec un taux de change.

Ensuite, il aurait était pertinent d'ajouter le niveau de français qui est susceptible d'affecter sensiblement l'employabilité. De même pour le degré de motivation, certes difficilement quantifiable. Les individus auraient pu également être interrogés, au travers d'une note de 1 à 10, sur leur volonté d'intégration dans le pays. Cette démarche est idéaliste dans la mesure où beaucoup d'individus auront tendance à sur-estimer cette note (si 10 est "je veux devenir français") ou simplement ne pas être en mesure d'en donner (il semble difficile de quantifier son envie d'intégration au travers d'une note). Nous pourrions cependant les interroger sur comment ceux-ci se sentent intégrés.

Enfin, nous souhaitons parler de la sélection des individus sondés. Vous l'avez dit dans l'énoncé, « les individus 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 **aléatoirement** 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** ». Plusieurs choses sont à retenir de cette brève description de la méthodologie de l'enquête. Tout d'abord, bien que les logements soient tirés de manière aléatoire, ce qui est optimal dans la recherche d'effets causaux, nous n'avons accès qu'aux individus n'ayant pas déménagé sur l'ensemble de la période, soit environ 2 ans. De fait, il n'est pas possible de capter des phénomènes d'enrichissement ou d'appauvrissement au travers de l'évolution du logement. Nous pouvons penser qu'un immigré non-européen occupait en $t = 1$ un logement précaire puis a trouvé un emploi et a pu déménager. L'absence de déménagement conditionne donc d'une certaine manière les résultats dans la mesure où les situations des individus sont moins amenées à fortement évoluer. Nuançons tout de même cette phrase en précisant bien que la période d'étude reste assez courte, 2 ans environ. Ensuite, les individus sondés sont ceux qui ont bien voulu répondre à l'enquête. En ce sens, ces derniers ont eu le choix. De fait, il existe ici un biais de sélection (peut-être pas compensé par la sélection aléatoire des logements) selon lequel les individus les plus pauvres ou les plus riches n'ont pas souhaité divulguer un certain nombre d'informations (la base contient tout de même 81 variables). Nous pouvons supposer que nous n'observons pas les situations plus extrêmes de pauvreté. Ceci peut avoir un impact sur les coefficients des variables d'origine ou de proportions de minorités. On s'attend à avoir un biais de sélection négatif en le sens où nous sous-estimons l'effet de l'origine sur le salaire ou l'employabilité. Cet aspect n'est pas négligeable.

Annexes

Annexe 1

Tableau 22 – Régression heckit de l'indicatrice de participation sur la population d'actifs et chômeurs

	(heckit) actop	(s.e)
Nombre d'enfants	-0.046*	(0.023)
pibhab_cat		
.1	ref.	(.)
.2	0.050	(0.111)
.3	-0.046	(0.104)
tuu_r		
- *Commune rurale*	ref.	(.)
- *Uu < 20 000 habitants*	-0.123**	(0.047)
- *Uu < 200 000 habitants*	-0.108*	(0.050)
- *Uu > 200 000 habitants*	-0.110*	(0.048)
- *Agglo. parisienne*	0.214	(0.121)
typvois		
- *Hors agglo.*	ref.	(.)
- *Lotissement, pavillons*	-0.058	(0.048)
- *Immeubles en ville*	-0.258***	(0.061)
- *Immeuble ou cité*	-0.277***	(0.071)
- *Habitat mixte*	-0.164*	(0.082)
zus	-0.194**	(0.064)
typmen		
- *1 personne*	-0.296***	(0.062)
- *Famille monop.*	-0.344***	(0.059)
- *Couple sans enfant*	-0.151*	(0.060)
- *Couple avec enfant(s)*	ref.	(.)
- *Ménages (av. parenté)*	-0.122	(0.246)
- *Ménages (ss. parenté)*	-0.338	(0.216)
- *Autres*	-0.342**	(0.121)
age	0.063***	(0.012)
age^2	-0.001***	(0.000)
homme	-0.030	(0.034)
cat_origine		
- *fra1pimmi_Neurope*	-0.300***	(0.076)
- *fra1pimmi_europe*	-0.044	(0.070)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.394***	(0.085)
- *immi_Nnatur_europe*	-0.310***	(0.092)
- *immi_natur_Neurope*	-0.432***	(0.082)
- *immi_natur_europe*	-0.068	(0.171)
csp_actif		
- *Non renseigné*	ref.	(.)
- *Cadres*	1.739***	(0.373)
- *Professions int.*	1.681***	(0.372)
- *Employés*	1.515***	(0.371)
- *Ouvriers*	1.328***	(0.371)
Nombre d'obervations	25 282	
Nombre de censures	1 065	
Mills		
lambda	0.33	
rho	0.95	
sigma	0.37	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Annexe 2

Tableau 23 – Régression heckit de l'indicatrice de participation sur la population d'actifs occupés et inactifs (dont chômeurs)

	(heckit) acteu	(s.e)
Nombre d'enfants	-0.069***	(0.018)
pibhab_cat		
.1	ref.	(.)
.2	0.051	(0.089)
.3	-0.023	(0.084)
tuu_r		
- *Commune rurale*	ref.	(.)
- *Uu < 20 000 habitants*	-0.114**	(0.038)
- *Uu < 200 000 habitants*	-0.115**	(0.040)
- *Uu > 200 000 habitants*	-0.109**	(0.039)
- *Agglo. parisienne*	00.166	(0.096)
typvois		
- *Hors agglo.*	ref.	(.)
- *Lotissement, pavillons*	-0.053	(0.038)
- *Immeubles en ville*	-0.253***	(0.049)
- *Immeuble ou cité*	-0.262***	(0.057)
- *Habitat mixte*	-0.171**	(0.065)
zus	-0.180***	(0.052)
typmen		
- *1 personne*	-0.356***	(0.050)
- *Famille monop.*	-0.424***	(0.046)
- *Couple sans enfant*	-0.146**	(0.049)
- *Couple avec enfant(s)*	ref.	(.)
- *Ménages (av. parenté)*	-0.308	(0.180)
- *Ménages (ss. parenté)*	-0.527**	(0.164)
- *Autres*	-0.478***	(0.091)
age	0.065***	(0.010)
age^2	-0.001***	(0.000)
homme	0.059*	(0.028)
cat_origine		
- *fra1pimmi_Neurope*	-0.311***	(0.061)
- *fra1pimmi_europe*	-0.080	(0.054)
- *fra2pfra*	ref.	(.)
- *immi_Nnatur_Neurope*	-0.444***	(0.066)
- *immi_Nnatur_europe*	-0.282***	(0.076)
- *immi_natur_Neurope*	-0.364***	(0.050)
- *immi_natur_europe*	-0.142	(0.069)
csp_actif		
- *Non renseigné*	ref.	(.)
- *Cadres*	2.006***	(0.288)
- *Professions int.*	1.878***	(0.286)
- *Employés*	1.677***	(0.286)
- *Ouvriers*	1.480***	(0.286)
Nombre d'obervations	35 869	
Nombre de censures	10 543	
Mills		
lambda	0.10	
rho	0.31	
sigma	0.33	

t statistique entre parenthèses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table des figures

Liste des tableaux

Références

[1] Afsa Cédric (2016), "Le modèle Logit. Théorie et application", Document de travail "Méthodologie Satistique" de la Direction de la Méthodologie et de la Coordination Statistique et Internationale, INSEE.

[2] Bruno Anne-Sophie (2010), "Analyser les écarts de salaires à l'aide des modèles de régression. Vertus et limites d'une méthode. Le cas des migrants de Tunisie en région parisienne après 1956", *Histoire & mesure* [En ligne], mis en ligne le 01 janvier 2011, disponible sur `http://journals.openedition.org/histoiremesure/3984`.

[3] Courgeau Daniel & Baccaïni Brigitte (1997), "Analyse mutli-niveaux en sciences sociales" [En ligne], p. 831-863, disponible sur `https://www.persee.fr/doc/pop_0032-4663_1997_num_52_4_6470`.

[4] Desplatz Rozenn & Ferracci Marc (2016), "Comment évaluter l'impact des politiques publiques ? Un guide à l'usage des décideurs et praticiens.", France Statégie.

[5] Letrait Muriel (2002), "L'utilisation par les chômeurs du temps libéré par l'absence d'emploi", *Économie et statistique*, N. 352-353, CNRS.

[6] Piketty Thomas (2001), "L'économie des inégalités", chap. 3, p. 64.