Ministère de l'éducation nationale, de la jeunesse et des sports

Direction de l'évaluation, de la prospective et de la performance

Document de travail

Série « Méthodes »

N°2021-M01

InserJeunes – la valeur ajoutée des établissements sur le taux d'emploi

Guilhem Deschamps
Loïc Midy

Ce document décrit la méthodologie de calcul de la valeur ajoutée des établissements sur le taux d'emploi, à la sortie des centres de formation d'apprentis et des lycées professionnels.

Introduction

La valeur ajoutée des établissements sur le taux d'emploi est un des indicateurs calculés par InserJeunes pour répondre à la loi du 5 septembre 2018 pour la liberté de choisir son avenir professionnel. Les établissements font ici référence aux centres de formation d'apprentis (CFA) et aux lycées professionnels. De même lorsque nous parlons d'élèves nous faisons allusion aux élèves et aux apprentis.

La première partie du document définit la valeur ajoutée et ses deux composantes, le taux d'emploi et le taux d'emploi attendu. Le taux d'emploi attendu est un taux d'emploi moyen issu de modèles économétriques. Ces modèles, appelés modèles multiniveaux, sont brièvement présentés en deuxième partie de ce document. L'objectif est d'avoir une compréhension suffisante de ces modèles pour appréhender le calcul du taux d'emploi attendu, indicateur déterminant dans la valeur ajoutée d'un établissement. La troisième partie détaille l'élaboration et le contenu des modèles multiniveaux utilisés dans InserJeunes pour le calcul de la valeur ajoutée des établissements sur le taux d'emploi. En annexes se trouvent les tableaux détaillés des modèles lors de leur élaboration ainsi qu'une description de leur mise en application sur le logiciel SAS et sur R.

Table des matières

1) La valeur ajoutée des établissements sur le taux d'emploi

Définition

La valeur ajoutée d'un établissement sur le taux d'emploi est un indicateur qui permet de comparer de façon pertinente le taux d'emploi des élèves sortants de cet établissement au taux d'emploi d'établissements similaires.

Une simple comparaison des taux d'emploi calculés entre établissements ne suffit pas, l'insertion dépendant en partie de facteurs extérieurs à l'établissement : profil des jeunes, type et niveau de formation, spécialité de formation et marché du travail local. Par exemple, comparer directement le taux d'emploi d'un établissement n'offrant que des formations en CAP avec celui d'un établissement n'offrant que des formations en BTS est peu pertinent.

Lorsque l'on se situe au niveau des établissements on souhaite plutôt savoir si un établissement a un impact sur l'insertion professionnelle de ses élèves, et si cet impact est plus ou moins important comparé à d'autres établissements similaires. Pour mesurer cet impact propre à l'établissement, la valeur ajoutée sur le taux d'emploi s'efforce d'éliminer l'incidence des facteurs d'insertion professionnelle extérieurs à l'établissement pour conserver ce qui est dû à l'action propre de l'établissement. Les facteurs extérieurs à l'établissement qui ont le plus d'impact sur l'insertion des lycéens professionnels et des apprentis sont les caractéristiques scolaires et sociodémographiques des élèves, le type et le niveau de formation, et le taux de chômage de la zone d'emploi de résidence de l'élève.

Une méthodologie similaire est déjà appliquée dans le cadre de la diffusion des indicateurs, nommés Indicateurs de valeur ajoutée des lycées (IVAL), produits par la Depp depuis de nombreuses années pour les lycées. Les IVAL mesurent la valeur ajoutée des établissements sur le taux de réussite au baccalauréat, le taux d'accès au baccalauréat et le taux de mentions au baccalauréat.

Calcul de la valeur ajoutée

Pour cerner l'action propre de l'établissement, il est nécessaire de calculer pour chaque établissement un taux d'« emploi attendu à 6 mois » qui correspond au taux d'emploi moyen des élèves accueillis dans des établissements comparables en termes de profil des élèves, de formations dispensées et domiciliés dans une zone d'emploi au taux de chômage comparable. Comme on s'intéresse au croisement de toutes les variables qui ont un impact sur le taux d'emploi, ce taux moyen est calculé à partir de modèles économétriques. Les modèles de régression par les moindres carrés ordinaires ne sont pas adaptés car ils ne permettent pas de mesurer l'effet des établissements sur les élèves tout en prenant en compte le taux de chômage de la zone d'emploi, c'est pourquoi nous utilisons des modèles multiniveaux.

A la première diffusion début 2021 des résultats d'InserJeunes, l'emploi est mesuré sur le champ du salariat privé en France. Il ne mesure donc pas l'emploi à l'étranger, l'emploi non salarié, l'emploi public, l'emploi auprès de particuliers employeurs ou à l'aide des titres emploi simplifié agricole (TESA). A terme, l'ensemble du champ salarié (public et privé, y compris auprès des particuliers employeurs ou réalisés à l'aide du TESA) sera couvert. Dans la suite de ce document on parlera simplement du taux d'emploi.

La valeur ajoutée est égale à **la différence** entre le taux d'emploi observé à 6 mois de l'établissement (le taux réel) et le taux d'emploi attendu à 6 mois (le taux issu d'un modèle statistique) :

Valeur ajoutée de l'établissement sur le taux d'emploi à 6 mois	=	Taux d'emploi observé à 6 mois de l'établissement	-	Taux d'emploi attendu à 6 mois de l'établissement

Le taux d'emploi observé à 6 mois est le taux d'emploi mesuré par InserJeunes via l'appariement des sortants de formation avec les contrats des salariés issus de la Déclaration Sociale Nominative (DSN).

Le taux d'emploi attendu à 6 mois est la moyenne, au niveau de l'établissement, des probabilités estimées par des modèles multiniveaux que les élèves de cet établissement soient en emploi à 6 mois.

La valeur ajoutée, différence entre deux taux, s'exprime en point de pourcentage. Par exemple, un établissement avec un taux d'emploi observé à 65 % et un taux d'emploi attendu à 60 % aura une valeur ajoutée de 5 (points).

Théoriquement, la valeur ajoutée peut prendre des valeurs comprises entre -100 et 100 : au maximum si le taux d'emploi observé est de 100 % et le taux d'emploi attendu est de 0 % la valeur ajoutée est égale à 100 ; et au minimum si le taux d'emploi observé est de 0 % et le taux d'emploi attendu est de 100 % la valeur ajoutée est égale à -100. Dans les faits, ces extrêmes ne sont jamais atteints dans InserJeunes : les valeurs ajoutées sont comprises entre 40 et -40 et environ 90 % des établissements et des CFA ont une valeur ajoutée comprise entre 12 et -12.

Interprétation

Indépendance entre les trois grands types d'établissements

Trois types d'établissements sont concernés par InserJeunes : les CFA, les lycées professionnels du ministère en charge de l'Education nationale et les lycées professionnels du ministère en charge de l'Agriculture. Le taux d'emploi attendu, et par conséquent la valeur ajoutée de l'établissement sur le taux d'emploi, est calculé **de manière indépendante pour chacun de ces trois types d'établissements.** En effet ces types d'établissement ont leur spécificité en matière de formation, de diplôme, d'insertion et débouchés professionnels, et les sources de données étant différentes les variables intégrées aux modèles peuvent varier.

Par construction, la moyenne des valeurs ajoutées est nulle pour chacun des trois types d'établissements.

Une mesure relative

La valeur ajoutée de l'établissement sur le taux d'emploi compare le taux d'emploi observé d'un établissement avec le taux d'emploi moyen des établissements qui lui ressemblent. Il s'agit donc d'une mesure relative qui peut prendre des valeurs positives ou négatives. Chaque établissement est ici comparé avec des établissements du même type : la comparaison des valeurs ajoutées entre CFA et lycées professionnels n'a donc aucun sens.

La valeur ajoutée est **positive** lorsque le taux d'emploi à 6 mois observé est supérieur au taux d'emploi à 6 mois attendu. Cela signifie que les élèves sortants de l'établissement s'insèrent, en moyenne, mieux que les élèves sortants ayant les mêmes caractéristiques individuelles, issus des mêmes formations et cherchant un emploi dans une zone d'emploi avec un taux de chômage similaire. Dans ce cas, l'apport propre de l'établissement dans l'insertion professionnelle de ses élèves est positif.

Elle est **négative** dans le cas contraire.

Par exemple, un établissement qui obtient des résultats d'insertion professionnelle moyens dans une zone d'emploi avec un taux de chômage élevé et avec des élèves socialement défavorisés joue probablement un rôle important et positif dans l'insertion de ces jeunes. Le taux d'emploi observé de cet établissement sera alors supérieur à son taux d'emploi attendu et sa valeur ajoutée sera positive.

2) Les modèles multiniveaux, généralités et intérêt

Le descriptif présenté dans cette partie vise à donner les principales caractéristiques des modèles multiniveaux et leurs principales différences avec les modèles usuels de régression par les moindres carrés ordinaires. Les explications, exemples et graphiques sont en majeure partie issus des articles de Pascal Bressoux (2007 et 2008) ainsi que du document de méthodologie statistique de l'INSEE, « Les modèles multiniveaux », de Pauline Givord et Marine Guillerm (2016). Les références précises sont disponibles dans la bibliographie.

Intérêts

Les modèles multiniveaux sont des modèles économétriques qui permettent de prendre en compte des données structurées selon des niveaux, typiquement dans le cas où des individus partagent un environnement commun qui peut affecter le comportement étudié. Ces modèles sont parfois désignés sous le terme de modèles hiérarchiques ou modèles mixtes et sont très utilisés dans le domaine des sciences de l'éducation.

Par exemple, les élèves d'une même classe ont leurs caractéristiques propres mais bénéficient aussi de conditions d'apprentissages communes. La prise en compte des différents niveaux permet d'expliquer une partie de l'hétérogénéité entre individus et de mettre en évidence l'effet du niveau sur l'estimation de la variable d'intérêt. Même si l'objectif d'une analyse ne porte pas sur l'effet du niveau dans une population, les modèles multiniveaux sont préconisés lorsque les données suivent une structure hiérarchique car ils donnent une meilleure estimation de l'impact des variables individuelles sur la variable d'intérêt.

Dans le cas d'InserJeunes la variable d'intérêt est binaire (être ou non en emploi à 6 mois) et le modèle économétrique utilisé est la régression logistique. Néanmoins, dans cette partie générale pour une lecture plus facile et pédagogique, nous détaillons les modèles avec une variable d'intérêt quantitative. En effet les hypothèses induites par l'utilisation des régressions par les moindres carrés ordinaires et les apports des modèles multiniveaux sont identiques que la variable d'intérêt soit binaire ou quantitative. La particularité lorsque l'on se situe dans le cas d'une variable binaire est qu'un faible nombre d'observation par groupe peut entrainer une mauvaise estimation de l'effet établissement et des coefficients du modèle. Nous ne sommes pas confrontés à ce problème dans Inserjeunes car la valeur ajoutée est diffusée uniquement s'il y a au moins 20 sortants dans l'établissement.

Différences avec la régression par les moindres carrés ordinaires

Pendant longtemps, il n'y a pas eu de méthode de régression spécifique pour prendre en compte l'effet de l'environnement sur les individus et la régression par les moindres carrés ordinaires (MCO) était le modèle le plus utilisé. Il est important pour apprécier l'apport des modèles multiniveaux de commencer par décrire la régression linéaire par MCO et ses limites.

La régression linéaire par les moindres carrés ordinaires

On se situe dans le cas le plus simple avec une seule variable explicative. Dans un modèle de régression linéaire par MCO on suppose que la variable d'intérêt quantitative Y est, pour chaque individu i, influencée par une variable explicative X

Le modèle peut être formalisé de la manière suivante :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Avec :

- l'indice i indique que l'équation s'applique à tous les individus i concernés.
- Y_i est la valeur observée de la variable Y pour l'individu i.
- β_0 est la constante du modèle (la valeur de la variable d'intérêt Y quand la valeur de la variable explicative X est nulle).
- β_1 est le coefficient directeur de la droite de régression (l'effet de la variable explicative X sur la variable d'intérêt Y).
- ε_i est le terme d'erreur du modèle pour l'individu i. Dans les faits on dispose des erreurs observées, aussi appelées résidus. L'erreur ε_i est donc l'écart entre la valeur estimée par le modèle $\widehat{Y}_i$ et la valeur observée Y_i.

La régression linéaire par MCO permet de trouver les valeurs des paramètres β_0 et β_1 qui minimisent la somme des carrés des écarts entre les valeurs observées de Y et les valeurs prédites $\widehat{Y}$ par le modèle.

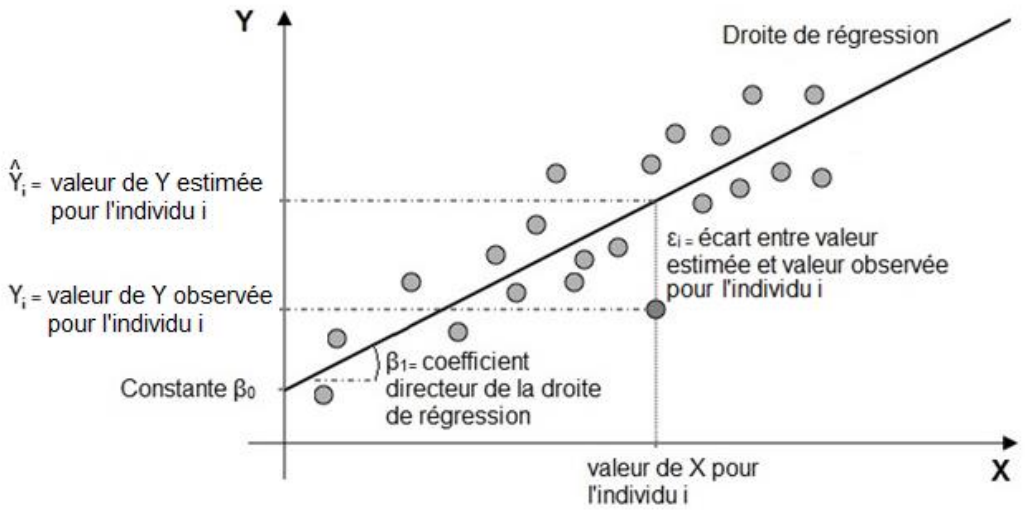
Le cas avec une variable est facilement généralisable au cas avec k variables X_1, X_2, ... X_k. Le modèle de régression linéaire est alors « multiple » et estime les valeurs des paramètres β_0, β_1, β_2... β_k.

Figure 1 : exemple illustratif

Pour chaque individu i on a une variable d'intérêt quantitative modélisée de façon suivante:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Sur le graphique suivant chaque individu i a une valeur $\widehat{Y}_i$ estimée par le modèle qui peut être différente de la valeur observée Y_i et l'erreur ε_i (le résidu) correspond à l'écart (la distance verticale) entre la réalité et sa prédiction par le modèle.



Par exemple on peut modéliser le niveau d'élèves en mathématiques (variable d'intérêt Y) en fonction de leur niveau en français (variable explicative X). Supposons que la majorité des élèves ont un niveau équivalent dans les deux matières, la relation sera positive et bien modélisée par la régression linéaire. Sur un graphique d'axes X et Y les points suivront à peu près une ligne droite et seront proches de la droite de régression (comme c'est le cas dans la figure 1). A l'inverse s'il n'y a pas de relation linéaire entre le niveau des élèves en mathématiques et en français, les points (X,Y) seront dispersés un peu partout sur le graphique et la distance (l'erreur) entre la plupart des points et la droite de régression sera importante.

Les hypothèses d'indépendance et d'homogénéité des erreurs

Deux des hypothèses de la régression linéaire par MCO sont invalidées si les individus partagent un environnement commun et que ce dernier a un impact sur la variable d'intérêt. Selon l'hypothèse d'indépendance des erreurs, les erreurs ε_i ne sont pas corrélées entre les individus i. Cela signifie que la connaissance de l'erreur liée à un individu i ne peut pas permettre de prédire l'erreur liée à un autre individu. Au contraire dans un modèle multiniveaux on suppose que l'environnement a un effet sur les individus de telle manière qu'une non-indépendance va s'établir entre eux du point de vue de la variable étudiée. Selon l'hypothèse d'homogénéité des erreurs (homoscédasticité), la variance des erreurs des individus est constante. Cette hypothèse est également invalidée lors de l'analyse des effets de l'environnement sur les individus.

Prenons par exemple le cas d'élèves dans une école élémentaire. On suppose que, quelles que soient les caractéristiques individuelles des élèves, chaque enseignant a un effet sur l'apprentissage des élèves de sa classe. Si un enseignant fait progresser ses élèves plus rapidement que d'autres, les acquisitions entre élèves de cette classe sont plus proches entre elles, et donc non indépendantes, par rapport aux acquisitions entre élèves de classes différentes.

Dans InserJeunes la valeur ajoutée vise à calculer l'apport propre de l'établissement sur le taux d'emploi de ses élèves sortants, compte tenu des variables individuelles et du taux de chômage de la zone d'emploi. Cet effet établissement ne peut pas être pris en compte par les modèles de régression par MCO, ce que permettent par contre les modèles multiniveaux.

Caractéristiques

Deux types de modèles multiniveaux existent : les modèles multiniveaux à effets fixes et les modèles multiniveaux à effets aléatoires. Ces deux types de modélisation ne reposent pas sur les mêmes hypothèses et ne permettent pas de répondre aux mêmes questions. Dans le tableau suivant on suppose que la variable d'intérêt quantitative Y est, pour chaque individu i dans un établissement j, influencée par une variable explicative individuelle X_{ij}. Il y a J établissements et l'erreur (le résidu) au niveau individuel s'écrit ε_{ij}.

	Modèle multiniveaux à effets fixes	**Modèle multiniveaux à effets aléatoires**
Hypothèse commune	Les erreurs de niveau individuel ne sont pas corrélées aux variables individuelles et sont indépendantes entre elles.	

Hypothèse spécifique	Pas d'hypothèse supplémentaire.	Hypothèse sur les effets aléatoires : il y a indépendance entre les erreurs de niveau établissement et les variables individuelles. Il n'existe pas de test statistique permettant de valider ou d'invalider cette hypothèse.
Modélisation	$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \sum_{k=1}^{J} \delta_k I_{k=j} + \varepsilon_{ij}$ Le niveau établissement est modélisé grâce aux J variables indicatrices I qui valent 1 lorsque k correspond à l'établissement j, 0 sinon. Les coefficients δ_j prennent en compte l'ensemble des caractéristiques de niveau établissement, il n'est par conséquent pas possible d'estimer l'effet propre d'une variable de niveau établissement sur la variable d'intérêt.	$Y_{ij} = \beta_0 + \beta X_{ij} + \gamma Z_j + \varepsilon_{ij} + \alpha_j$ Les variables de niveau établissement Z sont explicitement modélisés (coefficient γ) ainsi que les erreurs de niveau établissement α_j. Les erreurs α_j sont supposées suivre une loi normale dont les paramètres sont estimés.
Résidus du modèle	Il n'y a qu'un seul résidu (ε_{ij}).	Le modèle comporte deux résidus : un résidu élève (ε_{ij}) et un résidu établissement (α_j).
Conséquences	Les estimations sont supposées moins précises (plus de variance) mais sans biais. On ne peut pas déterminer quelle est la part de variance de chaque niveau dans la variance totale.	Les estimations sont supposées plus précises (moins de variance) mais si l'hypothèse sur les effets aléatoires n'est pas vérifiée les estimations des paramètres du modèle risquent d'être biaisées.
Cas principal d'utilisation	On souhaite contrôler l'effet du niveau établissement et des variables associées sans chercher à les distinguer, ou on pense que l'hypothèse sur les effets aléatoires est invalide.	On souhaite quantifier l'effet du niveau établissement et des variables associées ou on cherche une modélisation plus précise.

Nous avons décidé d'utiliser des modèles multiniveaux à effets aléatoires dans le cadre d'InserJeunes car seuls ces derniers permettent d'une part d'intégrer explicitement dans le modèle des variables au niveau établissement et d'autre part de séparer les résidus de niveau élèves et de niveau établissement.

<u>Détails des modèles multiniveaux à effets aléatoires</u>

On se place dans un cas avec deux niveaux (des d'individus dans des établissements), avec une variable d'intérêt Y quantitative et une seule variable explicative X. Comme l'effet du niveau établissement est explicitement modélisé il y a deux termes d'erreurs (résidus) dans le modèle : une erreur associée aux individus et une erreur associée aux établissements.

Cas 1 : modèle avec constante aléatoire

Dans un modèle avec constante aléatoire, la constante varie d'un groupe à l'autre et l'effet de la variable X sur Y est le même d'un groupe à un autre. La variabilité liée au groupe est ici uniquement sur la constante.

On peut écrire $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij} + \alpha_j$

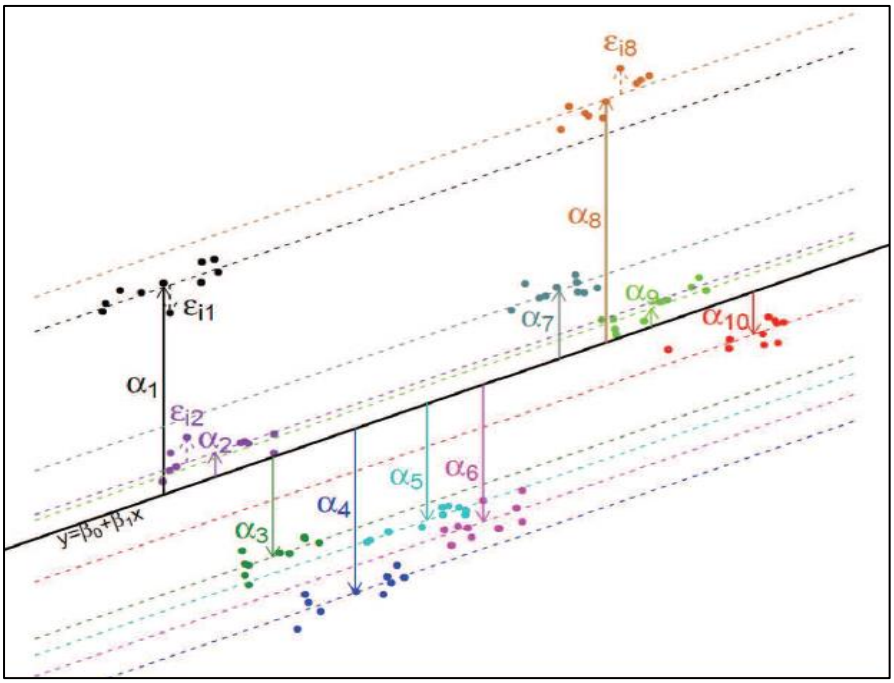
Avec :

- Y_{ij} est la valeur observée de la variable Y pour l'individu i appartenant à l'établissement j.
- β_0 est la valeur moyenne de Y sur l'ensemble des individus lorsque X est nul.
- β_1 est la pente de régression moyenne pour tous les établissements (l'effet de la variable X sur Y est supposé le même d'un établissement à un autre).
- ε_{ij} est l'erreur aléatoire du modèle (le résidu) pour l'individu i. Cette erreur est supposée suivre une loi normale de moyenne nulle.
- α_j est l'erreur aléatoire du modèle (le résidu) pour l'établissement j. Cette erreur correspond à l'effet de l'établissement j sur Y après prise en compte de la variable X. Cette erreur est supposée suivre une loi normale de moyenne nulle.
- $\beta_0 + \alpha_j$ correspond à la constante de l'établissement j.

Figure 2: modèle avec constante aléatoire

On représente sur le graphique suivant la droite de régression générale $Y = \beta_0 + \beta_1 X$ et les droites de régression de chaque établissement en pointillé $Y_j = \beta_0 + \beta_1 X_j + \alpha_j$

Les constantes varient entre les établissements : les droites de régression des établissements (en pointillé) ne sont pas à la même distance de la droite de régression générale (en trait plein). La distance verticale entre l'établissement j et la droite de régression générale est le terme α_j. La distance verticale entre l'individu i et la droite de régression de l'établissement j auquel il appartient est le terme ε_{ij}.



Ainsi, le groupe j=1 est éloigné de la droite de régression générale d'une distance verticale $\alpha_{1,}$ et l'individu i de ce groupe 1 est éloigné de la droite de régression par la distance verticale ε_{i1}

Cas 2 : modèle avec constante et pente aléatoires

En plus de la constante aléatoire, il est possible d'introduire de la variabilité dans les coefficients correspondant aux variables explicatives individuelles. Par exemple, la pratique d'un enseignant, nommée "effet maître", peut être plus bénéfique pour les bons élèves que pour ceux en difficulté, ou inversement. Ainsi dans un modèle avec constante et pente aléatoires, la constante et l'effet de la variable X sur Y peuvent varier d'un établissement à un autre.

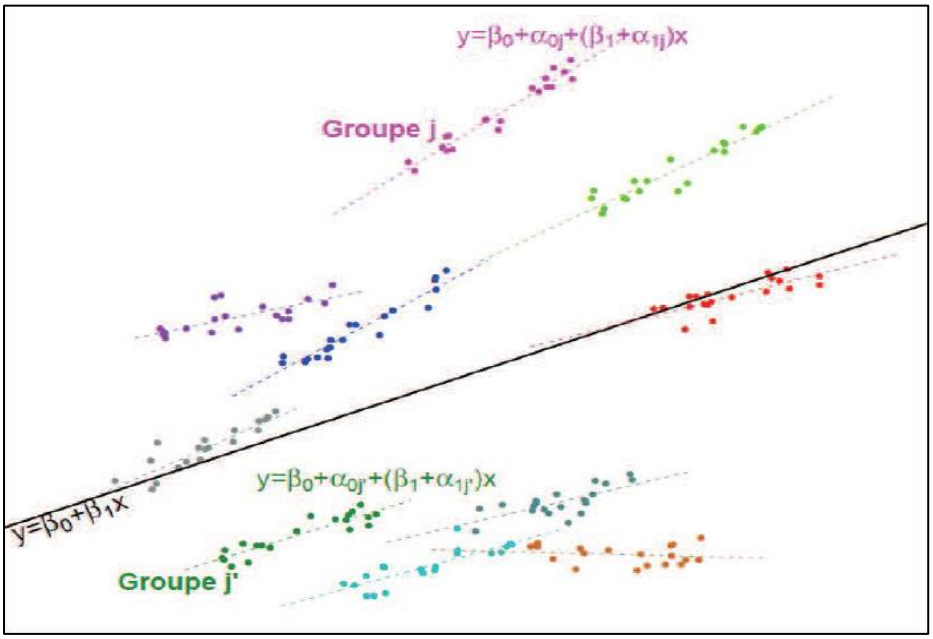
On peut écrire $Y_{ij} = \beta_0 + (\beta_1 + \alpha_{1j})\, X_{ij} + \varepsilon_{ij} + \alpha_{0j}$

Avec :

- Y_{ij} est la valeur observée de la variable Y pour l'individu i appartenant à l'établissement j.
- β_0 est la valeur moyenne de Y sur l'ensemble des individus lorsque X est nul.
- β_1 est la pente de régression moyenne pour tous les établissements.
- α_{1j} représente l'écart de l'établissement j à la relation moyenne (l'effet de la variable X sur Y est supposé être différent d'un établissement à un autre).
- ε_{ij} est l'erreur aléatoire du modèle (le résidu) pour l'individu i. Cette erreur est supposée suivre une loi normale de moyenne nulle.
- α_{0j} est l'erreur aléatoire du modèle (le résidu) pour l'établissement j. Cette erreur est supposée suivre une loi normale de moyenne nulle.

Figure 3: modèle avec constante et pente aléatoires

On représente sur le graphique suivant la droite de régression générale $Y = \beta_0 + \beta_1 X$ et les droites de régression de chaque établissement en pointillé $Y_j = \beta_0 + (\beta_1 + \alpha_{1j})\, X + \alpha_{0j}$



Les constantes varient entre les établissements.

Les pentes varient entre les établissements : les droites de régression des établissements (en pointillé) ont des pentes qui peuvent être différentes les unes des autres.

Emboîtement des niveaux

Dans de nombreux cas les données que l'on cherche à étudier suivent une structure strictement emboîtée ou hiérarchisée, par exemple les élèves dans une classe, les classes dans un collège, les patients dans un hôpital, les habitants dans un quartier… Le nombre de niveaux à prendre en compte dépend de la structure des données, de la pertinence de chacun des niveaux mais aussi de la puissance de calcul disponible. En effet, plus le modèle est complexe et plus sa résolution par un logiciel statistique peut prendre de temps. Lorsque la structure des niveaux n'est pas parfaitement emboîtée il est également possible d'utiliser des modèles multiniveaux, c'est le cas par exemple pour une analyse prenant en compte élèves, quartiers et écoles. Avec la carte scolaire, la majorité des élèves d'un quartier vont dans la même école mais une partie des élèves ira dans d'autres écoles.

Dans la figure 4 les élèves sont rattachés à une et une seule classe, et chaque classe est au sein de la même école. Les niveaux sont ici emboîtés car les unités observées sont strictement hiérarchisées les unes par rapport aux autres.

Dans la figure 5 tous les élèves d'un même quartier ne vont pas dans la même école. Les niveaux ne sont pas emboîtés car les unités observées ne sont pas strictement hiérarchisées les unes par rapport aux autres.

Figure 4 : modèles multiniveaux avec niveaux emboîtés

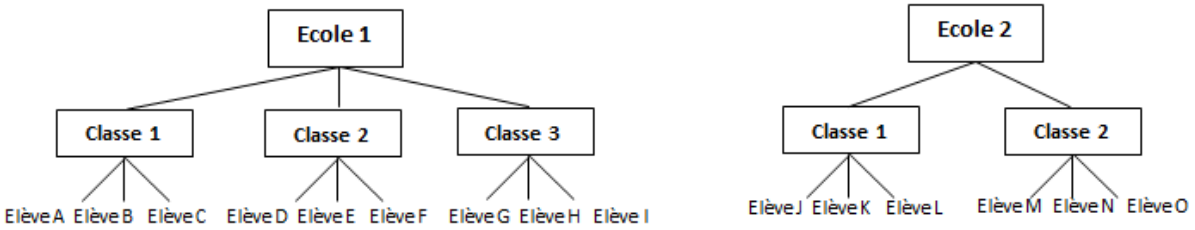
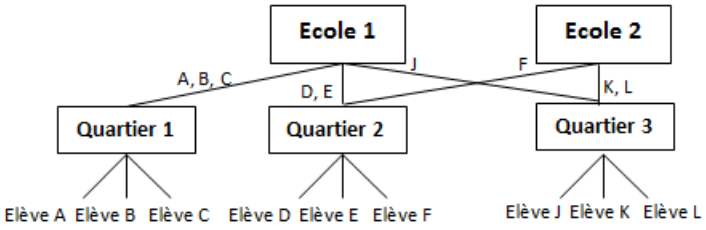


Figure 5 : modèles multiniveaux avec niveaux non emboîtés



3) Les modèles multiniveaux retenus dans le cadre d'InserJeunes

Le taux d'emploi attendu est la moyenne, au niveau de l'établissement, des probabilités estimées par modélisation multiniveaux que l'élève soit en emploi à 6 mois. Dans cette partie nous décrivons les caractéristiques des modèles multiniveaux utilisés dans InserJeunes ainsi que leur élaboration par ajout successif de variables.

Calcul du taux d'emploi attendu

Niveaux pris en compte

Dans le cas de l'insertion professionnelle mesurée par InserJeunes, les élèves sortants d'un même établissement ont été dans le même environnement scolaire pendant leur formation (cours et travaux pratiques, professeurs et équipe encadrante…) et cet environnement peut avoir un impact sur l'insertion professionnelle, qui est une caractéristique individuelle. De même, sous l'hypothèse que l'élève sortant cherche un emploi en priorité dans la zone d'emploi dans laquelle il réside, le marché du travail au niveau de la zone d'emploi et le taux de chômage associé ont un impact sur l'insertion professionnelle de l'élève sortant.

Ainsi trois niveaux sont pris en compte pour estimer la probabilité d'être en emploi salarié : le niveau individuel, le niveau établissement et le niveau zone d'emploi. Les élèves sont au sein d'établissements (lycée ou CFA) clairement identifiés. La zone d'emploi prise en compte est celle de la résidence de l'élève sortant lorsque la donnée est disponible, ce qui est le cas pour 98 % des sortants d'apprentissage et plus de 99 % des sortants de voie professionnelle scolaire. Dans le cas contraire nous prenons en compte la zone d'emploi de l'adresse de l'établissement.

Le choix de la zone d'emploi de résidence de l'élève implique que les modèles ne sont pas strictement emboîtés car les élèves d'un même établissement n'habitent pas forcément dans la même zone d'emploi. Une comparaison des résultats obtenus avec des modèles strictement emboîtés a été réalisée en prenant en compte la zone d'emploi de l'établissement. Cette comparaison n'a montré que de faibles différences entre les résultats des modèles et les valeurs ajoutées.

Spécification des modèles

Les modèles multiniveaux mis en œuvre dans le cadre d'InserJeunes ont les caractéristiques suivantes :

- Ce sont des modèles multiniveaux avec constante aléatoire : la constante varie d'un établissement à l'autre mais les effets des variables explicatives sur la variable expliquée sont les mêmes d'un établissement à un autre.
- Les modèles prennent en compte les niveaux élèves, établissement et zone d'emploi : il y a donc trois termes d'erreurs dans les modèles. Les niveaux établissement et zone d'emploi ne sont pas strictement emboîtés.
- Toutes les variables individuelles, de niveau établissement et de niveau zone d'emploi disponibles dans les bases de données élèves du champ d'InserJeunes ont été testées, mais seules les variables les plus significatives ont été gardées. Les variables finalement retenues sont le regroupement de spécialités de formation, le regroupement de codes NAF de l'établissement d'apprentissage (pour les apprentis), l'âge, le sexe, la profession et catégorie socio-professionnelle du responsable légal, la situation avant l'apprentissage (pour les apprentis), l'obtention du diplôme, les résultats à l'examen, la reconnaissance de handicap (au niveau individuel pour les apprentis et au niveau établissement pour la voie pro scolaire) et le taux de chômage de la zone d'emploi.

- La variable d'intérêt est binaire (être ou non en emploi à 6 mois) et le modèle économétrique utilisé est la régression logistique, aussi appelé modèle logit. Ce type de modèle permet en effet d'estimer pour chaque élève sortant la probabilité d'être en emploi. Ces modèles ne sont pas décrits ici mais font l'objet du document de méthodologie statistique de l'INSEE, « Le modèle Logit : théorie et applications », de Cédric Afsa (2016).

Pour modéliser le fait d'être en emploi, on utilise la variable Logit(P) où P est la probabilité d'être en emploi comprise entre 0 et 1. Logit(P) correspond à la formule suivante :

$$\mathrm{Logit}(\mathrm{P}) = \mathrm{Log}(\frac{\mathrm{P}}{1-\mathrm{P}})$$

Pour chaque individu i appartenant à un établissement j et dans une zone d'emploi k nous avons la régression suivante :

$$\mathrm{Logit}(\mathrm{P_{ijk}}) = \beta_0 + \beta_1 \mathrm{X_{ijk}} + \beta_2 \mathrm{W_j} + \beta_3 \mathrm{Z_k} + \varepsilon_{ijk} + \alpha_j + \gamma_k$$

Avec :

- β_0 est la valeur moyenne de $\mathrm{Logit}(\mathrm{P_{ijk}})$ sur l'ensemble des individus lorsque X est nul.
- $\mathrm{X_{ijk}}$ représente l'ensemble des variables individuelles du modèle (sexe, âge, diplôme, catégorie socio-professionnelle des parents, obtention du diplôme…)
- $\mathrm{W_j}$ correspond à la part d'élèves en situation de handicap au sein de l'établissement j. Cette variable est présente uniquement pour les modèles des lycées professionnels sous tutelle du ministère de l'éducation nationale.
- $\mathrm{Z_k}$ correspond au taux de chômage de la zone d'emploi
- ε_{ijk} est l'erreur aléatoire du modèle (le résidu) pour l'individu i. Cette erreur est supposée suivre une loi normale de moyenne nulle.
- α_j est l'erreur aléatoire du modèle (le résidu) pour l'établissement j. Cette erreur est supposée suivre une loi normale de moyenne nulle.
- γ_k est l'erreur aléatoire du modèle (le résidu) pour la zone d'emploi k. Cette erreur est supposée suivre une loi normale de moyenne nulle.

Comme le niveau établissement n'est pas strictement emboîté dans le niveau zone d'emploi, le résidu de niveau établissement s'écrit α_j et ne dépend pas de k.

Mode de calcul de la valeur ajoutée de l'établissement

Trois façons de calculer la valeur ajoutée de l'établissement sur le taux d'emploi sont envisageables:

1. La valeur ajoutée de l'établissement correspond au résidu de niveau établissement, soit α_j. Comme cette valeur ne correspond à aucune échelle et est difficile à expliquer cette définition n'a pas été retenue.

2. On prédit le taux d'emploi d'un établissement à partir des résultats des modèles multiniveaux en faisant la moyenne des probabilités d'être en emploi des sortants de cet établissement. On définit ensuite la valeur ajoutée de l'établissement telle que :

Valeur ajoutée de l'établissement	=	Taux d'emploi prédit avec résidu établissement α_j	-	Taux d'emploi prédit sans résidu établissement α_j

La valeur ajoutée correspond ici à une différence entre deux taux issus de modèles statistiques, et le taux d'emploi prédit avec résidu établissement peut différer du taux d'emploi observé. Cette définition n'est pas la plus simple à expliquer et n'a donc pas été retenue.

3. Nous avons retenu l'approche suivante : on considère que la valeur ajoutée d'un établissement correspond à la différence entre le taux d'emploi observé et le taux d'emploi prédit sans résidu établissement α_j.

Valeur ajoutée de l'établissement	=	Taux d'emploi observé	-	Taux d'emploi prédit sans résidu établissement α_j

Le taux d'emploi prédit sans résidu établissement α_j est calculé en faisant la moyenne des probabilités individuelles d'être en emploi après avoir soustrait le résidu établissement α_j. Ce taux est nommé « taux d'emploi attendu ». Ce dernier est calculé suivant le processus suivant en 3 étapes :

1. On modélise avec des modèles multiniveaux le fait d'être en emploi en fonction des caractéristiques de l'élève, de sa formation, de son établissement et du taux de chômage de la zone d'emploi de son lieu de résidence. On obtient pour chaque élève la probabilité estimée d'être en emploi.

2. Les modèles multiniveaux utilisés dans InserJeunes modélisent le logit de la probabilité P en prenant en compte le résidu de l'établissement α_j. On souhaite déterminer la probabilité individuelle d'être en emploi sans le résidu établissement α_j. On nomme cette probabilité Pa (c'est cette probabilité qui est prise en compte dans la définition du taux d'emploi attendu).

Dans un premier temps on part de la définition du Logit et on exprime Pa en fonction de $\mathrm{Logit(Pa)}$

$$\mathrm{Logit(Pa)} = \mathrm{Log}\left(\frac{\mathrm{Pa}}{1-\mathrm{Pa}}\right)$$

$$e^{\mathrm{Logit(Pa)}} = \frac{\mathrm{Pa}}{1-\mathrm{Pa}}$$

$$e^{\mathrm{Logit(Pa)}}(1-\mathrm{Pa}) = \mathrm{Pa}$$

$$e^{\mathrm{Logit(Pa)}} - \mathrm{Pa}\, e^{\mathrm{Logit(Pa)}} = \mathrm{Pa}$$

$$e^{\mathrm{Logit(Pa)}} = \mathrm{Pa} + \mathrm{Pa}\, e^{\mathrm{Logit(Pa)}}$$

$$\mathrm{Pa}\,(1 + e^{\mathrm{Logit(Pa)}}) = e^{\mathrm{Logit(Pa)}}$$

On obtient la formule (1) $\mathrm{Pa} = \frac{e^{\mathrm{Logit(Pa)}}}{1+ e^{\mathrm{Logit(Pa)}}}$

On peut ensuite écrire que le logit estimé par le modèle avec résidu établissement est égal à la somme du logit estimé sans résidu établissement et du résidu établissement.

On a donc : $\mathrm{Logit(P)} = \mathrm{Logit(Pa)} + \alpha j$

On remplace $\mathrm{Logit(Pa)}$ par $\mathrm{Logit(P)} - \alpha j$ dans la formule (1)

$$\mathrm{Pa} = \frac{e^{\mathrm{Logit(P)} - \alpha j}}{1 + e^{\mathrm{Logit(P)} - \alpha j}}$$

$$\mathrm{Pa} = \frac{e^{\mathrm{Log}(\frac{P}{1-P}) - \alpha j}}{1 + e^{\mathrm{Log}(\frac{P}{1-P}) - \alpha j}}$$

$$\mathrm{Pa} = \frac{P}{1-P}\ \frac{1}{e^{\alpha j}}\ \frac{1}{1 + \frac{P}{1-P}\ \frac{1}{e^{\alpha j}}}$$

$$\mathrm{Pa} = \frac{P}{1-P}\ \frac{1}{e^{\alpha j}(1 + \frac{P}{1-P}\ \frac{1}{e^{\alpha j}})}$$

$$\mathrm{Pa} = \frac{P}{1-P}\ \frac{1}{e^{\alpha j} + \frac{P}{1-P}}$$

On aboutit à la formule (2) $\mathrm{Pa} = \frac{P}{e^{\alpha j}(1-P) + P}$

A partir de la probabilité P et du résidu établissement α_j on obtient la probabilité Pa d'être en emploi sans résidu établissement. Cette probabilité prend en compte les caractéristiques individuelles de l'élève, la zone d'emploi et le taux de chômage associé.

3. Enfin on calcule le taux d'emploi attendu comme étant la moyenne des probabilités Pa des élèves de l'établissement.

Nombre total de modèles

Les niveaux de diplômes étant déterminants pour l'insertion professionnelle des sortants, des modèles par niveau de diplôme sont réalisés lorsque les effectifs sont suffisants.

De plus, nous exécutons également un modèle spécifique pour les apprentis en formation dans les lycées (EPLE). En effet ces apprentis sont comptés deux fois dans les indicateurs : une fois en tant qu'apprentis liés à leur CFA et une fois dans les lycées professionnels dans lesquels ils sont accueillis. Ainsi pour un lycée professionnel ayant des apprentis nous diffusons la valeur ajoutée de l'ensemble de l'établissement, ainsi que la valeur ajoutée des deux populations qui le composent : les lycéens et les apprentis.

Au total, 9 modèles ont été estimés pour calculer la probabilité d'être en emploi des élèves d'InserJeunes:

- 4 modèles pour les sortants d'apprentissage :
    - CAP, mentions complémentaires et autres formations de niveau V
    - BP
    - Bac pro, mentions complémentaires et autres formations de niveau IV
    - BTS et autres formations de niveau III
- 1 modèle pour les sortants d'apprentissage en EPLE
- 3 modèles pour les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'éducation nationale :
    - CAP, mentions complémentaires et autres formations de niveau V
    - Bac pro, mentions complémentaires et autres formations de niveau IV
    - BTS et autres formations de niveau III

- 1 modèle pour les sortants de voie professionnelle scolaire d'établissements sous tutelle du ministère de l'agriculture

Dans le modèle pour les sortants d'apprentissage en EPLE et le modèle pour les sortants de voie professionnelle scolaire sous tutelle du ministère de l'agriculture, la variable niveau de diplôme est introduite dans les variables explicatives et a un fort pouvoir explicatif.

Logiciel utilisé

La mise en œuvre des modèles multiniveaux a été réalisée en R (librairie LME4). Des comparaisons avec le logiciel SAS ont été réalisées et le logiciel R s'est avéré beaucoup plus performant que SAS. Des détails sur l'implémentation sur SAS et sur R sont disponibles en annexe 3.

Elaboration des modèles multiniveaux

Indicateurs de qualité des modèles

Pour chacun des modèles, on cherche à obtenir un modèle parcimonieux qui explique une part importante de la variance établissement. En effet, plus la variance de niveau établissement est faible, plus les variables intégrées dans le modèle expliquent « bien » le fait que l'élève soit en emploi ou non, et moins la valeur ajoutée sera dispersée entre les établissements.

Nous cherchons à obtenir des modèles parcimonieux et robustes pour plusieurs raisons :

- les modèles élaborés sont intégrés dans la chaîne de production d'InserJeunes et sont fixés pour plusieurs années, ils doivent donc avoir des performances stables dans le temps. Un modèle avec trop de variables pourrait expliquer une plus grande part de variance sur les données qui ont servies à son élaboration mais s'avérer de moins bonne qualité sur les données des millésimes suivants,
- le taux d'emploi attendu et la valeur ajoutée étant des indicateurs largement diffusés, la méthode utilisée doit pouvoir être expliquée facilement.

Elaboration des modèles par ajout successif de variables

L'élaboration de modèles de façon imbriquée permet de facilement comparer les modèles les uns avec les autres et de maîtriser l'impact et la significativité de chaque variable ajoutée. De plus, pour utiliser la déviance il faut que les modèles soient emboîtés : on compare la déviance entre deux modèles uniquement si un modèle est inclus dans l'autre.

Dans le cadre des modèles multiniveaux, nous commençons par un modèle vide (sans variables explicatives) avec constante aléatoire pour le niveau établissement. Commencer par un modèle sans variables explicatives permet de s'assurer de la présence de l'effet des niveaux : il s'agit alors d'une simple décomposition de la variance en variance inter-classes (variance entre établissements) et variance intra-classes (variance entre individus). Dans le cas des modèles logistiques, la variance d'une observation individuelle est fixée à $\pi^2/3$.

Nous ajoutons ensuite successivement le niveau zone d'emploi, les variables individuelles et en dernier les variables de niveau établissement et zone d'emploi. L'ajout de chaque variable est évalué avec les critères suivants :

Baisse de la déviance et sa significativité

La déviance correspond à la valeur -2 * Log (L) ou L est la vraisemblance du modèle. Plus la déviance diminue par rapport au modèle de référence, mieux le modèle décrit les données. Par définition, la

déviance diminue avec le nombre de paramètres ajoutés au modèle et ne pénalise pas les modèles trop complexes. Par contre, comme la diminution de la déviance d'un modèle à l'autre suit une loi du Chi2 il est possible de tester la significativité de l'ajout successif de chaque variable. Le nombre de degrés de liberté de la loi du Chi2 est déterminé ici par le nombre de paramètres supplémentaires à estimer dans le modèle le plus complexe.

Baisse de l'AIC

L'AIC (Le critère d'information d'Akaike) est un indicateur qui prend en compte la diminution de la déviance mais est pénalisé par deux fois le nombre k de paramètres ajoutés, soit :

$$AIC = - 2 * \log(L) + 2 * k$$

L'AIC représente donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible). En choisissant le modèle avec l'AIC le plus faible on s'assure que le modèle est robuste et ne contient que les variables ayant un pouvoir explicatif important.

Diminution de la variance de niveau établissement

Les modèles multiniveaux permettent d'estimer la variance des différents niveaux et leur part dans la variance totale du modèle. Par rapport au modèle sans aucune variable explicative, plus on ajoute de variables individuelles et plus la variance de niveau établissement diminue.

Dispersion de la valeur ajoutée

Nous regardons d'abord l'écart type de la valeur ajoutée, plus il est faible et moins la valeur ajoutée est dispersée. Puis en analysant la répartition de la valeur ajoutée en valeur absolue par classe, nous nous attendons à voir de moins en moins d'établissements avec des valeurs ajoutées extrêmes.

D de Somers

Le D de Somers est un indicateur basé sur les rangs. On considère toutes les paires d'observations ayant des valeurs observées de Y différentes, soient 1 et 0 et on les répartit dans 3 groupes :

- les paires concordantes : celles pour lesquelles l'observation où Y =1 a une probabilité estimée que Y =1 plus grande que l'observation où Y =0
- les paires discordantes : celles pour lesquelles l'observation où Y =1 a une probabilité estimée que Y =1 plus faible que l'observation où Y =0
- les paires « ex-aequo » : celles pour lesquelles l'observation où Y =1 a une probabilité estimée que Y =1 égale à celle de l'observation où Y =0

Le D de Somers est définit comme suit : $\frac{\text{nombre de paires concordantes} - \text{nombre de paires discordantes}}{\text{nombre de paires ayant des valeurs observées de Y différentes}}$

Le D de Somers varie dans [- 1, + 1], il est égal à 0 quand il n'y a pas d'association, il tend vers 1 lorsque l'association est très forte et positive et vers -1 lorsque l'association est très forte et négative.

Examen de la significativité des coefficients ajoutés

Nous conservons les variables quantitatives qui sont significatives dans au moins un des modèles par niveau de diplôme. De même nous conservons les variables qualitatives lorsqu'au moins une des modalités est significative dans un des modèles par niveau de diplôme.

Examen des Odds Ratio

Les Odds Ratio donnent une information sur l'ampleur de la relation entre une variable explicative et la variable d'intérêt. L'examen des Odds Ratio permet de s'assurer que les variables que l'on garde dans les modèles ont un impact important sur la variable d'intérêt.

Exemple : sortants d'apprentissage de l'été 2017, CAP et autres formations de niveau V

Dans InserJeunes, quatre modèles ont été élaborés pour les sortants d'apprentissage. On présente ici le cheminement de l'élaboration du modèle pour les CAP et autres formations de niveau V. Au niveau individuel les sortants de CAP sont comparés entre eux, et par conséquent la valeur ajoutée compare ici les CFA uniquement sur cette sous-population.

Modèle		**1**	**2**	**3**
Ajout de variables		modèle vide avec niveau établissement	Ajout niveau zone d'emploi	Ajout spécialité formation
Critères ajustement	deviance (-2 log vraisemblance)	61773	61767	61289
	AIC (préférer les petites valeurs)	61777	61773	61305
	diminution deviance	X	7	478
	diminution AIC	X	5	468
	significativité ajout paramètre	X	**	***
D de Somers		0,240	0,240	0,245
Valeur Ajoutée	écart type	14,8	14,1	10,9
	Valeur absolue (effectif >=10)			
	[0,5[	31%	34%	45%
	[5,10[	23%	23%	26%
	[10,15[	16%	17%	14%
	[15,25[	19%	17%	10%
	[25,max[	10%	9%	5%
Estimation variance *	Variance établissement	0,23	0,20	0,08
	variance zone emploi	X	0,02	0,03
	Total variance niveau 2 et 3	0,23	0,23	0,11

** Dans le cas des modèles logistiques, la variance d'une observation individuelle est fixée à $\pi^2/3$, ce qui équivaut à environ 3,28. L'ordre de grandeur de la variance individuelle est bien plus important que les variances de niveau établissement et zone d'emploi.*

Le modèle 1 est le modèle vide (pas de variables explicatives) avec constante aléatoire pour le niveau établissement.

L'ajout de la zone d'emploi dans le modèle 2 a les conséquences suivantes:

- baisse faible mais significative de la déviance
- baisse de l'AIC
- baisse de la variance de niveau établissement de 0.23 à 0.20
- baisse de la dispersion de la valeur ajoutée : 34% des CFA ont une valeur ajoutée comprise entre]-5,5[contre 31% pour le modèle 1.

L'ajout du regroupement de spécialités de formations dans le modèle 3 a les conséquences suivantes:

- baisse très importante et significative de la déviance
- baisse très importante de l'AIC
- baisse de la variance de niveau établissement de 0.20 à 0.08

- baisse de la dispersion de la valeur ajoutée : 45% des CFA ont une valeur ajoutée comprise entre]-5,5[contre 34% pour le modèle 2.
- Le D de Somers augmente de 0,240 à 0,245

Bien que faible par rapport à l'ajout des variables individuelles, l'ajout du niveau zone d'emploi est significatif. L'ajout du regroupement de spécialités de formation en 5 classes est significatif, a un très fort impact sur les indicateurs présentés dans cette partie et fait diminuer la variance de niveau établissement de plus de moitié.

Modèles finaux

Principaux résultats

La variance entre établissements diminue entre les modèles vides (pas de variables explicatives) et les modèles complets (elle est au minimum divisée par deux).

Le regroupement de spécialités de formation (voir définition ci-après) et le regroupement de NAF (nomenclature d'activités française) pour les CFA (voir définition ci-après) sont les variables qui ont le plus fort impact sur les modèles : diminution de la déviance, de l'AIC, de la variance établissement et de la dispersion de la valeur ajoutée.

L'ajout au niveau établissement de variables agrégeant des données individuelles (part des élèves de l'établissement selon l'âge, selon les formations) est significatif mais a été abandonné pour limiter la complexité du modèle.

L'ajout du taux de chômage de la zone d'emploi est tout le temps significatif.

Les résultats détaillés des modèles sont disponibles en annexe.

Les variables retenues

Regroupement de spécialités de formation

Les spécialités de formation sont trop nombreuses pour être intégrées telles qu'elles dans le modèle. Il a donc été décidé de les regrouper de manière automatique en fonction des taux d'insertion professionnels observés dans InserJeunes.

Pour ce faire, nous calculons dans un premier temps les taux d'emploi par nomenclature des spécialités en 100 postes ainsi que l'effectif de sortants associé.

Puis dans un second temps nous effectuons une Classification Ascendante Hiérarchique (CAH) avec le taux d'emploi par nomenclature des spécialités en 100 postes pondéré par l'effectif de sortant. La méthode utilisé dans la CAH est la méthode de Ward, elle cherche à minimiser l'inertie intra-classe et à maximiser l'inertie inter-classe afin d'obtenir des classes les plus homogènes possibles. Comme InserJeunes est un système d'information où tous les calculs sont automatisés, le nombre de groupes (cluster) choisis doit être fixe et suffisamment petit pour que chaque groupe ait toujours un nombre suffisant d'élèves. En effet, cette variable étant utilisée dans les modèles multiniveaux il est important d'avoir un effectif par modalité suffisamment important.

In fine, le regroupement de spécialités de formation en 100 postes comprend 4 modalités classées du taux d'emploi le plus faible au taux d'emploi le plus élevé.

Regroupement de codes NAF (nomenclature d'activité française)

De manière similaire au regroupement de spécialités de formation, les codes NAF de l'établissement d'apprentissage sont regroupés de manière automatique en fonction des taux d'insertion professionnels observés dans InserJeunes. Finalement, le regroupement de codes NAF comprend 3 modalités classées du taux d'emploi le plus faible au taux d'emploi le plus élevé.

Regroupement âge en classes

Les âges étant différents selon les différents modèles réalisés par niveau de formation, les regroupements d'âge sont spécifiques à chaque modèle, mais très proches entre l'apprentissage et la voie professionnelle scolaire.

Champ : apprentissage

diplôme	**bornes**	**libellé**
BTS et autres diplômes de niveau III	[min,20]	regroupement 1 (les plus jeunes)
	[21,21]	regroupement 2
	[22,23]	regroupement 3
	[24,max]	regroupement 4 (les plus âgés)
BP	[min,19]	regroupement 1 (les plus jeunes)
	[20,20]	regroupement 2
	[21,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)
Bac pro et autres diplômes de niveau IV	[min,19]	regroupement 1 (les plus jeunes)
	[20,20]	regroupement 2
	[21,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)
CAP et autres diplômes de niveau V	[min,18]	regroupement 1 (les plus jeunes)
	[19,19]	regroupement 2
	[20,21]	regroupement 3
	[22,max]	regroupement 4 (les plus âgés)

Champ : apprentis en lycée (EPLE)

diplôme	**bornes**	**libellé**
tous	[min,19]	regroupement 1 (les plus jeunes)
	[20,20]	regroupement 2
	[21,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)

Champ: élèves en voie professionnelle scolaire

diplôme	**bornes**	**libellé**
BTS et autres diplômes de niveau III	[min,20]	regroupement 1 (les plus jeunes)
	[21,21]	regroupement 2
	[22,22]	regroupement 3
	[23,max]	regroupement 4 (les plus âgés)
Bac pro et autres diplômes de niveau IV	[min,18]	regroupement 1 (les plus jeunes)
	[19,19]	regroupement 2
	[20,20]	regroupement 3
	[21,max]	regroupement 4 (les plus âgés)

CAP et autres diplômes de niveau V	[min,17]	regroupement 1 (les plus jeunes)
	[18,18]	regroupement 2
	[19,19]	regroupement 3
	[20,max]	regroupement 4 (les plus âgés)

Mention complémentaire

On ajoute une variable indicatrice telle que si le diplôme correspond à une mention complémentaire la valeur vaut 1 ; 0 sinon.

Sexe de l'élève

On ajoute une variable indicatrice telle que si le sexe correspond à une fille la valeur vaut 1 ; 0 sinon.

Profession et catégorie socio-professionnelle

Nous prenons en compte la PCS en une position du responsable de l'élève. Nous n'effectuons pas d'imputation des non réponses (moins de 2 % des élèves) mais créons une modalité « non réponse ».

libellé
agriculteurs
artisans, commerçants, chef d'entreprise
cadres, professions intellectuelles supérieures
professions intermédiaires
employés
ouvriers
retraités
sans activité professionnelle
non réponse

Situation avant l'apprentissage (pour les apprentis)

libellé
collège
second cycle général et technique, enseignement supérieur
second cycle professionnel
stage, emploi, contrat de professionnalisation
chômage
autres situations

Obtention du diplôme et résultats à l'examen

Obtention du diplôme	**moyenne**	**libellé**
Non réponse		Non réponse
non		diplôme non obtenu
oui	[min,12[	diplôme obtenu, note = [10,12[
	[12,14[	diplôme obtenu, note = [12,14[
	[14,max[	diplôme obtenu, note = [14,20[

Handicap

Les modèles multiniveaux prennent également en compte la spécificité de l'insertion professionnelle des jeunes sortants en situation de handicap sous la forme suivante :

- La reconnaissance de la qualité de travailleur handicapé (RQTH) pour les sortants d'apprentissage est introduite au niveau individuel.
- Pour les modèles sur les jeunes sortant pour la voie professionnelle scolaire en lycée, la part d'élèves en situation de handicap a été introduite. Cette variable est issue d'une enquête recensant annuellement les élèves ayant un Projet Personnalisé de Scolarisation au sein des établissements. Elle est disponible sur le champ des élèves en année terminale de formation pour les niveaux de diplômes suivants : CAP en deuxième année, terminale professionnelle, BTS en deuxième année.

L'ajout de ces variables dans les modèles multiniveaux a été testé et a un impact positif et important sur la valeur ajoutée des CFA qui accueillent majoritairement des apprentis avec RQTH ainsi que des lycées professionnels avec une part importante d'élèves en situation de handicap.

Taux de chômage

Nous avons retenu pour les modèles le taux de chômage annuel au niveau de la zone d'emploi car les taux de chômage trimestriels ne sont pas disponibles pour les départements d'outre-mer (les DOM, excepté Mayotte, font partie du champ d'InserJeunes). La comparaison, au niveau France métropolitaine, de l'ajout du taux de chômage annuel par rapport au trimestriel a montré un impact très modéré.

Références bibliographiques

Afsa Cédric, « Le modèle Logit, théorie et applications », Méthodologie Statistique, 2016

Bressoux Pascal, « L'apport des modèles multiniveaux à la recherche en éducation », Éducation et didactique, 2007

Bressoux Pascal, « Modélisation statistiques appliquée aux sciences sociales », De Boeck, 2008

Bringé Arnaud, Golaz Valérie, « Manuel pratique d'analyse multiniveau », Ined, 2007

Evain Franck, Evrard Laetitia, « Une meilleure mesure de la performance des lycées », refonte de la méthodologie des IVAL (session 2015), Education & Formations, N° 94, 2017

Evain Franck, « Indicateurs de valeur ajoutée des lycées, du pilotage interne à la diffusion grand public », Courrier des statistiques N° 5, 2020

Givord Pauline, Guillerm Marine, « Les modèles multiniveaux », Méthodologie Statistique, 2016

Annexe 1 : détails de l'élaboration des modèles imbriqués

Les résultats présentés ici sont issus des sortants d'apprentissage 2017 pour les CAP et les Bac pro. Il y a 3 modèles retenus :

- le modèle A comprend toutes les variables (individuelles, niveau établissement zone d'emploi) qui font diminuer la déviance de façon significative (même faiblement).
- le modèle B est le modèle A auquel on ajoute une pente aléatoire pour le regroupement des spécialités de formation (variable qui a le plus fort impact sur l'insertion professionnelle)
- le modèle C est le modèle A auquel on enlève le niveau zone d'emploi et le taux de chômage associé.

Les deux premiers sont des modèles à 3 niveaux imbriqués, le dernier en contient 2.

				Ajout de variables individuelles						Ajout variables niveau CFA			modèles finaux		
													A	B *B comparé à A*	C *C comparé à A*
Niveaux		2 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	3 niveaux	2 niveaux
Ajout de variables (modèles emboîtés: chaque modèle est comparé au précédent)		modèle vide	modèle vide	+ specialité form-ation	+ age	+ mention complém-entaire (ref=non)	+ sexe (ref= homme)	+ pcs resp. (ref=cadre)	+ situation avant apprentissage (ref=collège)	+ part des spécialités de formation	+ Nombre de sites de formation	+ part d'apprentis selon classes d'age	+ taux de chômage zone emploi	+ pente aléatoire specialité formation	modele A sans niveau zone emploi et tx chomage
Critères ajustement	deviance (-2 log vraisemblance)	61773	61767	61289	60973	60751	60606	60444	60397	60370	60357	60346	60310	60277	60376
	AIC (préférer les petites valeurs)	61777	61773	61305	60991	60771	60628	60482	60445	60428	60417	60412	60378	60355	60440
	diminution deviance	X	7	478	317	221	146	162	47	27	13	10	37	32	-66
	diminution AIC	X	5	468	315	219	144	146	37	17	11	4	35	22	-62
	significativité ajout paramètre	X	**	***	***	***	***	***	***	***	***	*	***	***	X
Prédiction	Somers' D	0,240	0,240	0,245	0,274	0,285	0,288	0,294	0,296	0,296	0,296	0,296	0,294	0,304	0,298
	augmentation Somers'D	X	0,000	0,005	0,028	0,012	0,003	0,006	0,002	0,000	0,000	0,000	-0,002	0,010	0,003
Valeur Ajoutée	ecart type effectif CFA >=10	14,8	14,1	10,9									10,2	10 ,0	11,2
	écart absolu VA														
	[0,5[	31%	34%	45%									47%	48%	38%
	[5,10[	23%	23%	26%									27%	27%	28%
	[10,15[	16%	17%	14%									13%	12%	19%
	[15,25[	19%	17%	10%									9%	10%	11%
	[25,max[	10%	9%	5%									3%	3%	4%
Estimation variance	variance établissement	0,23	0,20	0,08	0,10	0,10	0,08	0,07	0,07	0,07	0,07	0,06	0,06	0,06	0,10
	variance zone emploi	X	0,02	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,02	0,02	X
	total	0,23	0,23	0,11	0,13	0,13	0,12	0,11	0,10	0,10	0,10	0,10	0,08	0,08	0,10

Analyse :

La variance entre établissements diminue entre les modèles vides et les modèles complets (elle est au minimum divisée par deux).

Le regroupement de spécialités (regroupement ad-hoc selon le taux d'emploi par NSF) et le regroupement de NAF (même méthode) sont les variables qui ont le plus fort impact sur les modèles (diminution de la déviance, de l'AIC, de la variance établissement et de la dispersion de la valeur ajoutée).

L'ajout au niveau établissement des variables individuelles (part des élèves de l'établissement selon l'âge, selon les formations) est significatif mais leur contribution est assez faible.

L'ajout du taux de chômage de la zone d'emploi est significatif

L'ajout de la pente aléatoire pour la spécialité de formation est significatif mais complexifie le modèle.

La suppression du niveau zone d'emploi et du taux de chômage associé entraine une perte d'information significative par rapport au modèle complet et fait augmenter la dispersion des valeurs ajoutées.

Avec le modèle final A, 74 % des établissements ont un écart de valeur ajoutée inférieur à 10 points (contre 54 % pour le modèle vide à 2 niveaux).

C'est le modèle A qui sera choisi sans les variables de niveau établissement.

Annexe 2 : détails des modèles finaux pour les sortants 2018

Indicateurs d'ajustement des modèles

		Apprentissage				Apprentis en EPLE	Voie professionnelle scolaire		
Niveau de diplôme		**niveau V**	**niveau IV BP**	**niveau IV autres**	**niveau III**	**tous niveaux**	**niveau V**	**niveau IV**	**niveau III**
Taux d'insertion 6 mois après la sortie		54%	73%	64%	69%	65%	26%	37%	55%
Nombre d'élèves		41 248	14 256	18 584	28 368	12 591	26 213	75 374	56 444
Nombre d'établissement		619	415	601	629	721	1 637	2 046	1 734
Nombre d'élèves (étab avec effectif >=20)		40 986	14 127	18 360	27 683	12 412	24 978	73 296	52 698
Nombre d'établissements (étab avec effectif >=20)		581	396	567	551	683	1 487	1 830	1 416
Critères ajustement	deviance (-2 log vraissemblance)	53 285	15 440	22 493	33 057	14 869	26 819	95 103	76 143
	AIC	53 349	15 502	22 557	33 119	14 939	26 869	95 153	76 191
Estimation variance	variance établissement	0,05	0,04	0,05	0,04	0,06	0,11	0,06	0,07
	variance zone emploi	0,02	0,04	0,02	0,04	0,01	0,04	0,03	0,02
	total	0,07	0,08	0,07	0,08	0,07	0,14	0,09	0,09
filtre 20 sortants minimum par établissement									
Valeur Ajoutée (étab avec effectif >=20)	écart type VA	13,3	14,6	14,9	12,7	20,0	16,0	11,1	14,5
	Distribution écart absolu VA								
	[0,5[	39%	38%	33%	41%	25%	29%	39%	33%
	[5,10[	30%	27%	26%	27%	22%	25%	29%	27%
	[10,15[	13%	14%	18%	13%	17%	18%	17%	18%
	[15,25[	12%	13%	15%	13%	17%	18%	11%	15%
	[25,max[	6%	8%	8%	6%	20%	10%	4%	8%

Odds ratio des modèles

Niveau de diplôme		Apprentissage				Apprentis en EPLE	Voie professionnelle scolaire		
		niveau V	niveau IV BP	niveau IV autres	niveau III	tous niveaux	niveau V	niveau IV	niveau III
niveau de diplôme (ref= CAP)	niv 4 BP					2,89 ***			
	niv 4 bac pro					1,86 ***			
	niv 3 bts					1,53 ***			
regroupement specialités de formation (ref=2)	regroupement 1 (tx emploi faible)	0,91 .	0,98	0,74 **	0,76 ***	0,95	0,76 ***	0,80 ***	0,59 ***
	regroupement 3	1,18 ***	1,04	1,13 **	1,14 ***	1,15 *	1,13 **	1,19 ***	1,33 ***
	regroupement 4 (tx emploi fort)	1,70 ***	1,04	1,53 ***	1,45 ***	1,29 ***	1,60 ***	1,75 ***	1,83 ***
regroupement NAF (ref=2)	regroupement 1 (tx emploi faible)	0,45 ***	0,30 ***	0,36 ***	0,51 ***	0,38 ***			
	regroupement 3 (tx emploi fort)	1,50 ***	1,75 ***	1,74 ***	1,59 ***	1,86 ***			
regroupement age croissant (ref= 1)	regroupement 2	1,48 ***	0,90 .	1,18 **	0,96	1,00	1,72 ***	1,18 ***	0,98
	regroupement 3	1,48 ***	0,85 **	1,09 .	0,87 ***	0,96	2,06 ***	1,12 **	0,92 *
	regroupement 4	1,24 ***	0,76 ***	1,09	0,84 ***	0,79 **	2,01 ***	1,11 .	0,80 ***
mention complémentaire (ref=non)		1,70 ***		1,38 ***		1,20 .	1,92 ***	1,37 ***	
sexe (ref=homme)		0,77 ***	1,04	0,94 .	1,14 ***	0,96	0,65 ***	0,96 *	1,12 ***
PCS responsable (ref=cadres prof. intellectuelles sup.)	agriculteurs	1,14	1,33 .	1,05	1,24 *	1,39	1,43	1,35 *	1,28 *
	artisans, commerçants, chef Entreprises	1,22 **	1,33 **	1,14	0,97	0,86	1,40 **	1,23 ***	1,26 ***
	Professions intermédiaires	1,13 *	1,23 *	1,12 .	1,02	1,22 *	1,10	1,07	1,24 ***
	employés	1,14 *	1,21 *	1,10	0,96	1,05	1,10	1,10 *	1,24 ***
	ouvriers	1,15 **	1,37 ***	1,18 *	1,08	1,21 *	1,02	1,12 **	1,29 ***
	retraites	0,86	0,81	0,86	1,04	1,16	0,84	0,91	0,96
	sans activité professionnelle	0,97	1,10	0,90	0,88 *	0,99	0,88	0,89 **	1,05
	non réponse	1,25 ***	0,91	0,85 *	0,79 ***	0,88	1,37 ***	1,00	1,05
situation avant apprentissage (ref=collège)	2nd cycle gt, enseignement sup	1,00	0,86 **	0,80 ***	0,84 *	0,98			
	2nd cycle professionnel	1,07 *	0,87 *	0,81 ***	0,90	1,00			
	stage, emploi, ctt pro.	1,04	0,88	0,92	0,86 .	1,18			
	chomage	1,04	0,65 ***	0,67 ***	0,72 **	0,86			
	autres situations	1,10 *	0,84 *	0,81 ***	0,79 **	1,06			
obtention du diplôme et résultats à l'examen (ref=diplôme non obtenu)	non réponse	1,28 ***	1,07	1,75 ***	1,03	1,28 **	1,21 **	1,06	1,04
	diplôme obtenu, note = [10,12[	1,64 ***	1,84 ***	1,62 ***	1,03	1,68 ***	1,41 ***	1,41 ***	1,03
	diplôme obtenu, note = [12,14[	2,06 ***	2,10 ***	2,05 ***	1,13	2,26 ***	2,12 ***	1,65 ***	1,10
	diplôme obtenu, note = [14,20[	2,65 ***	2,15 ***	2,36 ***	0,96	2,56 ***	3,07 ***	2,06 ***	0,94
bénéficie reconnaissance travailleur handicapé (ref=non)		0,77 **	0,77	0,68 *	0,64 **	0,80			
part d'élèves en situation de handicap							0,99 ***	0,99 *	1,00
taux de chômage de la zone d'emploi		0,93 ***	0,94 ***	0,95 ***	0,95 ***	0,96 ***	0,90 ***	0,91 ***	0,94 ***

significativité: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Annexe 3 : mise en place de modèles multiniveaux avec SAS et R

Dans le cadre d'InserJeunes deux solutions logicielles ont été testées et comparées pour réaliser la modélisation multiniveaux : SAS (procédures glimmix et nlmixed) et R (package lme4).

Sur les nombreux tests effectués les procédures SAS et R donnent des résultats identiques, excepté lorsque l'on ajoute de nombreux paramètres de pente aléatoire : on constate alors des différences assez minimes sur l'estimation des paramètres et des résidus.

En termes de temps de traitement, R s'est avéré plus rapide que les procédures SAS dès que le modèle devient complexe ou que le nombre de modalités devient important. Par exemple sur une table avec 46 000 individus voici les temps de traitement avec SAS et R installé sur poste (Intel Core i5, 8Go de RAM):

- Pour un modèle à 2 niveaux avec 8 variables indicatrices et effet aléatoire sur la constante, SAS met 39 secondes et R met 14 secondes.
- Pour un modèle vide à 3 niveaux et effet aléatoire sur la constante, SAS met 3 minutes et R met 7 secondes.
- Pour un modèle à 3 niveaux avec 8 variables indicatrices et effet aléatoire sur la constante, SAS met 53 minutes et R met 15 secondes.
- Pour un modèle à 3 niveaux avec 19 variables indicatrices et un effet aléatoire sur la constante, SAS met plus de 10h et ne converge que partiellement (tous les paramètres ne sont pas estimés) tandis que R met trois minutes et n'a pas de problème de convergence.

Pour R, les temps de traitements mentionnés ont été optimisés en ajoutant l'option suivante dans la fonction glmer: *control = glmerControl(optimizer = "nloptwrap", calc.derivs = FALSE*.

1) Instructions sur SAS et sur R

Exemple : modèle à 3 niveaux emboîtés (individus, établissements, zones d'emploi). La variable numero_uai identifie l'établissement et la variable zone_emploi identifie la zone d'emploi.

- Variable d'intérêt : en_emploi (insertion professionnelle obtenu après appariement des élèves et des salariés de la DSN)
- Variable explicative : fille (indicatrice du sexe de l'élève)
- Constante aléatoire

Rappel instructions SAS

```
proc nlmixed data = table ;
/*modele avec effet aléatoire établissement*/
y=b0 + bsexe*fille + u01 + u02;
p=1/(1+exp(-y));
model en_emploi ~ binary(p);
/*résidus établissement u01 suivent une loi normale de moyenne 0 et de
variance établissement sigma_uai*/
random u01 ~ normal(0,sigma_uai) subject=numero_uai(zone_emploi);
/*les résidus zone d'emploi u02 suivent une loi normale de moyenne 0 et de
variance zone d'emploi sigma_ze*/
random u02 ~ normal(0,sigma_ze) subject=zone_emploi;
run;
```

Instructions R

Plusieurs formulations sont possibles pour la procédure glmer pour indiquer comment les niveaux sont imbriqués:

Formula	Alternative	Meaning
`(1 | g)`	`1 + (1 | g)`	Random intercept with fixed mean.
`0 + offset(o) + (1 | g)`	`-1 + offset(o) + (1 | g)`	Random intercept with *a priori* means.
`(1 | g1/g2)`	`(1 | g1)+(1 | g1:g2)`	Intercept varying among `g1` and `g2` within `g1`.
`(1 | g1) + (1 | g2)`	`1 + (1 | g1) + (1 | g2).`	Intercept varying among `g1` and `g2`.
`x + (x | g)`	`1 + x + (1 + x | g)`	Correlated random intercept and slope.
`x + (x || g)`	`1 + x + (1 | g) + (0 + x | g)`	Uncorrelated random intercept and slope.

Source: https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf

Comme pour la procédure SAS on indique ici que l'établissement (variable numero_uai) est emboité dans la zone d'emploi (variable zone_emploi).

Les deux formulations suivantes sont strictement équivalentes à l'instruction SAS présentée dans le paragraphe précédent:

```
modele <- glmer(en_emploi ~ 1
                + fille
                +(1| zone_emploi/numero_uai),
                data=df,
                family=binomial(link=logit))
summary(modele)
```

```
modele <- glmer(en_emploi ~1
                +fille
                +(1|zone_emploi)
                +(1|zone_emploi:numero_uai),
                data=df,
                family=binomial(link=logit))
summary(modele)
```

Pour accélerer les temps de traitements il est possible d'ajouter l'instruction :
control = glmerControl(optimizer = "nloptwrap", calc.derivs = FALSE)

2) Instructions dans le cadre du calcul de la valeur ajoutée

Pour cet exemple nous modélisons le fait d'être en emploi (variable « en_emploi ») selon le sexe dans un modèle à 3 niveaux emboîtés (élèves, établissements, zones d'emploi de la commune de l'établissement).

Les différentes étapes des programmes pour calculer la valeur ajoutée sont :

- Modèle multiniveaux avec export des probabilités et des résidus établissements
- Calcul taux brut de l'établissement
- Calcul taux attendu sans résidu établissement = moyenne des probabilités individuelles sans résidu établissement
- valeur ajoutée de l'établissement = taux brut – taux attendu

Instructions SAS

pb = probabilité sans résidu établissement

p = probabilité avec résidu établissement

```
proc nlmixed data = table ;
/*initialisation des parametres*/
parms b0=0.5 sigma_uai=0.15 bsexe=-0.3 sigma_ze=0.05;

/*modele sans effet aléatoire établissement*/
logitpb = b0 + bsexe*fille;
pb = exp(logitpb) / (1 + exp(logitpb));

/*modele avec effet aléatoire établissement*/
logitp = logitpb + u01;
p = exp(logitp) / (1 + exp(logitp));

model en_emploi ~ binary(p);

/*les résidus etablissement u01 suivent une loi normale de moyenne 0 et de variance sigma_uai*/
/*export table "residus_etab"*/
random u01 ~ normal(0,sigma_uai) subject=numero_uai(zone_emploi) out = residus_etab;

/*les résidus zone d'emploi u02 suivent une loi normale de moyenne 0 et de variance sigma_ze*/
random u02 ~ normal(0,sigma_ze) subject=zone_emploi;

/*export table "proba" avec estimation:
        - des résidus etablissements u01
        - des proba pb (sans effet etablissement)
        - des proba p (avec effet etablissement)*/
id u01 pb;

predict p out = proba;
run;
```

Le modèle sort directement la probabilité « pb » sans résidu établissement, ce qui permet de calculer le taux attendu par établissement (sans résidu établissement) puis la valeur ajoutée de l'établissement.

Instructions R

Contrairement au calcul de la valeur ajoutée sur SAS la procédure R ne donne pas directement la probabilité « pb » sans résidu établissement. On doit donc calculer la proba « pb » à partir de la proba « p » estimée par le modèle.

On part de : $p = \frac{e^{logit\ pb+u}}{1+e^{logit\ pb+u}}$ pour obtenir $pb = \frac{p}{(1-p)*e^{u}+p}$

```
library(lme4)
library(tidyverse)

modele <- glmer(en_emploi~1
                +fille
                +(1| zone_emploi)
                +(1| zone_emploi:numero_uai),
                data=df,
                family=binomial(link=logit),
                control=glmerControl(optimizer="nloptwrap",
calc.derivs=FALSE))


#extraction probabilités prédites du modèle avec résidu établissement
proba<-fitted(modele)

#ajout colonne proba a la table individu
df_indiv<-cbind(df, proba)

#extraction residus niveau 2
residus<-ranef(modele)

#recodage des residus niveau établissement
residus_uai <- residus[["numero_uai:zone_emploi"]]
residus_uai <- residus_uai %>%
  rownames_to_column(var="numero_uai") %>%
  rename(residus_uai = "(Intercept)") %>%
  mutate(numero_uai=substr(numero_uai,0,8))

#ajout colonne résidus à la table individu
df_indiv <-
left_join(
      mutate(df_indiv, numero_uai=as.vector(numero_uai)),
      residus_uai,
      by="numero_uai")

#table individu: taux brut et taux attendu (modele pb sans résidu
établissement)
df_indiv<-df_indiv %>%
  mutate(proba_sans_u=proba/((1-proba)*exp(residus_uai)+proba))

#vérification
head(select(df_indiv,proba,proba_sans_u))
```

```
#table par établissement: calcul de la valeur ajoutée
df_uai<-df_indiv %>%
  group_by(numero_uai) %>%
  summarise(effectif=n(),
            insertion=sum(en_emploi),
            tx_brut=insertion/effectif,
            tx_attendu=mean(proba_sans_u),
            va=round((tx_brut-tx_attendu)*100,digits=2)) %>%
  inner_join(residus_uai,by="numero_uai")

#dispersion de la valeur ajoutée
summary(df_uai$va)
ggplot(df_uai, aes(va, stat(density))) + geom_histogram(binwidth = 5)
ggplot(df_uai, aes(y=va)) + geom_boxplot() + expand_limits(y=c(-50, +50)) +
theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```