

## Article

---

« L'analyse des données biographiques au moyen des modèles linéaires à effets aléatoires. Le cas des carrières des acteurs professionnels »

Benoît Laplante et Benoît-Paul Hébert

*Cahiers québécois de démographie*, vol. 30, n° 1, 2001, p. 115-145.

Pour citer cet article, utiliser l'adresse suivante :

<http://id.erudit.org/iderudit/010301ar>

Note : les règles d'écriture des références bibliographiques peuvent varier selon les différents domaines du savoir.

---

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter à l'URI <http://www.erudit.org/apropos/utilisation.html>

---

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche. Érudit offre des services d'édition numérique de documents scientifiques depuis 1998.

Pour communiquer avec les responsables d'Érudit : [erudit@umontreal.ca](mailto:erudit@umontreal.ca)

## **L'analyse des données biographiques au moyen des modèles linéaires à effets aléatoires. Le cas des carrières des acteurs professionnels**

Benoît LAPLANTE et Benoît-Paul HÉBERT \*

*Les démographes font aujourd'hui un usage abondant des microdonnées longitudinales. Pour les analyser, ils utilisent généralement un ensemble de techniques et de modèles regroupés sous le nom d'analyse des biographies. Les modèles linéaires utilisés dans ce contexte appartiennent à la famille des modèles à risques proportionnels dont le plus courant est le modèle semi-paramétrique à risques proportionnels. Malgré leurs qualités, ces modèles ne permettent que l'étude des processus qui peuvent être décrits et mesurés facilement comme des suites d'états ou de changements d'état. Les auteurs montrent comment on peut utiliser deux modèles linéaires à effets aléatoires — le modèle logit à effets aléatoires et le modèle tobit à effets aléatoires — pour étudier des phénomènes longitudinaux comme l'évolution, au fil du temps, de l'appartenance à l'une ou l'autre des deux catégories d'une variable dépendante binaire, et l'évolution, au fil du temps, d'une variable continue dont une partie de la distribution n'est pas observée. L'approche est appliquée à l'étude de la dynamique des carrières des acteurs professionnels français. English abstract at the end of the article.*

**D**epuis le milieu des années 1980, les démographes se sont mis à faire un usage abondant des microdonnées longitudinales dans leurs travaux empiriques. Pour analyser ces données, ils utilisent généralement un ensemble de techniques et de modèles regroupés sous le nom d'analyse des biographies. Les modèles linéaires habituellement utilisés dans ce contexte appartiennent à la famille des modèles à risques proportionnels dont le plus courant est le modèle semi-paramétrique à risques proportionnels (Cox, 1972; Courgeau et Lelièvre, 1989; Blossfeld et al., 1989). L'intérêt que les démographes portent à ces modèles se comprend facilement : leur variable dépendante est le risque instantané de changer d'état, ce qui signifie que

---

\* Benoît Laplante : Centre interuniversitaire d'études démographiques et Institut national de la recherche scientifique; Benoît-Paul Hébert : Institut national de la recherche scientifique.

l'usage des modèles à risques proportionnels leur permet d'étudier, dans le cadre d'un modèle qui s'apparente à la régression, le risque qu'ils calculent dans les tables d'extinction, l'un des outils les plus fondamentaux de la démographie. Au cours des quinze dernières années, l'analyse des biographies, pratiquée avec ces modèles, a été utilisée pour étudier la plus grande partie des phénomènes démographiques : la fécondité, la nuptialité et la mortalité, bien sûr, mais aussi les comportements liés à la santé reproductive, les différents aspects de la migration, les trajectoires conjugales, etc. L'usage des modèles à risques proportionnels a, de toute évidence, rendu possibles des contributions à l'avancement des connaissances dont on voit mal comment elles auraient pu être réalisées autrement.

Malgré toutes leurs qualités, les modèles à risques proportionnels ont des limites qui tiennent à leur nature même : ils ne permettent que l'étude des processus qui peuvent être décrits et mesurés facilement comme des suites d'états ou de changements d'état. Il est pourtant évident que les phénomènes auxquels s'intéressent les démographes ne sont pas tous des changements d'état.

Nous nous sommes intéressés aux modèles à effets aléatoires pour données longitudinales dans le cadre d'un programme de recherche sur l'insertion professionnelle des comédiens français. Selon le droit du travail français, les comédiens sont des travailleurs salariés intermittents à employeurs multiples. Par définition, ils n'ont que très rarement un lien d'emploi unique et stable. Au contraire, leur carrière se construit dans un contexte où la fragmentation de l'emploi est la règle : succession d'engagements de durées variées auprès d'employeurs différents, engagements qui ne sont pas des emplois à temps plein, cumul fréquent d'emplois au même moment. Comme ces emplois sont généralement de courte durée, qu'ils sont rarement à temps plein, que le cumul est fréquent et que l'on sait par ailleurs que les notions de dates de début et de fin des engagements sont à tout le moins imprécises dans un tel contexte, il est à toutes fins utiles irréaliste de chercher à concevoir la carrière d'un comédien comme une suite de passages d'un état d'emploi à un état de non-emploi réalisés à des instants précis, ou encore d'étudier l'évolution de son revenu comme une suite relativement ordonnée de passages d'un niveau de revenu à un autre niveau de revenu à des instants précis correspondant aux moments des changements d'emploi. En revanche, on peut sans trop de difficulté concevoir

et mesurer des sommes, des moyennes ou d'autres variables analogues qui décrivent l'état moyen de la carrière de ce comédien au cours d'une période donnée. On construit donc ainsi des variables longitudinales, par exemple le fait d'avoir travaillé ou non au cours d'une année mesuré année après année, le revenu total annuel tiré de l'exercice du métier mesuré année après année ou bien la quantité de travail effectuée au cours d'une année mesurée par le nombre de journées de travail rémunérées au cours de l'année. Des variables de ce genre permettent d'étudier l'évolution de l'insertion de chaque individu et de l'ensemble des individus au fil du temps, en tenant compte du fait que cette évolution n'est pas à sens unique : celui qui a travaillé une année peut ne pas travailler l'année suivante et les revenus n'augmentent pas nécessairement chaque année, au contraire, ils peuvent évoluer en dents de scie. Ces considérations seraient évidemment d'intérêt secondaire si elles ne concernaient que les carrières des comédiens, mais on peut imaginer sans peine d'autres situations qui posent des difficultés analogues et qui pourraient gagner à être étudiées avec des outils adaptés. Ces outils existent (voir notamment Green, 1997; Hsiao, 1986; Liang et Zeger, 1986; Neuhaus et al., 1991; Pendergast et al., 1996), et parmi les plus importants d'entre eux figurent les modèles à effets aléatoires dont nous traitons dans cet article.

Nous présentons tout d'abord la notion d'effet aléatoire et ses liens avec les données longitudinales. Nous présentons ensuite deux modèles à effets aléatoires : le modèle logit à effets aléatoires, qui permet d'étudier l'évolution, au fil du temps, de l'appartenance à l'une ou l'autre des deux catégories d'une variable dépendante binaire, et le modèle tobit à effets aléatoires, qui permet d'étudier l'évolution, au fil du temps, d'une variable continue dont une partie de la distribution n'est pas observée. Nous montrons ensuite comment, dans le cadre de nos recherches, nous avons utilisé le premier modèle pour étudier l'évolution du simple fait de travailler ou non chaque année à partir du début de la carrière, et le second pour étudier l'évolution du revenu annuel tiré de l'exercice du métier de comédien compte tenu du fait que, chaque année, de nombreux comédiens ne travaillent pas et donc ne disposent pas d'un revenu de comédien observable. Nous profitons de cette démonstration pour montrer comment ces modèles permettent d'utiliser sans difficulté les variables indépendantes variant dans le temps de même que d'intégrer, parmi les variables

indépendantes, des mesures de contexte qui varient elles-mêmes dans le temps, comme la taille de l'offre et de la demande de travail. Nous expliquons également quelle stratégie nous avons utilisée pour vérifier les effets des variables déterminantes pour nos hypothèses alors que nous étions confrontés à des problèmes de colinéarité assez complexes. Finalement, nous comparons les résultats que l'on obtient en utilisant les modèles à effets aléatoires, les modèles similaires qui ne tiennent pas compte des liens entre les observations faites auprès des mêmes individus et les modèles qui en tiennent compte par un jeu d'effets fixes. Dans la conclusion, nous énumérons un certain nombre d'autres modèles à effets aléatoires qui peuvent être utilisés pour étudier des phénomènes différents de ceux que nous avons abordés.

## **EFFET ALÉATOIRE ET DONNÉES LONGITUDINALES**

### **La notion d'effet aléatoire**

La notion d'effet aléatoire est généralement peu familière aux démographes, qui utilisent soit des données de recensement, soit des données échantillonnées et les analysent au moyen de modèles linéaires dérivés de la régression. Elle est mieux connue des chercheurs qui recueillent leurs données selon un devis expérimental et qui les soumettent à une analyse de la variance.

Les démographes cherchent habituellement à obtenir des échantillons dont la composition est identique à celle de la population dont ils sont tirés. Les variables auxquelles ils s'intéressent ont un nombre fini et connu de catégories qui sont toutes représentées dans les échantillons : le sexe a deux catégories et un échantillon de la population d'une société contiendra des hommes et des femmes; on peut diviser les Québécois en francophones, anglophones et allophones, et un échantillon de la population québécoise contiendra des représentants des trois groupes.

Lorsqu'un échantillon comprend des individus qui appartiennent à toutes les catégories d'une variable, on peut dire que la population des individus a été échantillonnée, mais que la population des valeurs possibles de cette variable n'a pas été échantillonnée : toutes les valeurs possibles de cette variable sont représentées dans l'échantillon. On peut cependant imaginer des circonstances où toutes les valeurs possibles d'une

variable ne sont pas présentes dans un échantillon, c'est-à-dire des cas où l'échantillonnage porte à la fois sur les individus et sur les catégories d'une variable. Le cas le plus simple à imaginer est celui où l'on s'intéresse à l'effet que l'interviewer peut avoir sur les réponses données à un questionnaire. L'enquête sera faite par un certain nombre d'interviewers, mais il est clair qu'elle aurait pu être faite par des interviewers différents. Lorsque l'interviewer devient une variable indépendante dans une analyse qui vise à estimer l'effet des interviewers sur les réponses données par les répondants, chaque répondant se trouve à être affecté d'une valeur pour cette nouvelle variable. On comprend sans difficulté que l'univers des valeurs possibles de cette variable ne se limite pas au groupe d'interviewers qui a été embauché; au contraire, l'univers des catégories de cette variable comprend tous les individus qui auraient pu être embauchés comme interviewers. Le fait de se trouver dans une situation où l'on ne retrouve pas, dans les données, l'ensemble des valeurs possibles de cette variable conduit à se demander ce qu'aurait pu être l'effet de cette variable si on l'avait estimé à partir d'un échantillon différent de ses catégories. Il aurait fort bien pu être différent, ce qui signifie que, lorsqu'on se trouve dans une situation où les catégories d'une des variables indépendantes sont échantillonnées, l'estimation de l'effet de cette variable, en plus d'être affectée de l'imprécision propre aux estimations calculées à partir de données d'échantillon, est affectée d'une imprécision aléatoire supplémentaire. Cette imprécision fait en sorte que l'erreur type de l'estimation du paramètre de cette variable ne peut plus être calculée de la manière habituelle, mais cela n'est ni la conséquence la plus importante de l'existence de cette nouvelle source de variation aléatoire, ni le problème le plus complexe à résoudre : les différents estimateurs de variances robustes permettent de régler ce problème sans grande difficulté. La conséquence la plus importante est que l'estimation de l'effet de la variable obtenue dans de telles circonstances est en réalité échantillonnée au sein d'une population d'estimations *qui varient en fonction des échantillons de catégories* et que ce fait ajoute un *second processus stochastique* au processus stochastique qui est au cœur de tous les avatars du modèle linéaire. Dans le cas le plus simple, celui de la régression linéaire, cela revient à dire que la variance résiduelle de la régression ne peut plus être interprétée simplement comme la portion de la variance de la variable dépendante qui n'est pas expliquée par le modèle, mais bien

comme la somme des résultats de deux processus aléatoires distincts. Autrement dit, la part de la variance de la variable dépendante qui n'est pas expliquée par les effets des variables indépendantes doit être comprise comme la somme de 1) la part de la variance de la variable dépendante provenant du fait que l'effet d'une des variables indépendantes varie lui-même de manière aléatoire en fonction des catégories de cette variable qui sont présentes dans l'échantillon, et de 2) la part de la variance de la variable dépendante qui n'est pas expliquée par le modèle, effets aléatoires compris. En bref, dans le cas de la régression, l'existence d'effets aléatoires force à inclure un second terme aléatoire dans le modèle.

### **Effets aléatoires et données longitudinales**

Le lien entre données longitudinales et effets aléatoires peut paraître lointain, mais il est en réalité immédiat et fondamental.

Dans les études menées à partir d'échantillons longitudinaux, les individus composant ces échantillons sont observés plusieurs fois. En conséquence, les observations réalisées sur un même individu ne sont pas indépendantes les unes des autres. Ce problème, que l'on nomme parfois *corrélacion des données*, a au moins deux conséquences importantes.

La première découle du fait que toute la théorie de l'inférence statistique repose sur le postulat de l'indépendance des observations. Le fait d'inclure un individu dans un échantillon longitudinal détermine les informations que l'on recueillera à tous les moments où l'on observera cet individu. En d'autres termes, ajouter des observations à l'échantillon en ajoutant des informations recueillies auprès d'individus déjà observés n'augmente pas la puissance inférentielle de l'échantillon. Cela est tout à fait analogue à la situation que crée un échantillon de grappes : la puissance inférentielle d'un échantillon formé, par exemple, de tous les membres d'un nombre donné de ménages choisis de manière aléatoire n'est pas fonction du nombre des individus qui composent cet échantillon, mais bien du nombre des ménages qui ont été tirés au hasard.

Deuxièmement, le fait de constituer un échantillon longitudinal fait en sorte que l'on crée une situation analogue à celle des interviewers de l'exemple décrit plus haut. Estimer un modèle linéaire à partir d'observations provenant de mesures répétées auprès des mêmes individus oblige à tenir compte

d'une manière ou d'une autre du fait qu'une partie de la variance de la variable dépendante provient selon toute vraisemblance de la variation intra-individus ou, peut-être de manière plus exacte, de la corrélation intra-individuelle. Si l'on néglige ce fait, les conclusions de l'analyse risquent d'être faussées; dans une régression logistique, par exemple, l'erreur type des variables qui conservent la même valeur dans le temps tend à être sous-estimée alors que celle des variables qui fluctuent dans le temps tend à être surestimée (Fitzmaurice et al., 1993).

## MODÈLES À EFFETS ALÉATOIRES

### Le modèle logit à effets aléatoires

L'usage du modèle logit, également connu sous le nom de régression logistique, est assez répandu. Les démographes qui l'utilisent sont généralement habitués à le voir représenté de la manière suivante :

$$(1) \quad \ln \left( \frac{P(y_i = 1 | x_i)}{1 - P(y_i = 1 | x_i)} \right) = \alpha + x_i' \beta$$

et à interpréter les effets des variables indépendantes en termes de rapports de chances ou en rapports de cotes : lorsque le coefficient associé à une variable indépendante est supérieur à 0, tout accroissement d'une unité de la variable indépendante accroît les chances ou la cote qu'a un individu d'avoir 1 comme valeur de la variable dépendante, alors que lorsque ce coefficient est inférieur à 0, tout accroissement d'une unité de la variable indépendante accroît les chances ou la cote qu'a un individu d'avoir 0 comme valeur de la variable dépendante.

Dans le cadre des modèles linéaires, la manière la plus simple de tenir compte de l'existence d'effets aléatoires est tout simplement de supposer qu'il existe une ordonnée à l'origine différente pour chaque individu et d'estimer les effets des « véritables » variables indépendantes compte tenu de ces multiples ordonnées. Cela se réalise tout simplement en créant une variable dichotomique pour tous les individus de l'échantillon sauf un. L'équation d'une régression de ce genre estimée à partir d'un échantillon de 1000 personnes contiendrait donc 1000 ordonnées à l'origine différentes, c'est-à-dire l'ordonnée à l'origine « conventionnelle », qui serait devenue l'ordonnée de

l'individu qui aurait été choisi comme catégorie de référence, et les 999 ordonnées obtenues au moyen des 999 variables dichotomiques représentant les sources d'information de l'échantillon. Un modèle de ce genre pourrait s'écrire comme suit :

$$(2) \quad \ln \left( \frac{P(y_{it} = 1 | x_{it})}{1 - P(y_{it} = 1 | x_{it})} \right) = \alpha_i + x'_{it} \beta,$$

où  $y_{it}$  représente la valeur de la variable dépendante  $y$  pour l'individu  $i$  au temps  $t$ ,  $\alpha_i$  représente la valeur de l'ordonnée à l'origine du modèle pour l'individu  $i$  (notons que cette valeur est la même pour tous les moments où l'individu  $i$  est observé),  $x_{it}$  est le vecteur des valeurs des variables indépendantes pour l'individu  $i$  au temps  $t$  et  $\beta$  est le vecteur des effets des variables indépendantes.

Les modèles de ce genre sont connus sous le nom de modèles à *effets fixes* parce qu'on y tient compte de l'effet causé par l'échantillonnage des sources d'observation sans modéliser celui-ci comme un effet aléatoire. Cette appellation est équivoque parce que tous les modèles linéaires qui ne comportent pas d'effets aléatoires — comme la régression linéaire conventionnelle ou la régression logistique conventionnelle, par exemple — sont des modèles à effets fixes; les modèles où les sources d'information sont représentées par des ordonnées multiples ne semblent cependant pas être connus sous un autre nom que celui-là.

Parce que les effets de ces modèles sont traités comme s'ils étaient fixes et directement estimables alors qu'ils sont le produit d'un processus d'échantillonnage aléatoire, les estimations des modèles à effets fixes dépendent de l'échantillon où ils sont calculés et ne sont pas généralisables à la population dont est tiré l'échantillon. En d'autres termes, les estimations de ces modèles sont échantillonnées au sein d'une population d'estimations qui varie selon les échantillons de valeurs de la variable indépendante sans que rien, dans le modèle, ne tienne compte de ce fait. Plutôt que de représenter les différences entre les sources d'information par une série d'ordonnées différentes, les modèles à effets aléatoires les représentent en présumant qu'elles sont générées par un *processus stochastique* analogue à celui qu'on suppose responsable des résidus de la régression linéaire conventionnelle.

Dans sa forme la plus simple, le modèle logit à effets aléatoires peut s'écrire :

$$(3) \quad \ln \left( \frac{P(y_{it} = 1 | x_{it})}{1 - P(y_{it} = 1 | x_{it})} \right) = \alpha + x'_{it} \beta + v_i,$$

où le terme  $v_i$  représente l'effet individuel et aléatoire. Pour compléter le modèle, il est nécessaire de définir la loi de probabilité que suit l'effet aléatoire, comme il est nécessaire de définir la loi de probabilité que suivent les résidus d'une régression conventionnelle. On suppose habituellement que  $v_i$  est indépendant de  $\mathbf{x}_i$  et suit une distribution normale de moyenne 0 et de variance  $\sigma_v^2$ .

### Le modèle tobit à effets aléatoires

Nous avons expliqué plus haut que dans la population à laquelle nous nous intéressons, rien ne garantit que tous les individus aient un revenu d'emploi chaque année. Au contraire, le fait d'avoir un emploi est lui-même une variable dépendante, le résultat d'un processus déterminé par des facteurs semblables à ceux qui peuvent être utilisés pour expliquer le revenu. Les situations de ce genre ne sont pas rares. Une étude qui s'intéresserait à l'évolution du revenu d'emploi des femmes serait confrontée à une situation du même genre : il est fréquent que les femmes se retirent du marché du travail après la naissance d'un enfant.

Tobin (1958) a développé un modèle, appelé « tobit », afin de tenir compte de situations semblables. En termes simples, le modèle tobit est un modèle linéaire pour une variable dépendante en principe continue, comme le revenu, mais dont il est impossible d'observer la valeur pour une partie de la distribution et dont on sait que le processus qui gouverne l'impossibilité d'observer cette partie de la distribution est semblable à celui qui gouverne la variable dépendante continue. Formellement, le modèle de la régression tobit peut être présenté comme suit. Une variable  $z$  est présumée dépendre d'un certain nombre de variables indépendantes regroupées dans le vecteur  $\mathbf{x}$ , dont les effets sont regroupés dans le vecteur  $\beta$ . On présume que les valeurs observées de  $z$ , les  $z_i$ , sont la combinaison de la valeur prédite par la composante déterministe du modèle, c'est-à-dire  $x'_i \beta$ , et d'un résidu,  $\varepsilon_i$ , dont la valeur varie de manière aléatoire pour chaque individu. Jusqu'ici, ce modèle est identique à celui de la régression ordinaire. On suppose cependant que la variable  $z$  n'est pas observable directement, mais qu'on observe plutôt la variable  $y$ , qui vaut  $z$  lorsque  $z$  est plus grand

que 0, mais qui vaut exactement 0 lorsque  $z$  est inférieur ou égal à 0. Autrement dit, le modèle complet est décrit par les deux équations suivantes :

$$(4) \quad \begin{aligned} z_i &= \alpha + \mathbf{x}'_i \beta + \varepsilon_i, \\ y_i &= \begin{cases} z_i & \text{si } z_i > 0 \\ 0 & \text{si } z_i \leq 0. \end{cases} \end{aligned}$$

Ce modèle repose donc sur l'idée que la variable observée est le résultat de deux processus distincts qu'on suppose régis par les mêmes variables indépendantes. Le premier processus détermine la valeur de la variable *latente* continue ( $z$ ) et s'apparente à ceux que l'on modélise au moyen de la régression ordinaire. Le second processus détermine quant à lui si la variable *observée* est une version censurée ou non de la variable latente. Puisque ce second processus n'a que deux issues possibles (la variable observée est censurée ou non), son résultat est une variable dichotomique comme celles que l'on analyse au moyen de la régression logistique ou probit. Le modèle tobit a été développé en représentant le processus qui régit la dichotomie à l'aide de la régression probit.

Étant donné ce qui précède, on ne sera pas étonné de voir que le modèle tobit à effets aléatoires comprend tout d'abord une équation qui relie la variable dépendante du modèle,  $z$ , aux variables indépendantes, auxquelles s'ajoutent à la fois un effet aléatoire et un résidu :

$$(5) \quad z_{it} = \alpha + \mathbf{x}'_{it} \beta + v_i + \varepsilon_{it}.$$

Dans cette équation,  $z_{it}$  représente la valeur que prend la variable latente continue pour l'observation de l'individu  $i$  au temps  $t$ ,  $\alpha$  représente la valeur de l'ordonnée à l'origine,  $\mathbf{x}_{it}$  désigne l'ensemble des variables indépendantes telles que mesurées au temps  $t$  sur l'individu  $i$ ,  $\beta$  est le vecteur des coefficients affectant ces variables,  $v_i$  représente la valeur de l'effet aléatoire associé à l'individu  $i$  (rappelons que cet effet varie d'un individu à l'autre, mais ne prend qu'une seule valeur pour toutes les observations réalisées auprès du même individu) et  $\varepsilon_{it}$  constitue l'erreur du modèle, qui diffère pour chaque observation. Comme précédemment, on présume que  $z_{it}$  provient d'une distribution continue et que la valeur observée  $y_{it}$  est une version « censurée à gauche » de  $z_{it}$ .

Les manuels qui accompagnent les progiciels permettant l'estimation de modèles à effets aléatoires donnent habituellement les détails de la méthode d'estimation employée. Hamerle

et Ronning (1995) présentent pour leur part un exposé concis sur l'estimation par maximum de vraisemblance des modèles abordés dans cet article.

## **L'UTILISATION DES MODÈLES À EFFETS ALÉATOIRES**

### **Un modèle de la dynamique de la carrière des acteurs**

Comme nous l'avons expliqué plus haut, nous nous sommes intéressés aux modèles linéaires à effets aléatoires dans le cadre d'un programme de recherche sur les carrières des comédiens français. Dans nos analyses, les variables dépendantes sont le fait de travailler ou non durant chacune des années qui suivent l'année du premier contrat et le revenu annuel total tiré du métier d'acteur durant chacune de ces années. Par définition, chaque individu a travaillé et obtenu un revenu au cours de l'année du début de sa carrière, mais, au cours de chacune des années subséquentes, chacun peut travailler ou non et tirer ou non un revenu du métier d'acteur. Le revenu est nul les années où l'individu ne travaille pas et supérieur à zéro les années où il travaille; lorsqu'il n'est pas nul, il peut prendre n'importe quelle valeur positive.

Nous présentons ailleurs en détail (Laplante et al., à paraître) notre modèle de la dynamique des carrières des acteurs professionnels. Pour les besoins de la discussion, nous en donnons ici une version schématique.

La carrière des acteurs est le résultat d'un processus d'action sous contraintes. Les contraintes sont diverses : certaines sont liées à la nature de la profession (l'attrition est très forte en début de carrière, l'auto-sélection jouant ici le rôle que la sélection formelle joue dans la plupart des autres professions), d'autres à la vie personnelle des individus (notamment à la présence de personnes à charge, qui peut changer ce que l'individu attend de son travail), d'autres à la structure dans laquelle doivent se dérouler les carrières (en particulier l'évolution de l'offre et de la demande de travail). L'action est en partie individuelle et en partie collective. Les acteurs et à tout le moins certains des employeurs ont en commun la volonté de faire en sorte que les œuvres soient réalisées en dépit d'intérêts autrement divergents et en dépit de la difficulté générale de trouver les moyens nécessaires à la réalisation des œuvres. Les traces de l'action individuelle se constatent surtout dans l'intention de faire carrière comme comédien, dans le fait d'adopter

le mode d'action qui permet de mener une carrière faite d'une suite de contrats éphémères, notamment dans la capacité de se construire un capital social en se développant un réseau personnel d'employeurs potentiels, et dans la manière d'utiliser le régime d'assurance-chômage des intermittents du spectacle qui, en France, compte pour une part non négligeable du revenu des comédiens professionnels. Les traces de l'action collective, conçues comme les résultats « visibles » d'un ensemble de normes partagé par les acteurs et leurs employeurs, se constatent surtout en comparant l'efficacité de deux formes de valorisation dans la sélection des individus par les employeurs : l'expérience, mesurée par le nombre de jours travaillés en carrière, et la valeur marchande du travail, mesurée par le cachet quotidien moyen perçu. Ces deux critères renvoient à deux modes de sélection différents : dans le premier cas, les employeurs choisissent les acteurs en fonction de ce qu'ils ont fait alors que dans le second, les employeurs choisissent les acteurs en fonction de ce qu'ils ont récemment obtenu de leurs employeurs et donc, en principe, en fonction de ce que leur présence dans une production devrait rapporter à l'employeur qui les embauchera.

Nous suivons, de 1988 à 1996, les comédiens qui ont commencé leur carrière entre 1987 et 1995. Au cours de cette période, l'offre et la demande de travail n'ont évidemment pas été constants. Les histoires individuelles se construisent donc dans un contexte où l'histoire collective se déroule elle aussi. La nature des contraintes structurelles qui nous intéressent ne change pas — nous présumons que les effets de ces contraintes sont constants —, mais il est évident qu'elles évoluent au fil du temps et qu'elles sont elles-mêmes des variables longitudinales. Formellement, notre modèle de la carrière des comédiens doit donc tenir compte à la fois de variables individuelles qui varient dans le temps, comme l'expérience ou le fait d'avoir des personnes à charge, et de variables de structure qui varient dans le temps, comme l'offre et la demande de travail.

### **Données**

Nous utilisons les données administratives de la Caisse des congés spectacle, l'organisme français qui gère la paye de vacances des intermittents du spectacle. Les données originales tiennent en trois fichiers différents pour chaque année

d'observation. Le premier fichier, le fichier des employés, contient des informations sur les individus comme le sexe, l'année de naissance, le métier exercé au cours de l'année, le lieu de résidence, le nombre de personnes à charge, etc. Le deuxième fichier contient des informations sur chacun des employeurs. Le troisième fichier, le fichier des contrats, contient une ligne d'information pour chaque contrat conclu entre un intermittent et un employeur au cours de l'année. Chaque ligne de ce fichier contient deux identifiants qui permettent de relier les contrats à l'employé et à l'employeur concernés, les dates de début et de fin du contrat, le nombre de jours rémunérés, le cachet minimum prévu par les conventions ainsi que le cachet réellement versé.

Le fichier de données constitué aux fins de cette étude est d'une forme très semblable à celle des fichiers de données préparés pour les analyses de biographies en temps discret. La différence essentielle est que ce fichier est constitué d'une ligne d'information pour chacune des périodes pendant lesquelles l'individu a été observé et non pas d'une ligne pour chacune des périodes qui précèdent ou contiennent le moment où se fait le changement de valeur de la variable dépendante. Chaque ligne d'information contient la valeur de la variable dépendante au cours de cette période, ainsi que la valeur de chacune des variables indépendantes au cours de la même période. Comme dans les modèles d'analyse des biographies, les variables indépendantes peuvent être fixes ou au contraire varier dans le temps. La construction des premières ne pose aucun problème particulier. La construction des secondes demande un travail analogue à celui qui est fait pour préparer les variables indépendantes variant dans le temps des analyses de biographies en temps discret.

La mise en forme des données exige que l'on synthétise l'information pour la période que l'on utilise comme unité de temps. L'unité de temps choisie est l'année civile, en partie pour des raisons pratiques — les données sont traitées par la Caisse sur une base annuelle — et en partie pour des raisons théoriques : le revenu se calcule généralement sur la base de l'année fiscale, le droit à l'assurance-chômage est déterminé par le nombre d'heures de travail rémunéré effectué au cours d'une période de douze mois, l'activité du monde du spectacle est organisée autour de la notion de saison, etc.

Les modèles longitudinaux à effets aléatoires permettent d'inclure des variables qui décrivent aussi bien l'histoire de

l'individu que l'évolution du contexte dans lequel il se trouve. Une variable indépendante longitudinale qui décrit l'histoire d'un individu peut être une simple description de l'état dans lequel se trouve un individu sous un certain rapport à un moment donné (avoir ou non une personne à charge au cours de la période, par exemple) ou, encore, représenter l'état d'un processus en cours pendant la période (l'expérience de travail depuis le début de la carrière, par exemple, qui ne peut pas décroître et dont la valeur est susceptible d'augmenter chaque année). Une variable indépendante longitudinale peut également être la mesure d'un phénomène de contexte qui varie au fil du temps, comme la taille de la demande et de l'offre de travail au cours d'une année, qui changent d'une année à l'autre mais prennent la même valeur pour tous les individus pendant une période donnée. L'estimation des effets de caractéristiques de contexte affectant simultanément tous les individus est par définition impossible avec des données et des modèles transversaux, mais elle est relativement simple à réaliser avec des données longitudinales et les modèles qui permettent de les utiliser.

### **Problèmes liés à l'opérationnalisation du modèle**

Parmi les variables dont nous cherchions à estimer l'effet sur la dynamique du travail et du revenu des artistes dramatiques se trouvent des variables décrivant les parcours individuels et d'autres décrivant l'évolution du contexte dans lequel s'inscrivent ces parcours individuels. Il existe au sein de ces variables une structure de causalité complexe qui engendre, lorsque ces variables sont intégrées dans un même modèle, des problèmes de colinéarité importants. Dans ces circonstances, nous risquons de ne pas estimer correctement l'effet de chacune des variables indépendantes et d'attribuer à certaines d'entre elles une influence qui ne leur est pas propre. La comparaison de modèles emboîtés, qui permet généralement de faire apparaître la structure des relations entre les variables indépendantes et la variable dépendante (Davis, 1985), risquait de ne pas suffire pour démêler l'écheveau auquel nous étions confrontés. Nous avons donc adopté une approche différente, parfois utilisée en démographie (voir par exemple Wu, 1999). Pour résoudre les problèmes découlant d'une forte corrélation entre deux variables indépendantes dans notre modèle, nous avons ainsi remplacé l'une de ces variables par le résidu de la régression de

cette variable sur la seconde. Une fois intégré au modèle en tant que variable indépendante, ce résidu est affecté d'un coefficient qui équivaut à un coefficient de régression semi-partiel (Pedhazur, 1982). Par ce procédé, nous pouvons être plus assurés que l'effet estimé d'une variable indépendante importante dans l'une ou l'autre de nos hypothèses est vraiment un effet net. Les exemples qui suivent permettront de comprendre les avantages de cette technique dans l'étude des phénomènes dynamiques.

### *Le nombre d'artistes dramatiques actifs*

Le nombre d'acteurs actifs au cours d'une année, c'est-à-dire le nombre de ceux qui ont décroché un contrat, peut être considéré comme une mesure approximative du niveau d'activité qui a régné cette année-là sur le marché du travail des artistes dramatiques. Toutes choses égales par ailleurs, on doit normalement s'attendre à ce que la probabilité de travailler et le revenu d'emploi augmentent avec le nombre d'acteurs embauchés. Toutefois, le nombre d'acteurs employés est aussi lié à la valeur des cachets et au nombre de contrats. Avant d'intégrer le nombre d'acteurs embauchés à notre modèle explicatif, il faut d'abord faire en sorte que la part de la variation de la variable dépendante qui sera imputée à ce facteur sera bien due à ce que celui-ci mesure et non aux autres facteurs auxquels il est associé. Pour éviter cette ambiguïté, nous utilisons le résidu de la régression linéaire du nombre d'acteurs embauchés au cours d'une année sur les cachets et le nombre de contrats au lieu du nombre brut d'acteurs embauchés. Si le nombre d'acteurs embauchés dépasse le nombre moyen de ceux qui auraient été embauchés étant donné la valeur des cachets et le nombre des contrats de cette année (en d'autres termes, si la valeur observée est supérieure à la valeur prédite), le résidu sera positif et on l'associera à un niveau d'activité plus élevé que la moyenne. Inversement, si le nombre des acteurs qui ont été embauchés au cours d'une année est inférieur au nombre moyen qui aurait été embauché étant donné la valeur des cachets et le nombre des contrats, le résidu sera négatif. On suppose que ce résidu est lié positivement à la probabilité de travailler et au revenu des individus.

### *La demande de travail mesurée en occasions*

Bien que nous présumions que la probabilité individuelle de travailler s'accroît avec la demande globale de travail, la nature

même du travail d'acteur fait en sorte que la quantité de travail requise par les employeurs n'est pas divisée en emplois permanents à temps plein. La probabilité de travailler, que nous définissons comme la probabilité d'avoir obtenu au moins un contrat d'acteur au cours d'une année, dépend donc aussi très certainement du nombre de contrats qui sont octroyés à des acteurs au cours de la même année.

Cependant, en comparaison de la valeur totale des cachets, le nombre total de contrats ne peut être aussi directement assimilé à une mesure de la demande de travail dans l'univers particulier des artistes dramatiques. Si le rapport entre le nombre des contrats et leur valeur totale, exprimée en francs constants, demeurerait approximativement constant au cours de la période — autrement dit, si la valeur moyenne du contrat était approximativement constante —, les deux quantités seraient des mesures différentes d'une même chose et il vaudrait mieux n'en retenir qu'une dans les équations qui cherchent à modéliser les devenirs individuels. Le nombre de contrats conviendrait davantage que leur valeur dans le modèle qui cherche prédire la probabilité d'obtenir un contrat, alors que le choix de l'une ou l'autre mesure serait moins évident dans le cas de notre équation de prédiction du revenu, qui intègre formellement un modèle de sélection dichotomique, dans lequel le nombre de contrats serait plus approprié, et un modèle de prédiction du revenu, dans lequel la valeur totale des cachets serait a priori un choix plus judicieux.

Nous ne sommes cependant pas dans cette situation. Au cours de la période étudiée, le nombre de contrats n'évolue pas en raison directe de la valeur totale de ceux-ci : au contraire, le nombre de contrats augmente plus rapidement que leur valeur totale. En conséquence, les deux quantités contiennent des informations qui sont en partie corrélées, et donc redondantes, et en partie divergentes, et donc originales. La vertu consiste donc à chercher à utiliser toute l'information originale et à distinguer, dans la modélisation et dans l'interprétation, ce qui est corrélé de ce qui ne l'est pas. La chose n'est pas simple. Nous utilisons des données longitudinales, ce qui crée des corrélations intra-individuelles que nous contrôlons, autant que faire se peut, en utilisant des modèles à effets aléatoires, mais également des corrélations entre les données agrégées qui s'ajoutent aux corrélations découlant des relations causales qui existent entre les variables agrégées que nous utilisons comme facteurs explicatifs.

Plutôt que d'utiliser directement le nombre total de contrats réalisés au cours d'une année, nous utilisons le résidu de la régression linéaire du nombre de contrats sur le volume des cachets la même année. Dans le contexte qui prévaut au cours de la période que nous étudions, cette quantité n'est pas une mesure de la demande de travail. Au contraire, pour une année donnée, une valeur positive de cette quantité indique que davantage de contrats ont été conclus que ce qui aurait été prévisible étant donné le volume des salaires. Inversement, une valeur négative indique qu'au cours de l'année, on a conclu moins de contrats que ce que le volume des salaires aurait permis de prévoir. L'examen de l'évolution des caractéristiques du marché du travail montre que cette quantité varie à la hausse entre 1988 et 1996, traduisant une tendance vers des contrats de plus courte durée.

Pour un volume de cachets donné, la probabilité de décrocher un contrat devrait varier en fonction du nombre global des contrats. On s'attend à ce que la probabilité de travailler au cours d'une année varie en raison du résidu de la régression du nombre global des contrats sur le volume des cachets.

### *L'usage stratégique de l'assurance-chômage*

Nous nous intéressons à la place que l'usage stratégique de l'assurance-chômage peut avoir dans la dynamique du travail et du revenu. Les données à notre disposition ne renseignent cependant pas sur l'admissibilité ou sur la perception d'indemnités de chômage. Elles contiennent en revanche la durée, en jours, de chaque contrat.

Pour être admissible à des indemnités, le régime d'assurance-chômage des intermittents du spectacle exige que l'on ait exercé une activité salariée dans une ou plusieurs entreprises du spectacle durant 507 heures au cours des douze mois précédant la demande. Dans le cas les individus rémunérés au cachet, le nombre d'heures correspondant à l'engagement est calculé à raison de 8 heures par jour pour un contrat couvrant une période d'au moins 5 jours chez un même employeur, et de 12 heures par jour pour un contrat de 4 jours ou moins. En appliquant les règles de calcul du régime d'assurance-chômage aux durées contenues dans les données de la Caisse, nous avons pu déterminer si chaque acteur avait vraisemblablement accumulé ou non le minimum d'heures pour se qualifier au cours d'une année.

En elle-même, cependant, cette information est de peu d'intérêt : tout acteur qui a réussi son insertion professionnelle aura travaillé suffisamment pour obtenir des indemnités s'il en a besoin. Pour évaluer l'effet d'un éventuel usage stratégique de la qualification à l'assurance-chômage sur la probabilité de travailler et sur le revenu d'emploi, il est nécessaire non pas de savoir si un acteur a travaillé suffisamment pour avoir droit à des indemnités, mais plutôt de savoir s'il a obtenu le droit à des indemnités si rapidement qu'il est raisonnable de croire qu'il a aménagé son travail dans le but de les obtenir. Il est évidemment impossible d'obtenir une mesure directe d'une pareille intention, mais il est en revanche raisonnable de supposer que les acteurs qui agissent de la sorte parviennent, en moyenne, à atteindre le seuil des 507 heures plus rapidement que les acteurs qui accordent une moins grande importance à cette question. En comparant le fait de s'être ou non qualifié pour l'assurance-chômage au cours d'une année civile, mesuré comme nous l'avons expliqué ci-dessus, au nombre de jours travaillés par le même acteur au cours de la même année, on peut ordonner les acteurs selon qu'il est plus ou moins probable qu'ils aient aménagé leur temps de travail dans le but de se qualifier rapidement à l'assurance-chômage. En termes techniques, nous avons effectué la régression logistique avec effets aléatoires du fait de se qualifier ou non à l'assurance-chômage sur le nombre de jours ouvrés et obtenu ainsi la probabilité théorique qu'un individu se qualifie compte tenu du nombre de jours pendant lesquels il a travaillé. Par la suite, nous avons calculé la différence entre la qualification à l'assurance-chômage (codée 0 ou 1) et la probabilité théorique de se qualifier. Nous considérons cette différence comme un indicateur de la propension d'un individu à faire un usage stratégique de l'assurance-chômage. Il est supposé que, pendant une année donnée, les chances de décrocher un contrat et le revenu d'un individu sont liés à la propension qu'il a manifestée l'année précédente.

### *Autres facteurs*

La plupart des autres variables intégrées aux modèles que nous avons utilisés afin de décrire la dynamique du travail et du revenu sont des mesures directes qui ne nécessitent pas de descriptions supplémentaires. Trois facteurs méritent cependant d'être explicités, soit l'« insertion professionnelle », la « valeur marchande » et le « capital social ». Comme les varia-

bles présentées dans les sections précédentes, ces trois facteurs sont des mesures indirectes.

On peut raisonnablement supposer que la probabilité de travailler d'un individu et le revenu qu'il a gagné au cours d'une année donnée sont liés à son degré d'insertion professionnelle l'année précédente. Nous disposons de deux informations à ce sujet pour chaque individu constituant l'échantillon, soit le nombre annuel de jours de travail effectués en tant que comédien et le revenu tiré du métier de comédien. Ces quantités renvoient à des aspects importants et distincts de l'insertion professionnelle, mais elles sont évidemment étroitement corrélées. Plutôt que de retenir l'une au détriment de l'autre, nous avons construit une mesure de l'insertion professionnelle à l'aide de la première composante obtenue de l'analyse en composantes principales de ces deux quantités.

Les comédiens peuvent négocier le montant de leur cachet à chaque nouveau contrat. Certains ont acquis une notoriété ou ont un talent qui justifie une rémunération plus élevée et il est possible, avec les données de la Caisse, de suivre l'évolution de leur rémunération ou, en d'autres termes, de leur « valeur marchande ». Nous avons estimé que la rémunération quotidienne moyenne d'un comédien pendant une année pouvait constituer un bon indicateur de sa valeur marchande l'année suivante. L'examen des données a montré que l'évolution de la rémunération quotidienne moyenne est reliée de manière assez étroite à l'insertion professionnelle, ce qui rend un peu difficile l'estimation de l'effet propre de chacune des deux variables, bien que les deux notions soient clairement distinctes. Pour éviter les problèmes de colinéarité, nous avons utilisé la rémunération quotidienne moyenne nette de l'effet de l'insertion professionnelle comme mesure de la valeur marchande du comédien. On obtient la rémunération quotidienne moyenne nette en utilisant le résidu de la régression de la rémunération quotidienne moyenne sur l'insertion. La mesure ainsi obtenue peut être interprétée comme la part de la valeur marchande du comédien qui ne dépend pas de l'intensité de son activité professionnelle. Nous nous attendons à ce que cette variable ait un effet positif sur la probabilité de travailler et sur le revenu tiré du travail de comédien.

Enfin, il est plausible que le capital social d'un comédien (mesuré par exemple par la taille et la qualité de son réseau social et professionnel) influence ses chances de travailler et, donc, son revenu. En l'absence d'information sur les réseaux

des comédiens, nous avons construit une mesure approximative de l'étendue de leur réseau professionnel en utilisant le nombre d'employeurs différents pour lesquels ils ont travaillé depuis leurs débuts. Comme cette mesure est fortement associée à l'expérience (mesurée en jours cumulés de travail), nous l'avons remplacée par le nombre cumulé d'employeurs différents net de l'expérience, c'est-à-dire le résidu de la régression du nombre cumulé d'employeurs différents pour lesquels a travaillé un artiste sur son expérience, obtenu par régression de Poisson à effets aléatoires. Ces résidus ont été ajustés de manière à ce que leur valeur minimum soit zéro. Cette quantité est une mesure de l'importance relative du réseau professionnel du comédien compte tenu de son expérience, et son effet peut être interprété comme celui du réseau, net de l'effet de l'expérience. On s'attend à ce que la probabilité de travailler et le revenu augmentent en raison directe de cette mesure.

### **Comparaison des résultats**

Les tableaux 1 et 2 de même que la figure 1 illustrent certaines des propriétés de différentes méthodes d'estimation du modèle logit et du modèle tobit lorsqu'on les utilise avec des données longitudinales. Les exemples qui y sont présentés ont été construits à partir d'un échantillon aléatoire de la population des acteurs français qui ont entrepris leur carrière entre 1987 et 1995.

Nous utilisons le modèle logit pour estimer les effets de différentes variables indépendantes sur la probabilité d'obtenir au moins un contrat au cours de chacune des années qui suit l'année du début de la carrière. Le tableau 1 donne les résultats de l'estimation de quatre modèles logit différents : le modèle logit « ordinaire », le modèle logit à effets fixes, le modèle logit à effets fixes conditionnels et le modèle logit à effets aléatoires. Avec le modèle logit ordinaire, les estimations sont obtenues en maximisant la vraisemblance conjointe des observations comme si elles étaient indépendantes, alors qu'elles ne le sont pas.

Nous avons estimé le modèle à effets fixes et le modèle à effets fixes conditionnels, dont les résultats apparaissent dans les deuxième et troisième colonnes, à des fins illustratives. Nous avons créé le modèle à effets fixes en ajoutant « manuellement » une constante spécifique à chaque individu dans l'équation du modèle ordinaire. L'estimation des effets fixes

pose problème dans le cas des individus qui conservent d'une observation à l'autre la même valeur à la variable dépendante. Dans ces conditions, l'association entre la variable dépendante et l'effet individuel est parfaite et fait tendre la valeur de cet effet vers l'infini. Ces cas sont donc exclus de l'analyse, ce qui explique pourquoi les modèles à effets fixes ont été estimés avec 1416 observations, provenant de 251 individus, alors que les autres modèles ont été estimés avec 2952 observations réalisées auprès de 625 individus (l'échantillon complet contient 750 individus et 3904 observations). Il va sans dire que cette caractéristique des modèles à effets fixes réduit considérablement leur intérêt dans les études longitudinales puisqu'elle force à restreindre l'analyse aux seuls individus dont la valeur de la variable dépendante varie dans le temps. Nous n'avons pas inclus les estimations des effets individuels dans le tableau — il y en a 250 —, mais l'histogramme de la figure 1 permet de voir de quelle manière ils se distribuent. Pour fins de comparaison, la densité de probabilité normale centrée à zéro et de même variance que la distribution empirique est superposée à l'histogramme; on remarquera que les effets fixes suivent une distribution approximativement normale.

Le modèle à effets fixes conditionnels est une variante du modèle à effets fixes qui a certaines applications en contexte expérimental et dont on a proposé l'usage pour les études longitudinales. Le modèle à effets fixes, dont les résultats apparaissent dans la deuxième colonne, s'obtient en estimant une ordonnée à l'origine différente pour chaque individu, mais sa vraisemblance est tout de même calculée comme si toutes les observations avaient été obtenues de manière indépendante, ce qui n'est pas le cas ici. Le modèle à effets fixes conditionnels, que nous ne présentons pas en détail (voir plutôt Hamerle et Ronning, 1995), évite l'estimation des ordonnées individuelles et calcule la vraisemblance au niveau des individus; en cela, il est semblable au modèle à effets aléatoires. Il a cependant l'inconvénient, comme le modèle à effets fixes, de ne pas permettre l'usage d'informations provenant d'individus dont la valeur de la variable dépendante ne varie pas dans le temps. En outre, les variables indépendantes dont la valeur est constante pour chaque individu — ici l'âge en début de carrière et le sexe — sont également exclues, parce que la fonction de vraisemblance de ce modèle s'appuie sur les variations des variables indépendantes pour déterminer la probabilité de se trouver à chaque période dans un état plutôt que dans l'autre.





TABLEAU 2 — Résultats de l'estimation des modèles tobit quant à l'effet des différents facteurs sur le revenu annuel tiré des arts dramatiques <sup>a</sup>

	Tobit « ordinaire »		Tobit à effets aléatoires	
	$\beta$	$\hat{\sigma}_\beta$	$\beta$	$\hat{\sigma}_\beta$
<b>Variables contextuelles</b>				
Demande de travail	0,003	0,003	0,004	0,002
En espèces	-0,003	0,005	-0,001	0,005
En occasions de travail	0,112	0,036 ***	0,099	0,033 ***
Artistes dramatiques actifs				
Résidence [ <i>Île-de-France</i> ]	0,302	0,576	0,700	0,820
<i>Rhône-Alpes</i>	-1,421	0,625 **	-2,219	0,858 ***
<i>Provence-Côte d'Azur</i>	0,451	0,322	0,944	0,449 **
<i>Ailleurs en France</i>	-4,714	1,342 ***	-4,390	1,646 ***
<i>Hors de France</i>	-8,125	0,638 ***	-8,386	0,715 ***
<i>Inconnue</i>				
<b>Variables individuelles</b>				
Âge en début de carrière				
A	0,086	0,146	0,278	0,220
A <sup>2</sup>	-0,006	0,008	-0,017	0,012
A <sup>3</sup>	7,94E-05	1,17E-04	2,54E-04	1,78E-04
Sexe [M]	0,777	1,062	1,900	1,570
Interactions Âge et Sexe				
A×F	-0,024	0,203	-0,262	0,303
A <sup>2</sup> ×F	-0,003	0,011	0,009	0,016
A <sup>3</sup> ×F	7,77E-05	1,57E-04	-1,19E-04	2,36E-04
Personnes à charge [0]	0,794	0,331 **	1,272	0,439 ***
Durée de la carrière				
D	0,620	0,396	-0,252	0,388
D <sup>2</sup>	-0,175	0,129	0,048	0,121
D <sup>3</sup>	0,011	0,012	-0,006	0,011

Capital social, $t-1$	0,058	0,009 ***	0,086	0,012 ***
« Valeur marchande », $t-1$	0,691	0,115 ***	0,240	0,119 **
Expérience	0,218	0,174	-0,050	0,213
Propension à utiliser l'assurance-chômage, $t-1$	0,235	0,820	-0,057	0,779
Secteur d'activité <sup>b</sup> , $t-1$				
<i>Audiovisuel</i>	0,017	0,346	-0,094	0,347
<i>Cinéma</i>	-0,492	0,351	-0,270	0,353
<i>Spectacle vivant</i>	0,737	0,400 *	-0,120	0,424
<i>Autre</i>	0,267	0,476	0,399	0,473
Statut déclaré, $t-1$ [ <i>Acteurs</i> ]	-2,626	0,424 ***	-3,865	0,605 ***
Insertion professionnelle, $t-1$	4,642	0,276 ***	3,222	0,294 ***
<b>Autres variables</b>				
Année (tendance linéaire)	-0,409	0,232 *	-0,501	0,220 **
Constante	811,650	460,142 *	993,321	436,833 **
$\hat{\sigma}_\varepsilon$		5,776		5,032
$\hat{\sigma}_v$				3,433
$\hat{\rho}$				0,318
Observations		2952		2952
Observations censurées		1328		1328
Individus		625		625
Log-vraisemblance		-5905,77		-5867,42

a. Données de la Caisse des congés spectacle de France.

b. Les secteurs d'activité ne sont pas exclusifs : on peut travailler dans plus d'un au cours d'une année.

\* significatif à 0,10; \*\* à 0,05; \*\*\* à 0,01.

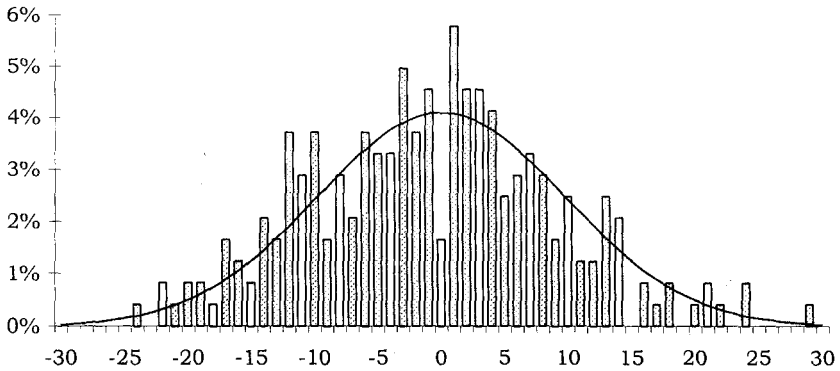


FIGURE 1 — Distribution empirique des effets estimés à l'aide du modèle logit à effets fixes

Les résultats du modèle à effets aléatoires se trouvent dans la quatrième colonne. Le modèle est estimé en utilisant tous les individus et toutes les variables, ce que les deux modèles à effets fixes ne permettent pas. Ces résultats ne peuvent donc être comparés qu'avec ceux du modèle ordinaire. Nous commentons brièvement leurs résultats ci-dessous.

Les effets des variables contextuelles sont similaires mais, dans la plupart des cas, les coefficients sont plus élevés (en valeurs absolues) dans le modèle à effets aléatoires que dans le modèle ordinaire. L'effet de la demande de travail mesurée en espèces (somme annuelle des cachets) devient significatif au seuil de 0,10 dans le modèle à effets aléatoires, mais l'ampleur de l'effet lui-même ne change pas. L'effet associé au nombre d'artistes dramatiques actifs au cours de l'année est significatif et varie peu d'un modèle à l'autre. La variation des effets imputés aux régions de résidence est plus marquée même si les tendances demeurent les mêmes : les effets estimés par le modèle à effets aléatoires sont plus importants dans des proportions qui varient de 19 % à 119 %. L'effet de la présence de personnes à charge est plus élevé de 80 % dans le modèle à effets aléatoires alors que celui du capital social l'est de 76 %. L'effet de la valeur marchande du travail de l'acteur au cours de l'année précédente, qui augmente significativement la probabilité de trouver du travail selon les résultats du modèle ordinaire, n'a plus d'effet significatif dans le modèle à effets aléatoires. L'expérience a un effet d'ampleur comparable dans les deux modèles, mais l'erreur type du coefficient du modèle à

effets aléatoires est nettement plus élevée. L'effet de l'intention de faire carrière comme acteur, mesurée par le fait de se déclarer acteur plutôt que figurant, est deux fois plus important dans le modèle à effets aléatoires. En résumé, utiliser le modèle ordinaire nous conduirait à des conclusions similaires à celles que permet le modèle à effets aléatoires, sauf en ce qui concerne deux variables : avec le modèle ordinaire, nous concluons que la demande de travail mesurée en espèces n'a aucune incidence sur la probabilité qu'a un comédien de décrocher un contrat, mais que la valeur marchande de cet individu augmente ses chances, alors que nous concluons le contraire avec le modèle à effets aléatoires.

Le tableau 2 indique toutefois que la valeur marchande d'un individu a une incidence positive sur son revenu, mais que la demande de travail en espèces n'en a pas, tant selon le modèle tobit ordinaire que selon le modèle à effets aléatoires. Le coefficient associé à la valeur marchande est cependant moins important dans le modèle à effets aléatoires. À deux exceptions près, les deux modèles inciteraient d'ailleurs à conclure que les mêmes variables jouent un rôle dans la dynamique du revenu des comédiens (bien que les coefficients affectant ces variables puissent varier de façon appréciable d'un modèle à l'autre). La première exception touche le fait d'avoir travaillé ou non dans le spectacle vivant au cours de l'année précédente : cette variable a un effet positif sur le revenu dans le modèle « ordinaire », mais n'en a aucun dans le modèle à effets aléatoires. La seconde a trait à l'influence du lieu de résidence. Dans le modèle ordinaire, les individus vivant en Provence-Côte d'Azur ou hors de France de même que ceux dont le lieu de résidence n'est pas connu ont un revenu significativement inférieur à celui des résidents de l'Île-de-France. Cela vaut également dans le modèle à effets aléatoires, mais dans celui-ci le fait de résider « ailleurs en France » (c'est-à-dire hors des trois autres grandes régions circonscrites) est associé à un revenu moyen significativement plus élevé que celui des résidents de l'Île-de-France.

Au bas des tableaux 1 et 2 apparaît également une estimation de la variabilité des effets aléatoires individuels,  $\hat{\sigma}_v$ , ainsi que la proportion de la variance totale attribuable aux effets individuels,  $\hat{\rho}$ . Cette proportion est d'environ 0,7 dans le modèle logit à effets aléatoires, plus du double de la proportion obtenue dans le modèle tobit à effets aléatoires. En général, plus la part de la variance attribuable aux effets aléatoires est

élevée, plus les résultats d'un modèle intégrant de tels effets sont susceptibles de différer de ceux d'un modèle qui n'en comporte pas.

## **CONCLUSION**

Les modèles à effets aléatoires permettent d'envisager l'étude de phénomènes longitudinaux qui ne peuvent être adéquatement décrits avec les modèles à risques proportionnels. C'est le cas des phénomènes dont la variable dépendante est quantitative. Dans cet article, nous avons effleuré la régression à effets aléatoires simplement pour préparer la présentation du modèle logit et du modèle tobit à effets aléatoires, mais il est bien évident que la régression à effets aléatoires est un modèle en soi très utile pour qui veut étudier un phénomène dont la variable dépendante est simplement quantitative. L'objet que nous étudions, comme beaucoup d'objets plus proprement démographiques, ne peut cependant pas se ramener à un processus dont la variable dépendante serait simplement quantitative. Les modèles logit et probit à effets aléatoires permettent d'aborder les phénomènes dont la variable dépendante est une dichotomie et peut être pensée comme un changement d'état, mais qu'il serait peut-être plus juste de concevoir comme une suite de passages d'un état à l'autre plutôt que comme un changement d'état unique. L'exemple que nous utilisons oppose le fait de travailler à celui de ne pas travailler, mais les cas similaires sont nombreux. Le fait de vivre en union ou pas, surtout au début de l'âge adulte, gagnerait peut-être à être étudié en définissant une période d'observation identique pour tous les individus, par exemple de 16 ou 18 ans à 25 ou 30 ans, et en cherchant à déterminer les facteurs qui font que de mois en mois, tout au long de cette période, hommes et femmes vivent seuls ou en couple. Les études sur l'insertion des migrants en milieu urbain gagneraient peut-être à étudier non plus simplement le temps nécessaire au migrant pour se trouver un premier emploi, mais aussi les facteurs qui font que, pendant une période d'observation qui peut varier d'un individu à l'autre, celui-ci passe du travail au chômage et vice-versa.

Comme nous l'avons expliqué plus haut, les modèles à effets aléatoires sont en fait une généralisation du modèle linéaire aux cas où les observations ne sont pas indépendantes. Cela signifie, entre autres choses, que la plupart des

avatars déjà connus du modèle linéaire peuvent, au moins en principe, servir à formuler un modèle à effets aléatoires. C'est ainsi que nous avons pu utiliser le modèle tobit à effets aléatoires pour étudier le problème particulier que pose le fait que le revenu de travail d'une personne n'est observable que lorsque cette personne travaille. On pourrait également employer la régression de Poisson à effets aléatoires pour analyser un phénomène dont la variable dépendante est une quantité discrète croissante, ou encore un modèle logit multinomial à effets aléatoires pour étudier, au fil du temps, les passages, réversibles ou non, entre les catégories d'une variable polytomique.

Le fait d'observer des individus à répétition et d'utiliser les observations comme autant d'unités dans un modèle linéaire crée le problème de corrélation entre les observations dont nous avons discuté plus haut. Nous avons montré comment les modèles à effets aléatoires permettaient de contourner ce problème. Observer les mêmes individus plusieurs fois n'est cependant pas la seule source de corrélation entre des observations. Utiliser les observations réalisées auprès de tous les membres d'un échantillon de ménages crée un problème analogue, tout comme le fait d'utiliser des observations réalisées auprès d'individus qui ont été sélectionnés aléatoirement mais dont on soupçonne qu'ils partagent des caractéristiques qui ne sont pas directement observées mais sont vraisemblablement reliées à leur appartenance commune à une unité observable, comme une école, un quartier ou une institution quelconque. Les situations de ce genre sont généralement traitées au moyen des modèles dits multiniveaux ou hiérarchiques (Goldstein, 1995; Bryk et Raudenbush, 1992) ou des modèles d'analyse spatiale qui sont, comme les modèles à effets aléatoires, des formes particulières de généralisation du modèle linéaire aux situations où les observations ne sont pas indépendantes. Les approches mathématiques qui sous-tendent ces méthodes sont différentes, mais chacune a pour but de permettre aux chercheurs d'obtenir les résultats les plus valables dans le contexte pour lequel elle a été développée. L'adaptation du modèle linéaire aux situations où le postulat d'indépendance des observations n'est pas réaliste est un domaine en développement. Il est raisonnable d'espérer que les démographes disposeront, au cours des prochaines années, de nouveaux outils pour l'étude des microdonnées longitudinales.

### RÉFÉRENCES BIBLIOGRAPHIQUES

- BLOSSFELD, H. P., A. HAMERLE et K. U. MAYER. 1989. *Event History Analysis*. Hillsdale, NJ, Lawrence Erlbaum Associates, Publishers.
- BRYK, A. S., et S. W. RAUDENBUSH. 1992. *Hierarchical Linear Models*. Newbury Park, CA, Sage.
- COX, D. R. 1972. « Regression models and life tables (with discussion) », *Journal of the Royal Statistical Society*, série B, 74 : 187-220.
- COURGEAU, D. L., et É. LELIÈVRE. 1989. *Analyse démographique des biographies*. Paris, Institut national d'études démographiques.
- DAVIS, J. A. 1985. *The Logic of Causal Order*. Beverly Hills, CA, Sage Publications.
- FITZMAURICE, G. M., N. M. LAIRD et A. G. ROTNITZKY. 1993. « Regression models for discrete longitudinal responses », *Statistical Science*, 8 : 284-309.
- FRISCH, R., et F. WAUGH. 1933. « Partial time regressions as compared with individual trends », *Econometrica*, 1 : 387-401.
- GOLDSTEIN, H. 1995. *Multilevel Statistical Models*. Londres et New York, Edward Arnold et Wiley, 2<sup>e</sup> édition.
- GREENE, W. H. 1997. *Econometric Analysis*. Upper Saddle River, NJ, Prentice Hall, 3<sup>e</sup> édition.
- HAMERLE, A., et G. RONNING. 1995. « Panel analysis for qualitative variables », dans G. ARMINGER, C. C. CLOGG et M. E. SOBEL, éd. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York, Plenum Press : 401-451.
- HSIAO, C. 1986. *Analysis of Panel Data*. New York, Cambridge University Press.
- KING, G. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York, Cambridge University Press.
- KISH, L., et M. R. FRANKEL. 1974. « Inference from complex samples », *Journal of the Royal Statistical Society*, série B, 36 : 1-37.
- LAPLANTE, B., B.-P. HÉBERT et P.-M. MENGER (à paraître). « Actors' careers: Individual and collective actions in a time of changing labour markets ».
- LIANG, K. Y., et S. L. ZEGER. 1986. « Longitudinal data analysis using generalized linear models », *Biometrika*, 73 : 13-22.
- MCCULLAGH, P., et J. A. NELDER. 1989. *Generalized Linear Models*. Londres, Chapman and Hall, 2<sup>e</sup> édition.
- NEUHAUS, J. M., J. D. KALBFLEISCH et W. W. HAUCK. 1991. « A comparison of cluster-specific and population-averages approaches for analyzing correlated binary data », *International Statistical Review*, 59 : 25-35.

- PENDERGAST, J. F., S. J. GANGE, M. A. NEWTON, M. J. LINDSTROM, M. PALTA et M. R. FISHER. 1996. « A survey of methods for analyzing clustered binary response data », *International Statistical Review*, 64 : 89-118.
- STATA CORP. 1999. *Stata Statistical Software: Release 6.0*. College Station TX : Stata Corporation.
- TOBIN, J. 1958. « Estimation of relationships for limited dependent variables », *Econometrica*, 26 : 24-36.
- WU, Z. 1999. « Premarital cohabitation and the timing of first marriage », *Canadian Review of Sociology and Anthropology*, 36 : 109-127.

## ABSTRACT

Benoît LAPLANTE and Benoît-Paul HÉBERT

### **EVENT HISTORY ANALYSIS USING RANDOM-EFFECTS LINEAR MODELS. THE CASE OF THE CAREERS OF PROFESSIONAL ACTORS**

*Demographers today often make use of longitudinal microdata. To analyze these data, they generally use a set of techniques and models associated with what is termed event history analysis. The linear models used in this context fall into the category of proportional hazards models, the most often used of which is the semiparametric proportional hazards model. Despite their advantages, these models only enable researchers to study processes which can be easily described and measured, such as sequences of states or changes of states. The authors show how two random-effects linear models—the random-effects logit model and the random-effects tobit model—can be used to study longitudinal phenomena such as changes over time in the categorization into either of the two categories of a binary dependent variable, and changes over time in a continuous variable where a portion of the distribution of this variable is unobserved. This approach is used to study the career dynamics of professional actors in France.*