CEPED

EHESS-INED-INSEE-ORSTOM-Université Paris VI

L'ANALYSE DES ENQUÊTES BIOGRAPHIQUES

à l'aide du logiciel STATA

Philippe BOCQUIER

Paris, juillet 1996

Documents et Manuels du CEPED n° 4

L'analyse des enquêtes biographiques

Philippe Bocquier
Documents et Manuels du CEPED N°4, Paris, juillet 1996

Note à l'attention des utilisateurs de Stata 5.0

Le manuel a été conçu pour Stata 4.0 et certains changements sont intervenus dans la nouvelle version 5.0 pour l'analyse des biographies. Les commandes **firstocc** et **censor**, disponibles sur disquette avec le manuel du CEPED, peuvent toujours être utilisées pour créer les indices de troncatures. Ensuite, la nouvelle version de Stata 5.0 exige de définir au préalable (avant l'exécution des commandes d'analyse descriptive ou de régression) les paramètres de temps, d'indice de troncature, etc., à l'aide de la commande **stset**. Mais la commande **stset** n'autorise pas de valeur manquante (*missing value*) pour l'indice de troncature. Il faut donc supprimer les observations pour lesquels l'indice de troncature est codé manquant (c'est-à-dire quand l'événement a déjà eu lieu, selon notre logique). Un fichier spécifique doit donc être créé pour chaque événement étudié.

Par exemple, dans les chapitres 5 et 6 du manuel du CEPED, au lieu des commandes **kapmeier**, **ltable**, **survsum**, **logrank** (qui n'existent plus) on exécutera :

```
. drop if job1==.
. save nomfichier                          (pour créer un nouveau fichier)
. stset Tjob1 job1, id(no)                 (pour fixer les paramètres)
. sts graph, by(strate)                    (pour les courbes de Kaplan-Meier)
. sts list, by(strate)                     (pour les tables de séjour)
. stsum, by(strate)                        (pour le résumé : médiane...)
. sts test strate, logrank                 (pour les tests)
```

Pour le modèle de Cox, on peut exécuter la commande :

```
. stcox gener2 gener3 ...
```

Mais si l'on veut utiliser les pondérations avec **aweight** (*analytic weights*), on peut exécuter :

```
. cox Tjob1 gener2 gener3 ... [aw=poids], dead(job1) tvid(no)
```

L'auteur,
Juin 1997

L'ANALYSE DES ENQUÊTES BIOGRAPHIQUES
à l'aide du logiciel STATA

Déjà parus dans la collection "Documents et Manuels du CEPED" :

- n° 1 : *La démographie de 30 États d'Afrique et de l'Océan Indien*, Valérie Guérin (éd.) (1994), 352 p.

- n° 2 : *Clins d'œil de démographes à l'Afrique et à Michel François*, Jacques Vallin (éd.) (1995), 244 p.

- n° 3 : *Manuel de sondages. Applications aux pays en développement*, Rémy Clairin et Philippe Brion (1996), 104 p.

Directeur de la publication : Jacques Vallin
Responsable scientifique : Christophe Lefranc
Composition et mise en page : Valérie Guérin-Mary et Sabine Joao

Couverture : Baga, génie de l'eau - collection particulière

Le CEPED, *Centre français sur la population et le développement*, est un « Groupement d'intérêt scientifique » (GIS) créé en 1988 par l'INED, l'INSEE, l'ORSTOM, l'Université Pierre et Marie Curie et l'École des hautes études en sciences sociales, pour conjuguer leurs efforts en matière de recherche, de formation et de coopération avec les pays du Sud dans le domaine de la population et de ses relations avec le développement. Ses activités de recherche portent essentiellement sur les facteurs de la dynamique des populations (santé, famille, fécondité, migrations), leurs relations avec les divers aspects du développement économique et social (éducation, emploi, activité économique, structures sociales, ...) ainsi que les méthodes d'observation et d'analyse appropriées. Ses travaux sont définis et conduits en étroite relation avec les organismes partenaires du tiers monde (offices statistiques, centres de recherche, universités). Le CEPED accueille régulièrement à Paris des chercheurs de ces pays, et met à la disposition du public un important centre de documentation sur les thèmes de sa compétence. Pour toutes ces tâches, le CEPED reçoit un large concours du Ministère de la coopération et du développement.

DOCUMENTS ET MANUELS DU CEPED N° 4

Philippe BOCQUIER

L'ANALYSE DES ENQUÊTES BIOGRAPHIQUES

à l'aide du logiciel STATA

Préface de Daniel Courgeau et Éva Lelièvre

Centre français sur la population et le développement
(EHESS-INED-INSEE-ORSTOM-Université Paris VI)

Juillet 1996

Éléments de catalogage :

L'analyse des enquêtes biographiques à l'aide du logiciel STATA, Philippe Bocquier. – Paris, Centre français sur la population et le développement, 1996, 208 p ; 24 cm.

ISBN : 2-87762-093-X ISSN : 1264-2487
Centre français sur la population et le développement
15, rue de l'école de médecine - 75270 PARIS Cedex 06 - FRANCE
Téléphone : (33) (1) 44 41 82 30 - Télécopie : (33) (1) 44 41 82 31

SOMMAIRE

PRÉFACE

Née de la conjonction du calcul actuariel et de l'analyse de régression, l'analyse des biographies a acquis depuis le début des années 1980 un rôle croissant dans le champ des études démographiques.

Le calcul actuariel, qui a connu ses débuts au milieu du XVII^{e} siècle avec les "Observations naturelles et politiques sur des bulletins de mortalité de la ville de Londres", de John Graunt, s'est développé après la seconde guerre mondiale pour devenir l'analyse longitudinale classique, appliquée à tous les phénomènes démographiques. Cette approche permet l'analyse au cours du temps de chacun de ces phénomènes, en éliminant l'effet perturbateur des autres. Elle fournit essentiellement la distribution de l'arrivée du phénomène étudié selon l'âge ou la durée écoulée depuis un événement initial.

L'analyse de régression est née au début du XIX^{e} siècle avec les travaux de Carl-Friedrich Gauss et Adrien-Marie Legendre, qui ont appliqué ces méthodes à des données astronomiques. Ce n'est que plus tard que les données humaines, en particulier biologiques, ont été analysées à l'aide de ces méthodes. Elles permettent de relier la mesure d'un caractère à un moment donné à diverses autres caractéristiques du même individu mesurées simultanément. Appliquées au départ à des mesures continues, telle que la taille d'un individu (régression linéaire), elles ont été généralisées par la suite à des mesures discrètes, telles que le fait qu'un individu soit marié ou non (régression logistique).

Si l'analyse longitudinale classique a pris une place de choix dans les méthodes d'analyse démographique, car elle faisait intervenir la dimension temporelle indispensable pour toute étude des populations, les méthodes de régression, s'appliquant à des données atemporelles, y ont joué longtemps un rôle très restreint. Ce n'est qu'en 1972, que David R. Cox a fait la synthèse des deux approches dans son article "Regression models and life tables" paru dans le Journal of the Royal Statistical Society. Ces nouvelles méthodes appliquées au départ à des données de statistique médicale, portant sur de très faibles échantillons, nécessitaient pour avoir une place de choix en démographie des enquêtes biographiques plus détaillées et portant sur des échantillons plus nombreux. Ce n'est qu'au cours des années 1980, que se multiplient des enquêtes de ce type, tant dans les pays développés que dans les pays en développement. Dans le même temps, les méthodes d'analyse se sont affinées pour permettre l'analyse des

événements non plus séparément les uns des autres, mais en interaction les uns avec les autres.

Ainsi se sont mises en place les méthodes d'analyse biographiques qui ont trouvé des applications dans toutes les sciences humaines : démographie, sociologie, économie, sciences naturelles, géographie humaine, statistique médicale, etc. Elles généralisent les méthodes classiques d'analyse longitudinale en permettant de considérer l'hétérogénéité des populations et l'interaction entre différents phénomènes.

Pour diffuser la connaissance de ces méthodes, nous avons publié en 1989, un manuel sur L'analyse démographique des biographies que le présent ouvrage vient parfaitement compléter en fournissant un exemple d'utilisation pratique très détaillé pour les utilisateurs de la micro-informatique, des méthodes que nous y avions présentées. Il faut bien voir que ces méthodes d'analyse ont pu se développer grâce à la puissance croissante des ordinateurs, capables de réaliser en quelques secondes des calculs d'une complexité telle qu'il faudrait des mois, voire des années pour les réaliser à la main, avec sans aucun doute, de nombreux risques d'erreurs. Un nombre croissant de logiciels sont à même maintenant de réaliser de telles analyses, mais la complexité des méthodes et des hypothèses à faire pour mener à bien une telle étude, nécessite une présentation détaillée avec de nombreux exemples d'utilisation de ces logiciels.

C'est ce que Philippe Bocquier réalise parfaitement ici. Après avoir posé les bases conceptuelles d'une telle démarche, il applique aussi bien les méthodes de régression linéaire multiple et les méthodes de régression logistique, que les méthodes d'analyse des biographies à un échantillon issu d'une enquête réalisée à Dakar par l'IFAN et l'ORSTOM sur l'insertion urbaine. Pour l'unité de son propos, il utilise ici les diverses commandes du logiciel STATA qu'il applique de façon très détaillée et très précise au traitement des données de son enquête. Il indique les pièges à éviter, commente les résultats et donne aux lecteurs un certain nombre d'exercices astucieux qu'ils peuvent réaliser grâce aux fichiers de données sur la disquette qui accompagne cet ouvrage.

Ce manuel ouvre au chercheur en sciences sociales l'accès à des techniques d'analyse complexes, par une présentation très claire et très didactique. Il devrait lui permettre d'exploiter lui-même les fichiers d'enquêtes biographiques qu'il a réalisées, d'en faire une analyse très complète et de présenter clairement les résultats qu'il a obtenus auprès de la communauté scientifique.

Daniel Courgeau
Éva Lelièvre
INED

AVANT-PROPOS

En 1989, des chercheurs de l'Institut fondamental d'Afrique noire (IFAN) et de l'Institut français de recherche scientifique pour le développement en coopération (ORSTOM) ont conduit sous la direction de Philippe Antoine une étude de l'insertion urbaine à Dakar. Lors de l'analyse des données de Dakar, j'ai utilisé plusieurs logiciels, pour retenir en définitive STATA, et j'ai mis au point de nombreuses procédures. En 1992, nos collègues du Centre d'études et de recherche sur la population pour le développement (CERPOD) et du Département de démographie de l'Université de Montréal ont réalisé, sous la direction de Dieudonné Ouedraogo et Victor Piché, une étude similaire à Bamako.

Ces deux enquêtes s'appuient sur un questionnaire biographique. Cependant peu de chercheurs de ces deux équipes maîtrisaient l'analyse de ce type de données. Très vite le besoin s'est fait sentir de rendre ces techniques plus accessibles, en particulier aux chercheurs non quantitativistes, pour que puisse s'instaurer un dialogue plus riche et une meilleure collaboration entre chercheurs. Avec l'appui financier du Réseau démographie de l'AUPELF-UREF, du CEPED, de l'ACDI, de l'ORSTOM, un stage de formation a été organisé en mai 1993 au CERPOD à Bamako. À la demande de mes collègues, j'ai repris et développé mes notes de cours pour écrire ce manuel.

Je tiens à remercier Daniel Courgeau et Éva Lelièvre, à qui revient le mérite de la diffusion de l'analyse des biographies dans le milieu de la recherche francophone. Ce manuel a beaucoup bénéficié de leur relecture attentive. Je remercie aussi Lucie Gingras pour ses avis critiques sur la première version de ce manuel.

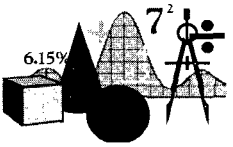
Mes remerciements vont aussi à Victor Piché, grâce à qui j'ai pu en de multiples occasions être invité au Département de Démographie de l'Université de Montréal, et à Dieudonné Ouedraogo, qui a tout fait pour faciliter mon insertion de plusieurs années au CERPOD. Je remercie enfin toute l'équipe du CEPED pour la qualité et la rapidité de son travail d'édition.

Je remercie enfin, pour leurs encouragements constants, Philippe Antoine, Benoît Laplante et Richard Marcoux, et tous mes collègues du CERPOD qui ont eu à subir mes cours.

Philippe Bocquier

Pour que le lecteur se repére plus facilement dans la lecture de ce manuel, nous avons inséré dans le texte plusieurs icônes auxquelles sont associées les significations suivantes :

Référence bibliographique importante, utile pour l'approfondissement des connaissances.



Formule statistique importante, nécessaire pour la bonne compréhension du modèle, notamment pour transmettre les résultats de la recherche dans une revue scientifique.

Référence à un logiciel ou à un programme contenu dans la disquette accompagnant ce manuel.

En outre, les lignes de programmes et les résultats produits par STATA figurent en encadré et en caractères courrier. Les encadrés en grisé sont réservés aux exercices qui figurent généralement en fin de chapitre. Dans le corps du texte, les noms des fichiers sont écrits en majuscules (par exemple JOB.DTA), et les noms des variables des fichiers de données sont entourés de guillemets de type " " (par exemple "strate"). Les guillemets de type « » sont réservés aux noms de commandes de STATA (par exemple « cox ») ou aux options de ces commandes (« slope »). Lorsqu'une commande ou une option nécessite un argument que l'utilisateur doit choisir, cet argument est écrit en italiques (par exemple « logit *liste_variables* » ou bien « dead(*variable*) »).

INTRODUCTION

L'analyse quantitative des biographies est réputée d'accès difficile. La difficulté tiendrait autant au recueil des données qu'aux techniques pour les analyser. L'analyse des biographies apparaît, aux yeux de beaucoup, plutôt comme un luxe qui n'intéresse qu'un petit nombre de chercheurs.

Pourtant, les techniques d'enquête rétrospective ont évolué et ont montré leur efficacité. En comparaison de leur coût, ce type d'enquêtes dégage une grande plus-value : il permet de recueillir en un seul passage des informations très riches. Les biais dus à la collecte d'informations rétrospectives (problèmes de mémoire en particulier) ont souvent été et continuent de faire l'objet des principales critiques à ce type d'enquêtes. Pourtant, les méthodes d'analyse actuelles sont peu sensibles aux erreurs de datation pourvu que la succession des événements soit bien établie. En annexe 3 de ce manuel figure d'ailleurs une méthode simple de recueil et de contrôle qui permet de s'assurer de la cohérence des dates. De ce point de vue, le recueil rétrospectif n'est plus aussi handicapant.

Ce manuel montre que la richesse de l'analyse des biographies réside en la prise en compte du temps dans l'explication causale. Avec les outils conceptuels et les programmes fournis dans ce manuel, le lecteur devrait arriver à une quasi-autonomie sur un matériel informatique courant.

1. Une technique au service des sciences sociales

Notre souhait est de rendre accessible aux chercheurs en sciences sociales de disciplines diverses, des techniques développées en démographie et qui peuvent être utilisées dès lors que les données font intervenir la dimension temporelle. Ce manuel s'adresse donc aussi bien à l'économiste, au sociologue, à l'anthropologue, au psychologue, au géographe, à l'historien... qu'au démographe. Tous ces

spécialistes (et d'autres encore) font intervenir le temps à un moment ou à un autre de leurs recherches.

Ce manuel ne traite pas de statistique mathématique. D'autres ouvrages existent qui ont pour but de fournir à l'étudiant et au chercheur les principes théoriques utiles pour le développement de techniques nouvelles : il s'agit par exemple de l'ouvrage de Daniel Courgeau et Éva Lelièvre "*Analyse démographique des biographies*", publié en 1989, le seul rédigé en français à ce jour, ainsi que de nombreux ouvrages en anglais, dont nous citerons pour mémoire John Kalbfleisch et Ross Prentice (1980), James Heckman et Burton Singer (1985), Karl-Ulrich Mayer et Nancy Tuma (1990), Perkragh Andersen, Lornulf Borgan, Richard Gill et Neil Keiding (1993). Notre objectif n'est pas de venir en bout de cette liste : ces ouvrages restent indispensables pour le statisticien. Mais il nous a semblé qu'il manquait un manuel, pratique pour l'analyse des biographies, notamment pour les utilisateurs de la micro-informatique, de plus en plus nombreux aujourd'hui.

Dans ce manuel, il ne sera pas fait de démonstration, et d'ailleurs nous n'aurons recours qu'à des formules très simples, seulement nécessaires pour permettre au lecteur de faire une présentation de ses résultats auprès de la communauté scientifique. Notre but n'est pas de former des statisticiens mais de contribuer à la recherche en sciences sociales en permettant à tout un chacun d'exploiter lui-même ses fichiers de données et d'en faire l'analyse de bout en bout. Mais nous serions déjà satisfaits si à l'issue de la lecture de ce manuel, un membre d'une discipline pouvait discuter avec un membre d'une autre discipline pour traiter d'un même problème relatif à des événements observés dans le temps.

C'est pourquoi, nous insisterons particulièrement sur le travail conceptuel qui précède les analyses. Certes, la technique d'analyse est statistique, mais ce n'est pas l'essentiel du travail. La difficulté réside bien plus souvent en amont qu'en aval de l'analyse statistique.

2. Vers la maîtrise de la causalité

Nous proposons d'abord au lecteur un travail conceptuel en lui soumettant une réflexion sur la temporalité et la causalité. Cette réflexion, qui s'applique à toutes les sciences sociales, est un préalable indispensable pour bien comprendre les mérites et les limites de l'analyse statistique.

La technique actuellement la plus appropriée pour analyser les données biographiques est sans doute le modèle semi-paramétrique mis au point par

David R. Cox en 1972. Un des buts spécifiques de ce manuel est donc de faire comprendre ce modèle et ses dérivés. Pour cela, il est indispensable de faire un détour par la régression en statistique et par les tables de survie en démographie.

La logique de l'analyse multivariée est expliquée à l'aide des régressions statistiques simples. Cet exposé montre pourquoi le modèle de régression linéaire est préférable à l'analyse descriptive (tableaux simples ou croisés), et quelles relations entre variables explicatives ce modèle permet de mettre en valeur dans les données.

La régression logistique est une extension de la régression linéaire qui permet de traiter des variables qualitatives (dichotomiques ou polytomiques) plutôt que des variables quantitatives (continue). Ce type de modèle nous rapproche de la problématique démographique, dès lors qu'il est question de "population soumise au risque de connaître un événement". Des techniques de diagnostic ont été développées qui permettent de juger l'adéquation entre le modèle et les données.

La démographie ajoute un élément à l'analyse statistique classique : la dimension temporelle. Cette prise en compte du temps nécessite des techniques d'analyse originales, qui permettent de traiter des données d'enquêtes dites biographiques. C'est à ce moment-là que commencent véritablement les problèmes de conceptualisation, qui se poseront jusqu'aux modélisations les plus complexes.

Les principes de la régression (notamment logistique) et des tables de survie sont combinés pour produire le fameux modèle semi-paramétrique, dit de Cox, qui fait parti des modèles "à risques proportionnels". Ce modèle est d'une incomparable richesse, qui en fait actuellement l'enfant chéri de la statistique démographique. C'est probablement avec lui que l'on maîtrise le mieux la dimension temporelle, et, de fait, l'analyse causale.

En dernier lieu, nous aborderons des questions plus générales relatives au recueil des données et à la notion même d'événement : il s'agira de s'interroger sur les mérites et limites de l'analyse des biographies.

CHAPITRE 1

ANTÉRIORITÉ, CAUSALITÉ ET CORRÉLATION

Ce chapitre doit beaucoup à l'ouvrage de Guillaume Wunsch, "*Causal Theory and Causal Modeling*" (1988). Malgré l'abondance de la littérature statistique, il n'existe que très peu de références traitant de la théorie de la causalité et de la modélisation statistique. Cet ouvrage est d'autant plus estimable qu'il concerne les sciences sociales. Par ailleurs, dans le cadre de la chaire Quételet, un colloque a été organisé en 1987 sur "*L'explication en sciences sociales - La recherche des causes en démographie*". Ses actes (Duchêne *et al.*, 1987) demeurent jusqu'à présent une référence incontournable dans le domaine. À l'aide de ces deux ouvrages, nous tenterons de préciser les concepts qui sont à la base de l'analyse des biographies. Nous justifierons d'abord l'importance du temps dans l'analyse causale en sciences sociales, pour ensuite définir l'événement et les différents types de causalité dans le temps. Enfin, le lien entre causalité et probabilité sera expliqué pour que le lecteur perçoive les limites conceptuelles de la démarche statistique qui lui sera proposée dans les chapitres suivants.

La réflexion sur la causalité est particulièrement importante parce qu'**il n'est guère possible en sciences sociales de contrôler l'analyse à un autre stade que celui de la définition conceptuelle**. En effet, bien des raisonnements se justifient par la répétition (confirmation ou infirmation des interprétations précédentes) plutôt que par l'expérimentation, qui est quasiment impossible en sciences sociales. Or, répéter ne justifie rien, car on peut répéter ses erreurs ou celles des autres. De plus, aucune source de données n'épuisera les interprétations possibles : différents résultats peuvent provenir du même corpus de données.

Il est impossible de départager deux raisonnements scientifiques sur la base des seules conclusions d'analyse : **sans la théorie, les faits ne sont pas**

interprétables, quel que soit d'ailleurs l'instrument d'analyse[1]. C'est au contraire dans les prémisses du raisonnement qu'il faut aller chercher les arguments contradictoires. Ce n'est pas facile, car les auteurs ne font généralement pas l'effort de rationaliser leur propre raisonnement. Les hypothèses sont rarement posées, et le modèle théorique qui met en rapport ces différentes hypothèses pour former un tout cohérent (une chaîne causale) est encore moins souvent explicité. La théorie est au mieux simplement rejetée dans une revue de littérature (qui fait se perdre le lecteur dans un dédale bibliographique), et au pire elle est supposée faire partie du bagage scientifique du lecteur. Certes, la bonne perception d'un phénomène est l'aboutissement d'un processus cumulatif d'expériences passées, dont la communauté scientifique est garante, à travers différentes procédures de contrôle, mais ces expériences doivent être d'abord rationalisées et validées par une armature théorique.

Pour reprendre après Guillaume Wunsch (1988) l'idée proposée par Paul Blalock en 1982, la recherche consiste à passer d'une théorie générale (concepts théoriques) à une théorie auxiliaire (spécifique ou partielle), et ensuite éventuellement à un modèle statistique causal (voir aussi Gérard *in* : Duchêne *et al.,* 1987). L'effort fourni dans la première étape (passage d'une théorie générale à une théorie auxiliaire) est beaucoup plus grand que celui fourni dans la deuxième étape de formalisation statistique. Quand au calcul statistique, il est rendu de plus en plus accessible par la programmation informatique. C'est pourquoi sans doute les auteurs font rarement l'effort de théorisation préférant passer tout de suite aux aspects techniques.

1. Temps et événement

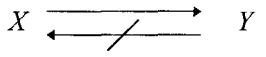
a) L'antériorité implicite ou la priorité temporelle de la cause sur l'effet

Qu'est-ce qu'une relation causale ? Pour qu'il y ait causalité, il faut nécessairement une cause (appelée conventionnellement *X*) et un effet (*Y*), la relation étant exprimée le plus simplement comme suit :

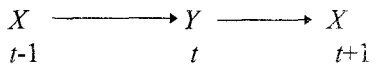
$$X \longrightarrow Y$$

1 L'analyse démographique n'a malheureusement pas le privilège de l'insuffisance théorique : les statistiques s'y substituent souvent au raisonnement scientifique mais le manque de théorisation se rencontre aussi dans les autres sciences sociales.

Si cette relation est vérifiée, alors la relation inverse, *Y* est la cause de *X*, ne peut l'être simultanément :

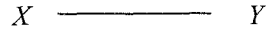


On appelle cette propriété l'**asymétrie causale**, ce qui signifie simplement que **si *X* cause *Y*, alors *Y* ne peut simultanément causer *X***. Dans cette formulation, le terme "simultanément" est en fait essentiel : c'est parce que le processus se déroule dans le temps que la relation causale est asymétrique. Si l'on veut forcer une relation à être symétrique, on est obligé d'abandonner la **priorité temporelle de la cause sur l'effet**, ce qui n'est guère envisageable dans le monde réel. Dire que *X* cause *Y*, c'est aussi dire que *X* est différent de *Y* et que *X* est antérieur à *Y*, c'est-à-dire que la cause *X* précède d'au moins un temps infinitésimal l'effet *Y*. Dès lors, *Y* peut à son tour causer *X* seulement dans un temps postérieur :

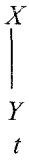


On voit que **le temps est à la base de la perception que l'on a de la causalité**. Pour percevoir une relation, il est nécessaire de faire évoluer cause et effet dans le temps. C'est dès la conceptualisation que l'on doit se poser la question du temps.

Une relation qui n'est pas définie dans le temps et qui est malgré tout vraie, est alors autre chose qu'une relation causale : c'est le cas par exemple lorsque *X* et *Y* sont simultanés et liés par une **dépendance fonctionnelle**. On aura alors la relation simultanée :



... que nous préférons noter sous une forme verticale, pour bien signifier que dans ce cas, *X* et *Y* évoluent dans le même temps[2] :



[2] Cette présentation n'est claire que si le lecteur a intégré la représentation conventionnelle du temps : une ligne tracée de gauche à droite. Cette représentation a un caractère conventionnel, qui tient sans doute au mode d'écriture occidental (de gauche à droite). Dans d'autres sociétés, notamment sans écriture, il est possible que la représentation du temps ne soit pas horizontale : elle peut être circulaire (sur le modèle du rythme des saisons ou du soleil autour de la terre) ou bien verticale (accumulation par couches successives, ce que comprendront aisément les archéologues et les pédologues).

La cause de X ou de Y, sur ce schéma, ne peut se situer qu'à gauche de t, en t-1 par exemple. Dans le cas d'une telle dépendance fonctionnelle, montrer la relation entre X et Y n'explique ni X ni Y, s'il n'est pas fait mention d'une cause antérieure.

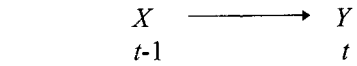
On notera au passage que X et Y sont considérés comme événements dans la relation causale : dès lors qu'on considère le temps, l'objet étudié ne peut être qu'un événement. Les deux notions de temps et d'événement sont indissociablement liées.

b) Événement et condition permanente comme causes

Pour que l'événement X soit considéré comme cause de Y, il faut nécessairement que X et Y soient tous deux définis dans le temps comme événements successifs. Cependant, un événement X est rarement la seule cause d'un événement postérieur Y. On identifie souvent un ensemble de causes, et non une cause unique, pour expliquer un événement. Parfois, les causes ne sont pas suffisamment bien définies ou forment un ensemble trop complexe pour être identifiées séparément.

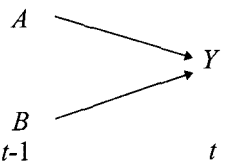
Les relations causales peuvent donc prendre différentes formes (nous reprenons la typologie des causes proposée par Wunsch, 1988) :

Cause unique :



Chaîne causale :

Pluralité causale disjonctive :



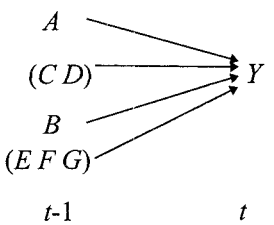
Pluralité causale conjonctive :

• de type INUS :

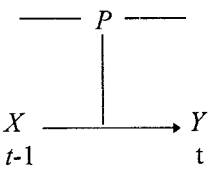
• de type interaction :

La pluralité causale conjonctive de type INUS (de l'anglais *Insufficient but necessary part of a condition which is itself unnecessary but sufficient*) signifie que la cause *A* n'est pas suffisante pour provoquer *Y* mais qu'elle est une partie nécessaire d'une condition (que *A* agisse ensemble avec *B*) qui est suffisante (sans être nécessaire) pour provoquer *Y*. La pluralité causale conjonctive de type interaction signifie que non seulement *A* et *B* agissent pour provoquer *Y* mais que s'y ajoute un effet combiné (ou interaction) des deux causes.

Bien entendu, plusieurs relations causales peuvent être présentes à la fois :



Ces relations causales supposent que chaque événement (cause ou effet) est bien défini. Or, on peut être confronté à certaines indéterminations. Par exemple, un événement *X* peut être clairement identifiable, mais n'être la cause de *Y* que dans un certain contexte, qui n'est pas nécessairement un autre événement : on parle alors de **condition permanente** P pour la réalisation d'un événement *Y* :

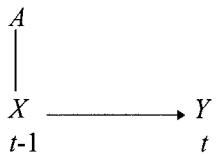


Cette condition permanente permet d'introduire la notion de relation causale sans événement explicatif, ou si l'on veut de **quasi-relation causale** : sachant le contexte *P*, on peut connaître *Y* dans la mesure où *P* est associé à un événement inconnu *X* ou à une chaîne d'événements complexe (*X*1, *X*2, *X*3, etc.). Ce qui est réellement évalué, par manque d'information sur *X*, est la relation suivante,

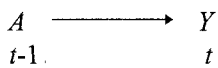
Dans ce cas, il n'est pas nécessaire de faire référence au temps. Mais c'est seulement à défaut de connaître *X* que l'on a recours à ce type de relation.

Par ailleurs, un événement A peut être connu et utilisé en raison de sa dépendance fonctionnelle avec un événement X mal identifié (ou peu mesurable).

On a alors la relation causale,



mais l'analyse portera, par simplification, sur la relation observable :



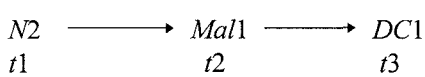
c) Association et indétermination temporelle

À l'heure actuelle, le principe de priorité temporelle de la cause sur l'effet n'est guère remis en cause. Cela semble un principe épistémologique adopté par l'ensemble de la communauté scientifique, quelle que soit la discipline. Cela ne signifie pas pour autant que toutes les analyses se conforment maintenant à ce principe fondamental. Sans même que parfois les auteurs en aient conscience, le temps est mal défini, de même que l'antériorité des causes, les dépendances fonctionnelles, les conditions permanentes, etc.

Prenons deux exemples de confusions dans la définition des relations causales : le cas où un événement est improprement considéré comme la cause d'un autre, alors que c'est un troisième événement qui est la cause probable ; et le cas où le temps n'a pas été défini convenablement menant ainsi à une interprétation fausse de la causalité.

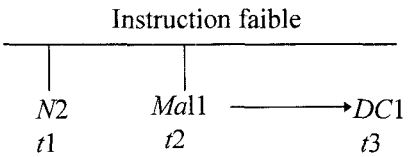
Premier exemple : survie de l'enfant et reprise de la fécondité de la mère

Considérons l'enchaînement suivant : un nouvel enfant naît, l'enfant précédent tombe malade, puis meurt. On aura le schéma :



avec $N2$: naissance du second enfant
$Mal1$: maladie du premier enfant
$DC1$: décès du premier enfant

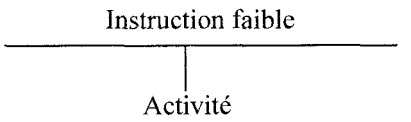
Or, l'antériorité de la reprise de la fécondité (*N*2) par rapport à la maladie de l'enfant (*Mal*1), qui donne l'impression que l'arrivée d'un nouvel enfant entraîne la maladie du premier enfant, peut simplement être due à un facteur commun ayant un effet sur les deux phénomènes. Prenons l'effet de l'instruction de la mère comme condition permanente :



La maladie du premier enfant ne serait alors pas due à la naissance du suivant, mais à la faible éducation de la mère.

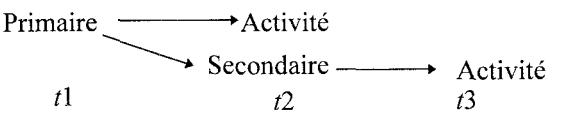
Deuxième exemple : l'effet de l'éducation sur l'emploi des jeunes femmes

Lors d'une enquête, on a déterminé une relation positive entre l'exercice d'une activité et le bas niveau d'instruction chez les jeunes femmes :



Or on pourrait penser le contraire, les femmes font des études pour obtenir un emploi et non pas pour rester inactives. En fait, on sait que les jeunes femmes instruites mettent aussi plus longtemps à trouver un emploi, non seulement en raison de la durée des études, mais aussi parce que la période de recherche d'emploi est plus longue chez les plus instruites.

On aura donc :



Selon l'âge où l'on étudiera l'activité de ces femmes, les conclusions seront très différentes : au temps *t*2, les événements "activité" et "poursuite des études dans le secondaire" sont en concurrence.

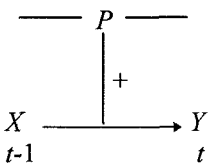
L'indétermination du temps d'observation est à l'origine de bien des erreurs d'interprétation. Dès lors que l'on étudie un événement, on devrait toujours utiliser un moyen d'observation qui permette de tenir compte du temps, afin de bien définir

les causalités possibles. Mais la difficulté est avant tout conceptuelle : aucun outil de mesure ne permet de se passer d'une réflexion sur le processus causal, où le temps tient une place fondamentale.

2. L'univers probabiliste

Dans la première partie de ce chapitre, il n'a pas encore été question de probabilité. Les relations entre événements ont été posées de manière déterministe, alors qu'il apparaît évident, surtout dans les deux derniers exemples, que la relation entre une cause et son effet n'est que probable. Toute nouvelle naissance n'entraîne pas le décès de l'enfant qui la précède, et toutes les femmes n'accèdent pas à l'emploi. On pourrait schématiser cette notion probabiliste ainsi :

Cela signifie qu'un événement X augmente la probabilité que Y survienne. De même, on aura la relation suivante où la condition permanente P augmente la probabilité que X cause Y :



En fait, l'événement Y peut avoir plusieurs causes. En d'autres termes, la relation entre X et Y, ou entre P et Y, n'est pas considérée comme fortement déterministe : X et P ne sont que suffisants. L'événement Y peut être causé par autre chose, ou ne pas survenir du tout. Ce type de lien de causalité est ce qu'on appelle un **déterminisme faible**.

Comme on le voit, penser en termes probabilistes ne signifie pas rejeter le déterminisme, mais le rendre plus complexe. D'ailleurs, si les résultats d'une expérimentation ne suivent pas exactement le modèle, dans un contexte déterministe, même faible, on peut invoquer le "bruit blanc" (*white noise*) dû aux instruments de mesure qui ne nous permettent pas de mesurer avec exactitude un phénomène.

Mais le bruit n'est pas toujours "blanc". Beaucoup plus fréquente dans le domaine des sciences sociales, est l'existence de l'indéterminisme fondamental et de l'indéterminisme conceptuel.

a) Indéterminisme conceptuel et indéterminisme fondamental

Même si le phénomène étudié est déterministe par nature, notre ignorance de toutes les causes qui conduisent à ce phénomène nous interdit le plus souvent une prévision parfaite. C'est particulièrement le cas en sciences sociales où on n'est jamais sûr de détenir toutes les variables pertinentes : on en oublie, on les recueille mal, et certaines sont même impossibles à saisir ou tout simplement inconnues, cachées. Par exemple, **les chercheurs font un usage fréquent des variables qualifiant des conditions permanentes**, telles que le niveau d'instruction, l'origine sociale, le contexte économique, etc., **pour éviter de décrire un système de causalités trop complexe, un enchaînement d'événements sans fin**. Ces inévitables simplifications, que l'on peut appeler comme Guillaume Wunsch (1988) des "**causalités floues**" (*fuzzy causalities*) aboutissent malheureusement à du "temps flou" (*fuzzy time*) : elle conduisent à "oublier" que toute action, tout phénomène, s'inscrit dans le temps.

Réduire l'indéterminisme conceptuel, dû aux insuffisances des chercheurs eux-mêmes et de leurs théories, **est bien le propre de toute démarche scientifique**. Le recueil des meilleures variables explicatives, la mise au point du meilleur système de causalité, sont des objectifs louables, qui prennent peut-être leur origine dans le besoin qu'a l'Homme de rejeter au plus loin les limites de l'inconnu. **Mais on ne peut pas échapper à une autre forme d'indéterminisme, ontologique cette fois-ci**, malgré tous nos efforts. Bien des phénomènes sont aléatoires par nature. C'est particulièrement le cas pour l'Homme qui doit choisir, sur la base de ses expériences passées et des informations dont il dispose, l'action à entreprendre. Même en tenant compte de la faible qualité de l'information disponible ou de l'irrationalité de certains comportements, il n'en reste pas moins que **la plupart des choix se font dans l'incertitude**. Qui peut prédire si la pièce envoyée en l'air tombera du côté pile ou du côté face ? Or, bien des actions humaines dépendent de ce genre de résultat : il faut y reconnaître une forme fondamentale d'indéterminisme.

Comme on l'a dit plus haut, le caractère déterministe de certains phénomènes n'est pas contradictoire avec l'approche probabiliste. En outre, cette approche est rendue nécessaire pour analyser la complexité des actions humaines, qui rend à la fois difficiles le recueil des informations et la conceptualisation. Mais c'est l'indéterminisme ontologique propre aux actions humaines qui nous fait adopter définitivement une approche probabiliste.

b) Déduction et induction

Le propre de la démarche scientifique est de vérifier des théories en les confrontant à des faits. Une théorie n'est pas vérifiable si elle ne peut être confrontée à des faits, et inversement la réalisation d'un fait particulier n'a d'intérêt scientifique que dans la mesure où ce fait infirme ou confirme une théorie. En d'autres termes, **une théorie est une explication causale qui met en relation des faits entre eux**.

On s'est souvent demandé si la démarche scientifique consiste à partir d'une théorie et à la vérifier par des faits (**approche hypothético-déductive**), ou bien à partir des faits pour construire une théorie (**approche inductive ou empiriste**). Cette question a certainement autant de sens que de savoir qui, de la poule ou de l'œuf, est venu en premier. En effet, il faut plutôt se représenter la démarche scientifique (et le processus de connaissance en général) comme une spirale ascendante qui se développe à l'infini et qui nous fait passer alternativement de droite à gauche, de la théorie aux faits, de l'explication causale à l'observation empirique.

L'induction et la déduction conduisent toutes deux à des excès. Par exemple, vouloir confirmer une théorie aboutit souvent à se limiter aux hypothèses produites par cette théorie : ce **théoricisme** conduit alors à voir seulement ce qu'on a envie de voir à travers le prisme conceptuel de la théorie qu'on soutient. À l'inverse, explorer les faits ne conduit souvent qu'à les compiler, sans se donner les moyens de les interpréter. L'**empirisme**, c'est-à-dire le manque de conceptualisation, incite à reproduire les théories anciennes, les modèles stéréotypés. Il s'agira alors de trouver un bon équilibre entre l'induction et la déduction : l'une et l'autre doivent se répondre et se corriger mutuellement.

Ainsi les enquêtes procèdent à la fois à la vérification de théories et à la recherche de nouvelles explications. La méthodologie de collecte des données est établie sur la base des expériences passées : les outils de collecte et d'analyse sont supposés être adéquats pour confirmer ou infirmer une théorie. Mais l'enquête peut aussi révéler des faits inattendus, ou présenter des faits connus sous un jour nouveau. Les faits peuvent ne pas confirmer une théorie mais susciter de nouvelles explications. Nous verrons à propos de la sélection des modèles statistiques que les approches exploratoire et confirmatoire, qui correspondent l'une à l'induction et l'autre à la déduction, sont intrinsèquement liées.

c) Signification et significativité

L'indéterminisme fondamental que nous avons décrit plus haut à propos des actions humaines, a une conséquence importante sur l'interprétation en termes

probabilistes. **Dans une situation déterministe, un événement devrait être prédit par les variables explicatives et par elles seules, dans la limite des erreurs de mesure.** Si ce n'est pas le cas, le modèle théorique utilisé pour la prédiction n'est pas validé. Même si ce modèle prédit l'événement à 80 %, il n'en reste pas moins inadéquat dans un contexte déterministe. On imagine en effet les conséquences d'une erreur de 20 % dans le cas où l'on veut envoyer un astronef sur Mars ! On cherchera plutôt une probabilité proche de 100 %, la différence étant attribuée aux erreurs de mesure.

Dans le cas d'un indéterminisme fondamental, caractéristique des sciences sociales mais aussi des sciences biologiques par opposition aux sciences physiques, il n'est pas nécessaire d'atteindre un tel niveau de probabilité. Un médicament qui améliore le taux de rémission de 20 % est un bon médicament. Une variable explicative qui améliore l'ajustement d'un modèle aux données mérite toujours attention. La **pertinence** d'un facteur se mesure par la différence entre les modèles sans et avec le nouveau facteur explicatif, c'est-à-dire par l'amélioration de la prédiction. Mais un facteur, ou un ensemble de facteurs, peut être pertinent sans être totalement explicatif.

Dans un contexte non déterministe, la prédiction ne peut pas constituer un critère définitif pour la sélection d'une bonne théorie. D'abord parce que la prédiction ne peut être totale, mais aussi parce que l'amélioration de la prédiction ne valide pas nécessairement les composantes du modèle théorique. Pourquoi cela ?

La pertinence (mesurée par l'amélioration de la prédiction) n'est pas un critère suffisant pour la sélection du meilleur modèle. Tout dépend de la **cohérence interne** des éléments de la théorie explicative. On peut facilement prendre une cause pour une autre. Nous l'avons vu plus haut à propos du décès de l'enfant. Même si la naissance du deuxième enfant prédit mieux le décès du premier, le niveau d'instruction de la mère peut fournir une meilleure explication au décès. La relation positive entre la nouvelle naissance et le décès du premier enfant n'est peut-être due qu'à un facteur commun : le niveau d'instruction de la mère influe à la fois sur la reprise des conceptions et sur la morbidité des enfants ; si la reprise des conceptions est plus rapide que le développement des maladies, alors on sera tenté de penser que l'un cause l'autre simplement parce que l'un précède l'autre.

Ceci nous mène à **distinguer la signification et la significativité**. La signification est le sens que l'on donne aux choses, tandis que la significativité est l'évaluation statistique d'une différence quantitative. Dans le premier cas, on fait appel à la cohérence interne de la théorie explicative, dans l'autre on fait appel à des lois générales, que l'on s'impose comme critère de validation extérieure. **Le degré de significativité ne pourra jamais nous renseigner sur la causalité, mais seulement sur la corrélation entre un facteur explicatif et le phénomène à expliquer.**

En statistique, la mesure des relations entre variables ne préjuge en rien des liens de causalité. L'analyse en terme de causalité devrait être bannie du commentaire statistique et pourtant, on arrive souvent à confondre corrélation et causalité. Cet abus de langage est cependant bien compréhensible, même s'il ne se justifie pas, dans la mesure où **c'est au fond un système de causalité qu'on essaie de mettre en lumière à l'aide de l'outil statistique, qui ne mesure pourtant que des corrélations**.

La distinction entre causalité et corrélation (ou simplement relation) est malgré tout essentielle, car elle nous rappelle justement que la statistique n'est qu'un outil, et que l'analyse en termes de causalité précède et dépasse tout à la fois la statistique. On retrouve la causalité dans toutes les sciences, la technique statistique ne faisant qu'ajouter dans l'analyse des éléments de justification ou de relativisation par la mesure. **L'explication n'est jamais statistique, elle est sociologique, économique, etc**.

Même le calcul statistique strict ne peut se passer d'une évaluation qualitative. Prenons un exemple : il est très difficile d'évaluer la différence en termes statistiques entre une proportion de 0 sur 50 et une proportion de 2 sur 50. On ne peut pas mesurer l'augmentation du phénomène de 0 % à 4 % (aucun chiffre ne peut multiplier 0 % pour donner 4 %) car il faudrait qu'il y ait eu au moins un individu dans le premier groupe. On pourrait alors mesurer une augmentation de 2 % (1/50) à 4 % (2/50), soit une multiplication du risque par 2. Mais si le phénomène est absent dans le premier groupe, on ne peut que constater la présence du phénomène dans le deuxième groupe.

Tout dépend de l'importance que l'on donne à la présence ou à l'absence du phénomène. Par exemple, la présence peut avoir une signification très importante pour l'étude du phénomène, comme elle peut n'avoir aucune signification si le chiffre de 2/50 n'est pas considéré comme suffisamment important au regard de 0/50. De plus, l'absence constatée d'un phénomène dans une population peut très bien être interprétée contradictoirement comme une erreur de mesure ou comme une absence réelle et signifiante.

D'après cet exemple, on peut voir que **même pour l'interprétation statistique, il est nécessaire de passer par une évaluation qualitative**. De tels cas d'ambiguïté ou d'indétermination quantitative ne sont pas aussi rares qu'on pourrait le penser.

CHAPITRE 2

LES DONNÉES ET LEUR TRAITEMENT INFORMATIQUE

Dans ce manuel, nous n'utiliserons que des données recueillies auprès des individus, c'est-à-dire auprès d'entités bien définies, et par définition, indivises. On peut imaginer des applications spécifiques sur des entités ou unités plus complexes, comme le couple, le ménage voire même un village ou une entreprise, qui ont aussi une "vie" plus ou moins longue. Le temps ne se mesure pas seulement au niveau individuel, et c'est d'ailleurs un problème conceptuel que nous aborderons dans la dernière partie de ce manuel. Mais, pour l'instant, le lecteur doit savoir que nous traiterons ici des données relatives aux individus.

Nous avons utilisé, autant qu'il était possible, des données réelles dans les exemples qui illustrent ce manuel, ceci pour rendre plus vivant, et moins théorique, un exposé parfois complexe. Les analyses concernent essentiellement l'emploi, ce qui est déjà une manière de montrer que des techniques développées initialement par la biologie et la démographie peuvent être utilisées dans n'importe quel autre domaine.

La structure du questionnaire et la collecte sur le terrain sont indissociables du traitement informatique. Ce chapitre présente le logiciel statistique utilisé pour toutes nos analyses, ainsi que les fichiers de données, et les programmes spécifiques qui accompagnent ce manuel.

1. L'enquête "Insertion urbaine" à Dakar

Les données d'où sont tirés les exemples d'analyses couvrent divers sujets : dans l'enquête "Insertion urbaine" à Dakar, on a recueilli la biographie des individus sous l'angle des changements d'emploi, de résidence et d'état matrimonial. Ce principe d'analyse tri-biographique (suivant le principe élaboré par Daniel Courgeau

pour l'enquête 3-B de 1981 en France (Riandey, 1985)) permet des analyses complexes d'interférences entre les événements de différents ordres connus par chaque individu. On peut faire une analyse, classique en démographie, de la nuptialité, aussi bien qu'une analyse de l'entrée dans la vie active ou de l'accès à la propriété, ou encore une analyse des interférences entre vie professionnelle et itinéraire matrimonial chez les femmes.

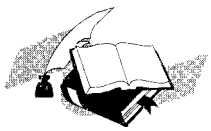
En effet, trois composantes de l'insertion en ville ont été retenues par l'équipe "Insertion urbaine" de Dakar : l'accès au travail, l'accès au logement, la constitution du ménage (à travers l'itinéraire matrimonial et le devenir des enfants) et son éventuel éclatement géographique. Ces composantes sont situées dans le contexte d'urbanisation rapide dans lequel migrants et non migrants évoluent. Le processus de l'insertion en ville est replacé dans l'ensemble des cheminements migratoires connus par les individus. Il s'agissait, pour l'équipe, de voir comment migrants et non migrants, arrivent à satisfaire un certain nombre de besoins, en particulier travail et logement, alors qu'ils ne possèdent peut-être ni les mêmes atouts, ni les mêmes exigences.

L'enquête "Insertion urbaine" de Dakar, tout comme l'enquête française 3-B, contient une richesse d'informations biographiques que nous ne pourrons exploiter ici dans son entier. Le choix de l'analyse de l'emploi tient à ce que l'auteur a par ailleurs écrit un ouvrage sur "*L'insertion et la mobilité professionnelles à Dakar*" (Bocquier, 1996). Le lecteur intéressé s'y référera pour les explications sociologiques ou économiques qui ne figurent pas dans les commentaires des résultats présentés à l'occasion dans ce manuel. Ces explications ne seront évoquées que dans la mesure où elles peuvent servir à l'exposé didactique.

L'échantillon IFAN-ORSTOM en quelques mots

Au cours d'une première phase de l'enquête (octobre 1989), nous avons enquêté plus de 2 100 ménages et 17 900 personnes de tous âges et de toutes catégories. Ces personnes constituent, une image représentative de la composition des ménages de l'agglomération, des cheminements migratoires jusqu'à Dakar et des activités économiques de la ville. L'enquête ménage sert de base au tirage d'un échantillon d'individus pour l'enquête biographique après stratification par sexe et par groupe d'âges. Le tirage de ces individus s'est fait selon une stratification par groupe d'âges (25-34 ans, 35-44 ans et 45-59 ans), de telle façon qu'un nombre à peu près égal de personnes de chaque groupe de générations soit interrogé. Cela était nécessaire pour réduire les problèmes d'effectifs insuffisants qui pouvaient se poser pour les comparaisons d'une génération à l'autre. Par ailleurs, deux fois plus d'hommes que de femmes ont été interrogés, car, l'analyse nous l'a confirmé, l'itinéraire des femmes est beaucoup moins complexe que celui des hommes, et par conséquent les analyses nécessitent moins de données pour les femmes que pour les hommes. Au bout du compte, 1 557 biographies ont été recueillies à Dakar, au cours du dernier trimestre 1989.

2. Le logiciel STATA



Il existe maintenant de nombreux logiciels pour traiter les données biographiques : SAS, BMDP, SPSS, GLIM, RATE, TDA, etc. Pour des analyses très avancées, certains logiciels sont plus performant que STATA (analyse des données multi-épisodes dans TDA, par exemple). On trouvera une revue complète des possibilités offertes par ces différents logiciels dans l'ouvrage d'Éva Lelièvre et Arnaud Bringé, *"Manuel pratique pour l'analyse statistique des biographies"* (à paraître).

Cependant, nous avons choisi, pour ce manuel, de n'utiliser qu'un seul logiciel pour présenter et traiter les données. Lorsque nous avons pris connaissance de STATA[3], il s'est immédiatement imposé à nous comme un des logiciels les plus performants du marché, car il combine plusieurs avantages : convivialité, capacité de mémoire, potentialité de programmation, rapidité d'exécution.

STATA possède à la fois des commandes interactives et un langage de programmation très évolué qui permet à l'utilisateur de concevoir ses propres commandes interactives. La facilité d'exécution des commandes rend son apprentissage très rapide.

Sa convivialité se mesure aussi à sa capacité de manipulation des fichiers. En utilisant la mémoire vive (RAM) étendue (*extended memory*), la taille des fichiers n'est plus limitée par le logiciel lui-même (le programme exécutable) mais par les capacités de l'ordinateur. De plus, on peut à volonté faire varier le nombre maximum de variables ou le nombre maximum d'observations selon la longueur (nombre d'observations) ou la largeur (nombre de variables) du fichier. L'utilisation de la mémoire vive permet en outre une rapidité d'exécution des commandes étonnante sur un simple micro-ordinateur, même comparée aux performances des autres logiciels sur mini ou gros systèmes (type VAX ou Unix ; signalons d'ailleurs qu'il existe une version STATA pour Unix).

Tout cela fait qu'avec un micro-ordinateur 386 avec un minimum de 4 méga octets de mémoire étendue, un co-processeur arithmétique et la version Intercooled de STATA (la plus performante), le lecteur sera parfaitement à l'aise pour traiter tous les problèmes exposés dans ce manuel. On peut toutefois utiliser la version normale de STATA sur un simple 286 avec 640 kilo octets de mémoire vive, mais en limitant la taille des fichiers et la rapidité d'exécution. Les fichiers de données

[3] Les références du logiciel figurent en annexe 1.

joints à ce manuel sont de taille peu importante : ils ont été prévus pour être utilisés sur n'importe quel ordinateur.

La convivialité de STATA a aussi été améliorée avec la nouvelle version sous Windows. L'éditeur ligne à ligne, les résultats, la liste des variables, les graphiques, la liste des commandes précédentes, apparaissent chacun dans une fenêtre, ce qui évite de nombreuses manipulations. Le fichier peut être facilement consulté sous la forme d'une feuille de données, et peut même être modifié comme dans une feuille de calcul. Ceci dit, les commandes fonctionnent exactement de la même manière sous DOS et sous Windows.

La convivialité de STATA s'exprime aussi à travers le *STATA Technical Bulletin* (STB) : ce bulletin bi-mensuel de liaison pour les utilisateurs du logiciel permet une circulation de l'information optimale et la diffusion rapide des nouvelles techniques statistiques proposées par les concepteurs et les utilisateurs eux-mêmes. En effet, du fait de l'accès au langage de programmation interne de STATA, chaque utilisateur est en mesure de proposer ses propres procédures statistiques aux autres utilisateurs. Cela fait de STATA un logiciel particulièrement évolutif, où l'on peut trouver les dernières techniques conçues par des chercheurs. Nous avons d'ailleurs proposé des modifications à certaines commandes de STATA et écrit quelques nouvelles commandes qui figurent dans le présent manuel et sur la disquette qui l'accompagne. Certaines de ces commandes ont été proposées aux concepteurs de STATA, qui les ont diffusées à travers le numéro 22 du *STATA Technical Bulletin*.

Enfin, un nombre important de commandes graphiques accompagne les commandes statistiques. Nous en ferons largement usage dans la mesure où l'image est souvent plus parlante et son interprétation plus immédiate qu'une statistique.

3. La saisie et le transfert des données

On peut saisir les données dans STATA mais ce logiciel n'a pas été conçu à cette fin. Il est donc conseillé de se servir d'un autre logiciel qui intégrera plus facilement des procédures de contrôle, et de transférer les données en format STATA ensuite. Pour notre part, nous avons utilisé les logiciels ISSA pour la saisie, SPSS et Stat/Transfer pour le transfert des données (annexe 2).

À l'origine, chaque module du questionnaire constitue un seul fichier, avec un numéro d'identifiant pour l'individu. Les questionnaires ont été saisis à l'aide du

logiciel ISSA[4] (*Integrated Statistical and Survey Analysis*) qui permet une saisie par occurrences successives (les différentes périodes d'emplois par exemple, se succèdent jusqu'à ce qu'un code indique que le module est entièrement saisi ; on passe alors au module suivant pour le même individu). Cette saisie permet la création de fichiers distincts qu'on peut ensuite fusionner en utilisant l'identifiant de l'individu comme clé.

Les commandes de STATA sont particulièrement séduisantes en ce qui concerne la manipulation des fichiers et leur traitement pour l'analyse des biographies. En effet, les données biographiques sont préférablement stockées sous la forme d'une ligne d'observation par période (d'emploi, de résidence...) à laquelle on peut associer les caractéristiques permanentes de l'individu (sexe, date de naissance, origines sociales...). Dans la plupart des logiciels, le traitement des données biographiques nécessite alors la reconstitution de fichiers individus où les périodes biographiques sont mises bout à bout sur une même ligne. Cette manipulation des données, tous les statisticiens vous le dirons si vous ne l'avez pas expérimentée vous-même, est particulièrement fastidieuse, et la programmation statistique aussi, notamment pour le traitement des variables indépendantes fonction du temps dans l'analyse des biographies.

STATA utilise les fichiers contenant une ligne d'observation par période. Le fichier utilisé est donc parfaitement rectangulaire (chaque observation est décrite par le même nombre de variables qui ne dépend pas du nombre de périodes vécues). Nous verrons que les manipulations de fichiers s'en trouvent considérablement facilitées, de même que la transformation des variables pour l'analyse.

4. Les fichiers utilisés et leur usage

Toutes les données utilisées pour les exemples des chapitres suivants figurent sous forme de fichiers sur la disquette accompagnant ce manuel. Comme nous l'avons dit précédemment, ils proviennent tous de l'enquête "Insertion urbaine" de Dakar. L'échantillon est constitué des personnes qui ont été interrogées lors de l'enquête biographique, c'est-à-dire trois groupes de générations, nées dans les périodes 1930-44, 1945-54 et 1955-64, et qui avaient donc respectivement 45-59 ans, 35-44 ans et 25-34 ans au 31 décembre 1989.

Les données illustreront successivement la régression simple, la régression logistique, la description non paramétrique d'un ou deux événements, et les différentes formes de régressions semi-paramétriques.

---

[4] Pour se procurer le logiciel, contacter directement ISSA (adresse en annexe).

a) La régression simple : l'usage d'une variable dépendante continue

Nous avons choisi une des rares variables continues saisie dans notre enquête, le salaire en francs CFA, pour illustrer le modèle de régression classique. Pour cela, la période du dernier emploi des hommes salariés au moment de l'enquête a été sélectionnée dans le fichier SAL.DTA. Les données se présentent sous forme d'un fichier où chaque ligne représente un individu.

b) La régression logistique et la régression semi-paramétrique : l'usage d'une variable dépendante dichotomique ou polytomique

Une illustration de l'usage d'une variable dichotomique est fournie par la création du fichier JOB1.DTA dans le chapitre concernant la régression logistique. Nous avons choisi l'itinéraire professionnel des femmes pour illustrer par des exemples ce qui constitue l'essentiel de ce manuel : l'analyse biographique. L'événement étudié ici est le premier emploi des femmes. L'âge minimum à partir duquel on a recueilli la biographie professionnelle est de 12 ans.

Le fichier à partir duquel a été créé le fichier JOB1.DTA s'appelle JOB.DTA et comporte toutes les périodes d'activité depuis l'âge de 12 ans jusqu'à l'enquête, et non pas seulement la première période d'emploi. C'est donc un fichier qui se présente tel qu'il a été saisi : le lecteur verra comment traiter un fichier réel sans qu'une sélection préalable n'ait été faite selon les individus ou les périodes.

c) La régression semi-paramétrique avec des variables indépendantes fonction du temps

Avec les variables évoluant dans le temps, on aborde à la fois les problèmes de création de fichiers et les problèmes d'analyse. Le lecteur se rendra compte que le traitement des autres types de variables est un préalable nécessaire, bien que le fichier JOB.DTA tel qu'il a été saisi suffise pour l'analyse des transitions.

Le dernier fichier utilisé dans ce manuel s'appelle MAR.DTA : il comprend toutes les périodes associées aux différents états matrimoniaux des femmes jusqu'à l'enquête. Ce fichier comprend les périodes de célibat, de mariage, de divorce ou de veuvage.

La commande pour fusionner deux fichiers contenant chacun des périodes successives est assez complexe. Il est préférable de pratiquer d'abord sur les fichiers simples. Nous verrons le moment venu comment le fichier MAR.DTA peut être

combiné au fichier JOB.DTA, pour former un nouveau fichier JOBMAR.DTA qui permet d'analyser l'influence réciproque des événements matrimoniaux et professionnels.

d) Les programmes STATA sur la disquette accompagnant ce manuel

Les fichiers programmes qui figurent sur la disquette accompagnant ce manuel ont pour extension « *.ado », et sont parfois couplés avec un programme ayant le même nom avec extension « *.hlp ». Par exemple, les programmes « censor.ado » et « censor.hlp » vont ensemble. Pour exécuter ces programmes disponibles sur la disquette, il faut procéder en deux étapes :

1. Créer un répertoire « c:/ado/ »,

2. Copier tous les programmes contenus sur la disquette (extension « *.ado » et « *.hlp ») sur le répertoire « c:/ado/ ».

5. Quelques lignes de commandes en guise d'initiation à STATA

Il n'est heureusement pas besoin de longs discours pour expliquer chacune des commandes proposées par le logiciel : leur usage est relativement intuitif. Il suffit de taper « help *nom_commande* » pour obtenir quelques explications sur la syntaxe d'une commande, mais évidemment le manuel STATA est irremplaçable.

En guise d'initiation, nous proposons au lecteur de découvrir lui-même le fichier SAL.DTA. Les commandes suivantes ne modifient pas les variables existantes mais en créent quelques nouvelles. Elles permettent de mieux comprendre comment est construit un fichier STATA. Pour cela, lisez attentivement les messages qui vous sont renvoyés après chaque commande.

(Configurer la mémoire vive)

```
. bmemsize sal
```

(Frapper par exemple la touche [F4])

```
. use sal
. describe
. exit
. describe
. use sal
```

(Création d'un fichier de résultat en format ASCII)

```
. log using resultat
```

(Explorer les données)

```
. list in 1/2
. list no v306 v311 v603 v611 v632 v633
```

(Frapper [page up] pour appeler la ligne précédente et la modifier en utilisant les touches de direction)

```
. list no v306 v602-v611
. list no v306 v60*
```

(Modifier les données)

```
. generate bidon=1
. exit
. drop bidon
. exit
. exit, clear
. use sal
. describe bidon
```

(Faire des tabulations simples)

```
. tabulate v611 v610
. tabulate v610 v611
. tabulate v311 v610 v611
```

(Trier le fichier et faire des tabulations plus complexes)

```
. list no v311 v306 in 1/10
. by v311 : tabulate v610 v611
. sort v311
. list no v311 v306 in 1/10
. by v311 : tabulate v610 v611
. exit
. save, replace
. desc, short detail
. by v311 : tab v610 v611
. exit
. use sal
```

(Fermer temporairement le fichier de résultat)

```
. log off
```

(La commande suivante se lit : générer une variable nommée "pub" qui sera égale à 1 si v610 est égale à 3 et à 0 sinon.)

```
. gen pub=cond(v610==3,1,0)
. tab pub v610
. gen pub=v610==3
```

(La commande suivante se lit : générer une variable nommée "public" qui sera égale à 1 si l'expression v610=3 est vraie et à 0 sinon.)

```
. gen public=v610==3
```

(Cette deuxième formulation, très concise, est particulièrement utile pour la création de variables dichotomiques, codées 0 ou 1. Elle est absolument équivalente à la première formulation pour la création de la variable "pub", comme on peut aisément le vérifier en comparant les résultats des deux calculs.)

```
. tab public pub
. compare public pub
```

(Traitement des valeurs manquantes)

```
. gen public1=v610==3 if v349==1
. compare pub public1
. list pub* v610 v349 in 1/10
. gen public2=cond(v610==3 & v349==1,1,0)
. gen public3=v610==3 & v349==1
. list pub* v610 v349 in 1/10
. desc pub*
```

(Économiser de la mémoire : remarquez la valeur de "width")

```
. describe, short detail
. compress
. desc pub*
. describe, short detail
```

(Rouvrir le fichier de résultat)

```
. log on
```

(Création d'une série de variables dichotomiques correspondant à chacune des modalités d'une variable polytomique)

```
. tab v610, gen(v610_)
. desc v610 v610_*
. tab v610 v610_4
. tab v610_4 pub
. compare v610_4 pub
```

(Fermer définitivement le fichier de résultat)

```
. log close
```

(Le fichier de résultat sera automatiquement fermé si l'on interrompt la session)

```
. exit
. exit, clear
```

CHAPITRE 3

LA LOGIQUE DE L'ANALYSE MULTIVARIÉE : LA RÉGRESSION STATISTIQUE

Dans ce chapitre, nous introduisons la régression statistique, son utilité mais aussi ses limites. Pour bien en comprendre les principes, nous verrons les différents types de causalité entre les variables que ces modèles permettent de prendre en compte. Quelques techniques de diagnostic de la régression seront exposées, de même qu'une discussion sur les procédures de sélection des modèles.

1. L'usage du contrôle statistique pour déterminer la structure des relations entre plusieurs variables

Comment passe-t-on de la description à l'analyse ? Le premier stade de la statistique est généralement la construction d'un tableau croisé à deux variables. Les résultats de ce croisement sont souvent considérés, à tort, comme "neutres", parce qu'ils sont "simples" à produire. Un exemple montrera que cette neutralité n'est qu'une apparence. L'analyse multivariée est indispensable pour sortir des contraintes de l'analyse descriptive.

a) Faire un tableau, c'est déjà construire un modèle

On veut savoir combien les salariés gagnent au moment de l'enquête, en 1989, selon leur sexe, leur statut migratoire et leur génération. Le salaire actuel en francs CFA est contenu dans la variable v633. Les salaires codés "888888" (plus d'un million de francs CFA) n'ont pas été pris en considération, de même que les salaires inconnus ("999999") ou nuls ("0") au moment de l'enquête ; ces cas représentent

5,3 % de l'échantillon. Pour ne pas faire de confusion, nous avons préféré créer une variable "salaire" où seuls les montants connus seront mentionnés, et éliminer du fichier les individus dont le salaire est manquant (*missing values*) :

```
. use sal

. gen salaire=v633 if v633 !=0 & v633 !=888888 & v633 !=999999
(28 missing values generated)

. drop if salaire==.
(28 observations deleted)

. save sal, replace
```

Les salariés se répartissent comme suit, par sexe, année de naissance et statut migratoire :

```
. sort v304

. by v304 : tab migrant strate

-> v304=masculin
           | strate
   migrant |    G30-44     G45-54     G55-64 |     Total
-----------+---------------------------------+----------
       oui |        73         70         38 |       181
       non |        67        115         82 |       264
-----------+---------------------------------+----------
     Total |       140        185        120 |       445

-> v304=feminin
           | strate
   migrant |    G30-44     G45-54     G55-64 |     Total
-----------+---------------------------------+----------
       oui |         4          8         11 |        23
       non |         3         10         20 |        33
-----------+---------------------------------+----------
     Total |         7         18         31 |        56
```

On peut calculer le salaire moyen ainsi que l'écart-type avec l'option « summarize() » de la commande « tabulate » :

```
. tab v304, summ(salaire)
           |          Summary of salaire
      sexe |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
  masculin |   102646.03   106632.59         445
   feminin |    72620.25   73278.621          56
-----------+------------------------------------
     Total |   99289.856   103814.13         501

. tab migrant, summ(salaire)
           |          Summary of salaire
   migrant |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
       oui |   78376.794   91848.349         204
       non |   113654.38   109137.45         297
-----------+------------------------------------
     Total |   99289.856   103814.13         501
```

```
. tab strate, summ(salaire)
           |          Summary of salaire
      strate|        Mean   Std. Dev.       Freq.
------------+------------------------------------
     G30-44 |   98485.476   92087.926         147
     G45-54 |    116687.6   119274.45         203
     G55-64 |   76683.907   87106.943         151
------------+------------------------------------
      Total |   99289.856   103814.13         501
```

Ces tabulations correspondent aux schémas de causalité d'une condition permanente (chapitre 1) :

Salaire
1989

où *P* représente alternativement le sexe, la génération ou le statut migratoire.

Les trois différentes variables ne sont pas des événements : elles ne sont pas situées clairement dans le temps et sont donc assimilables à des conditions permanentes. Même le salaire en 1989 a un statut ambigu : on ne sait pas quand ce montant a été atteint. Ce n'est donc pas un événement, mais le résultat de tout un processus qui a mené le salarié à toucher telle ou telle rémunération au moment de l'enquête. Ce qu'on mesure ici, ce sont des corrélations entre le salaire et différentes variables explicatives, mais on est loin du processus qui aboutit à fixer le salaire à tel ou tel niveau.

L'échantillon étant stratifié par sexe et par génération, le salaire calculé en contrôlant ces deux variables donnera une meilleure image de la réalité[5] :

```
. by v304 : tab migrant strate, summ(salaire) nost

-> v304=masculin
           Means and Frequencies of salaire

           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+---------------------------------+----------
       oui | 80415.658  97425.371  55173.026 | 81694.442
           |        73         70         38 |       181
-----------+---------------------------------+----------
       non | 119275.37  135711.37  88933.329 | 117010.57
           |        67        115         82 |       264
-----------+---------------------------------+----------
     Total | 99012.807  121224.77  78242.567 | 102646.03
           |       140        185        120 |       445
```

[5] Il est possible de pondérer les données, mais pour ne pas alourdir l'exposé et la présentation des données, nous n'utiliserons jamais les pondérations : les variables de stratification (sexe et année de naissance (variable intitulée "strate")) constituent le niveau minimum de désagrégation des données.

```
-> v304=feminin
          Means and Frequencies of salaire

           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+------------------------------------+----------
       oui |     58743      61000  43563.636 |  52268.348
           |         4          8         11 |         23
-----------+------------------------------------+----------
       non |  126866.67      77300    85548.1 |  86804.909
           |         3         10         20 |         33
-----------+------------------------------------+----------
     Total |  87938.857   70055.556  70650.387 |   72620.25
           |         7         18         31 |         56
```

Cette fois-ci, c'est l'interaction des trois variables sexe, génération et statut migratoire sur le montant du salaire que l'on a évalué :

Sexe Génér. Migr.

Salaire
1989

On imagine dès maintenant que le tableau deviendrait difficile à comprendre si l'on voulait évaluer l'interaction de plus de trois variables, à l'aide d'un tableau croisant d'autres variables encore. Au-delà de trois dimensions, l'esprit humain a tout simplement du mal à visualiser les choses. En fait, **un tableau à plus de quatre dimensions est souvent incompréhensible pour une personne normalement constituée** (y compris un statisticien).

Mais cherche-t-on vraiment à évaluer exactement le salaire pour telle ou telle case du tableau ? On veut plutôt savoir quelle est l'influence de telle ou telle variable sur le phénomène étudié. Pourquoi, alors, ne pas faire des tableaux séparément pour chaque variable ? Parce que les variables sont **corrélées** entre elles.

Reprenons l'exemple ci-dessus. On a, dans le tableau évaluant le salaire moyen selon le sexe, un écart de plus de 30 000 entre les hommes et les femmes. Si l'on va plus loin dans l'analyse, on voit que la génération et le statut migratoire jouent aussi un rôle. Le salaire des femmes n'est pas uniformément inférieur à celui des hommes. Par exemple, on peut voir que les femmes non migrantes des vieilles générations (1930-44) gagnent en moyenne plus que les hommes non migrants des mêmes générations. Les écarts sont en revanche nettement en défaveur des femmes, migrantes, ou non pour les générations qui suivent (1945-54).

Le tableau tenant compte de la génération et du statut migratoire est certainement très instructif, mais on voit tout de suite la limite d'un commentaire lorsque les cases du tableau ne contiennent que quelques individus, ce qui arrive fréquemment dans l'échantillon des femmes.

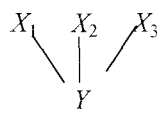
L'analyse au-delà de trois dimensions pose des problèmes à la fois physique (présentation d'un tableau) et physiologique (capacité humaine à concevoir des relations entre plus de trois variables). Pour l'analyse du salaire, les commentaires seront très différents avant et après introduction des variables statut migratoire et génération, mais ils seront peut-être encore très différents si l'on introduisait d'autres variables supplémentaires, pour former un tableau à plus de trois dimensions.

b) Le principe de la régression statistique

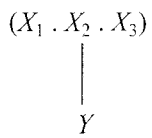
Par convention, dans l'analyse statistique, on appelle **variable dépendante** une caractéristique de la population que l'on cherche à expliquer par l'analyse statistique. Dans l'exemple précédent, il s'agit du salaire. On peut simplement décrire la répartition de cette caractéristique à travers l'échantillon, mais on préfère généralement établir une relation causale entre cette caractéristique et d'autres caractéristiques de la population. Dans ce cas, ces autres caractéristiques sont appelées explicatives, ou encore **variables indépendantes**, se sont le sexe, le statut migratoire et la génération dans l'exemple qui précède.

Comment fait-on quand on ne peut commenter ni les tableaux univariés (variable par variable), ni un grand tableau croisant toutes les variables dont on dispose, avec des effectifs très faibles dans chaque case du tableau ? On doit faire appel à des techniques d'analyse multivariée qui permettront de **mesurer l'influence d'une variable explicative tout en contrôlant l'influence des autres variables explicatives sur la variable expliquée**, quel que soit leur nombre. C'est le sens de la phrase rituelle "toutes choses égales par ailleurs" ou encore de l'expression latine *in ceteris paribus*.

On voudrait aboutir à la relation causale :



Notons qu'il ne s'agit pas de la même relation causale que dans le tableau croisé, où l'on mesure l'interaction des trois variables indépendantes X_i sur la variable dépendante Y :



Dans le modèle de régression, on suppose que les variables X_i agissent indépendamment les unes des autres : leurs effets s'ajoutent les unes aux autres.

Dans le tableau croisé, au contraire, on suppose que les variables X_i agissent en interaction, c'est-à-dire qu'elles ne sont pas indépendantes entre elles.

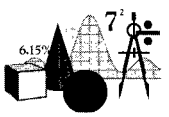
Avant de voir comment mettre en pratique l'analyse de régression, faisons une petite digression dans le domaine de l'histoire des statistiques. Savez-vous d'où vient le terme de "régression statistique" ? Vous vous êtes certainement interrogés comme nous sur la signification de ce mot, sans avoir jamais osé la demander aux statisticiens qui vous entourent. Eh bien, voilà la petite histoire : au siècle dernier, un chercheur a voulu vérifier l'hypothèse selon laquelle les enfants étaient plus grands que leurs parents. Il a donc collecté des données sur ses compatriotes et a mis au point une formule mettant en rapport la taille moyenne des générations. En fait, il a constaté que plus la taille moyenne des parents était élevée par rapport à la moyenne de l'échantillon, plus la taille moyenne des enfants étaient inférieure à celle de leurs parents. Il en a donc conclu à une *régression* de la taille des enfants par rapport à celle de leur parents. Les chercheurs qui l'ont suivi ont repris la technique en même temps que le terme, et c'est ainsi qu'est née "l'analyse de régression".

c) Les hypothèses de base pour appliquer le modèle de régression

La formulation d'un modèle de régression sans variable explicative est,

$$E(Y_j) = c,$$

où Y_j est la variable dépendante pour l'individu j, $E(Y_j)$ l'espérance mathématique de cette variable dépendante, et c une constante égale, par exemple, au salaire moyen pour l'ensemble de l'échantillon. Il ne s'agit cependant pas à proprement parler d'un modèle, dans la mesure où il n'y a pas de relation entre différentes variables.



La formulation du modèle avec une seule variable explicative est la suivante :

$$E(Y_j) = c + bX_j$$

où X_j définit une caractéristique connue pour l'individu j : c'est la variable indépendante. Cette fois c est une constante égale au salaire moyen chez les individus qui ne possèdent pas la caractéristique X. Le coefficient b représente la différence de salaire entre ceux qui possèdent la caractéristique X et les autres. Ce modèle correspond à la relation causale :

$$X \rightarrow Y$$

où l'on remarque que le temps n'intervient pas.

De façon générale, le modèle classique de régression linéaire nécessite de faire cinq hypothèses sur les données analysées :

1. **La variable dépendante est une fonction linéaire d'un ensemble de variables indépendantes, plus une erreur aléatoire.** La formulation du modèle est donc $E(Y_j) = c + \sum_i b_i X_{ij} + \varepsilon$. Les violations de cette hypothèse sont de trois ordres :

   - *les variables indépendantes ne sont pas convenablement choisies*, soit parce que des variables indépendantes ont été omises, soit parce que les variables indépendantes choisies ne sont pas pertinentes ;
   - *la fonction n'est pas linéaire* entre la variable dépendante et les variables indépendantes ;
   - *les variables indépendantes ne sont pas constantes* au cours de la période de collecte des données.

2. **L'espérance mathématique de l'erreur aléatoire est nulle**, c'est-à-dire que $E(\varepsilon) = 0$. Si ce n'est pas le cas, l'estimation est biaisée.

3. **L'erreur aléatoire a une variance unique pour tout l'échantillon, et les erreurs associées à chaque individu de l'échantillon ne sont pas corrélées entre elles.** Les violations de cette hypothèse sont de deux ordres :

   - *Hétéroscédasticité* : ce terme compliqué veut dire que les erreurs ne sont pas tirées de la même distribution pour tout l'échantillon ;
   - *Les erreurs sont autocorrélées entre elles*. Ce cas se présente surtout dans séries temporelles.

4. **Il est possible de retrouver le même effet des variables indépendantes avec un autre échantillon tiré dans la même population.** Les violations de cette hypothèse sont de trois ordres :

   - les variables n'ont pas été correctement mesurées ;
   - le modèle est autorégressif, c'est-à-dire qu'une variable indépendante est une simple transformation mathématique de la variable dépendante ;
   - la variable dépendante est le résultat d'une équation simultanée, c'est-à-dire que les variables indépendantes en interaction déterminent de manière mathématique (unique) la variable dépendante. Cette situation est la généralisation pour plusieurs variables de la violation précédente où une seule variable intervenait.

5. **Il n'y a pas de multicolinéarité entre les variables indépendantes**, c'est-à-dire qu'elles ne sont pas liées entre elles par une fonction linéaire, même approximativement.

Toutes ces hypothèses sont liées entre elles. Par exemple, l'hétéroscédasticité constatée dans l'échantillon peut être liée à la mauvaise spécification du modèle, et l'introduction d'une nouvelle variable pertinente peut corriger le problème. La vérification de la validité des hypothèses se fait dans un premier temps par un examen attentif des données, surtout en ce qui concerne le choix des variables pertinentes et l'existence de relations mathématiques ou de phénomènes de multicolinéarité entre variables. Une fois l'équation du modèle bien définie, on vérifiera la validité des autres hypothèses essentiellement par observation des résidus après modélisation, comme on le verra dans la suite de ce chapitre.

2. La pratique de l'analyse multivariée avec STATA

a) La création des variables indépendantes sous forme polytomique

Reprenons le salaire en francs CFA comme variable dépendante, et le sexe, le statut migratoire et la génération comme variables indépendantes, que l'on va introduire progressivement dans la régression.

Par exemple, pour le sexe féminin, on pourra créer la variable "femme" comme suit (voir l'initiation à STATA au chapitre précédent) :

```
. gen byte femme=v304==2
```

La génération est représentée par la variable "strate", mais elle doit figurer dans la régression sous la forme d'un codage polytomique. La catégorie de référence pour cette variable sera par exemple le premier groupe de générations (1930-44) et on mesurera l'écart moyen du salaire entre ce groupe et les suivants (1945-54 et 1955-64). Pour faire les calculs, on peut utiliser l'option « gen() » de la commande « tabulate » :

```
. tab strate, gen(gener)
     strate|      Freq.     Percent        Cum.
-----------+-----------------------------------
     G30-44 |        147       29.34       29.34
     G45-54 |        203       40.52       69.86
     G55-64 |        151       30.14      100.00
-----------+-----------------------------------
      Total |        501      100.00
```

On a ainsi créé les variables "gener1", "gener2" et "gener3" :

```
. desc gener*

103. gener1       byte    %8.0g             strate==G30-44
104. gener2       byte    %8.0g             strate==G45-54
105. gener3       byte    %8.0g             strate==G55-64
```

On vérifie que l'on a bien créé des variables dichotomiques correspondant à chacune des modalités de la variable "strate" :

```
. tab strate gener1

           | strate==G30-44
    strate |         0          1 |     Total
-----------+----------------------+----------
    G30-44 |         0        147 |       147
    G45-54 |       203          0 |       203
    G55-64 |       151          0 |       151
-----------+----------------------+----------
     Total |       354        147 |       501

. tab strate gener2

           | strate==G45-54
    strate |         0          1 |     Total
-----------+----------------------+----------
    G30-44 |       147          0 |       147
    G45-54 |         0        203 |       203
    G55-64 |       151          0 |       151
-----------+----------------------+----------
     Total |       298        203 |       501

. tab strate gener3

           | strate==G55-64
    strate |         0          1 |     Total
-----------+----------------------+----------
    G30-44 |       147          0 |       147
    G45-54 |       203          0 |       203
    G55-64 |         0        151 |       151
-----------+----------------------+----------
     Total |       350        151 |       501
```

Le statut migratoire est contenu dans la variable "migrant". On peut utiliser l'option « gen() », pour créer une variable dichotomique :

```
. tab migrant, gen(mig_)
    migrant |      Freq.     Percent        Cum.
------------+-----------------------------------
        oui |         204       40.72       40.72
        non |         297       59.28      100.00
------------+-----------------------------------
      Total |         501      100.00
```

b) L'interprétation des coefficients

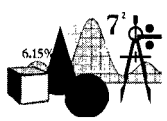
On estime le modèle de régression du salaire sur la génération,

Génération

Salaire
1989

```
. fit salaire gener2 gener3

  Source |       SS       df       MS                  Number of obs =     501
---------+------------------------------                F(  2,   498) =    6.58
   Model |  1.3870e+11     2  6.9352e+10                Prob > F      =  0.0015
Residual |  5.2500e+12   498  1.0542e+10                R-square      =  0.0257
---------+------------------------------                Adj R-square  =  0.0218
   Total |  5.3887e+12   500  1.0777e+10                Root MSE      =  1.0e+05

------------------------------------------------------------------------------
 salaire |      Coef.   Std. Err.       t     P>|t|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   18202.12   11119.66     1.637   0.102    -3645.106    40049.36
  gener3 |  -21801.57   11896.66    -1.833   0.067    -45175.41     1572.27
   _cons |   98485.48   8468.482    11.630   0.000     81847.12    115123.8
------------------------------------------------------------------------------
```

Dans ce tableau, la constante ("_cons") du modèle représente le salaire moyen d'un individu qui appartient à la catégorie de référence, c'est-à-dire à la génération 1930-44 ("gener1"). Les coefficients correspondant aux modalités "gener2" et "gener3", représentent le montant du salaire à ajouter ou à soustraire à la constante, pour les individus nés dans les années 1945-54 ("gener2") et 1955-64 ("gener3").



On a donc comme modèle,

$E(Y_j) = c + b_1X_{1j} + b_2X_{2j}$

qui équivaut à :

salaire= _coef[_cons] + _coef[gener2].gener2 + _coef[gener3].gener3

Pour évaluer le salaire dans la génération 1945-54, on peut faire le calcul :

```
. display _coef[_cons] + _coef[gener2]
116687.6
```

et pour la génération 1955-64 :

```
. display _coef[_cons] + _coef[gener3]
76683.907
```

Un tel modèle n'a cependant que peu d'intérêt par rapport à une simple moyenne des salaires par génération. Comparons le salaire prédit par le modèle pour chaque individu de l'échantillon avec le salaire moyen dans l'échantillon :

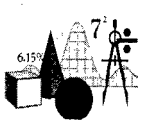
```
. predict salpred

. tab strate, summ(salpred)

           |          Summary of salpred
    strate |       Mean   Std. Dev.       Freq.
-----------+------------------------------------
    G30-44 |  98485.477          .          147
    G45-54 |   116687.6          .          203
    G55-64 |  76683.906          0          151
-----------+------------------------------------
     Total |  99289.856   16655.62          501
```

Dans ce tableau, l'écart-type pour chaque valeur de "salpred" est nul (à l'arrondi près : les valeurs manquantes "." doivent être ici interprétées comme nulles), ce qui signifie que la probabilité calculée par le modèle est la même pour tous les individus appartenant à une même catégorie de la variable "strate". L'écart-type du salaire prédit "salpred" pour le total de l'échantillon correspond à ce que serait l'écart-type si les données se conformaient parfaitement au modèle.

On peut comparer les résultats du modèle avec les proportions et les écarts-types réels dans l'échantillon :

```
. tab strate, summ(salaire)

           |          Summary of salaire
    strate |       Mean   Std. Dev.       Freq.
-----------+------------------------------------
    G30-44 |  98485.476   92087.926          147
    G45-54 |   116687.6   119274.45          203
    G55-64 |  76683.907   87106.943          151
-----------+------------------------------------
     Total |  99289.856   103814.13          501
```

Les valeurs observées et prédites par le modèle sont les mêmes parce que nous n'avons introduit qu'une seule variable (le groupe de générations) dans le modèle. Un modèle de ce type n'a en fait aucun intérêt, puisqu'il ne fait pas intervenir simultanément plusieurs variables explicatives pour calculer l'effet propre "toutes choses égales par ailleurs" de chacune des variables.



On va donc utiliser un modèle à plusieurs variables explicatives :

$$E(Y_j) = c + \sum_i b_i X_{ij}$$

où b_i et X_i sont respectivement les coefficients et les variables indicatrices.

On peut maintenant ajuster le salaire sur la génération, le sexe et le statut migratoire :

```
. fit salaire gener2 gener3 femme mig_1
  Source |       SS       df       MS                  Number of obs =     501
---------+------------------------------                F(  4,   496) =    8.05
   Model |  3.2854e+11     4  8.2135e+10                Prob > F      =  0.0000
Residual |  5.0601e+12   496  1.0202e+10                R-square      =  0.0610
---------+------------------------------                Adj R-square  =  0.0534
   Total |  5.3887e+12   500  1.0777e+10                Root MSE      =  1.0e+05
------------------------------------------------------------------------------
 salaire |      Coef.   Std. Err.      t    P>|t|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
 gener2  |   13855.27   11034.27     1.256   0.210    -7824.401    35534.93
 gener3  |  -25907.17   12082.96    -2.144   0.033    -49647.27   -2167.061
  femme  |  -21215.43   14630.77    -1.450   0.148    -49961.36    7530.508
  mig_1  |   -37383.8   9310.886    -4.015   0.000    -55677.44   -19090.16
  _cons  |   119077.7   9667.409    12.317   0.000     100083.6    138071.8
------------------------------------------------------------------------------
```

Dans cette régression, la constante du modèle représente l'estimation du salaire moyen (soit 119 077,7 francs CFA) d'un individu qui appartiendrait à la catégorie de référence pour chaque variable indépendante : il s'agirait d'un homme de la génération 1930-44 non migrant. C'est pourquoi les variables "homme", "gener1" et "mig_2" ne figurent pas dans le modèle. On peut d'ailleurs éliminer les variables superflues associées à la catégorie de référence :

```
. drop gener1 mig_2
```

Chaque coefficient permet de voir quelle variation de salaire peut être attribuée à une différence de caractéristique du salarié : par exemple, un salarié qui aurait les mêmes caractéristiques que le salarié de référence à l'exception d'être migrant, verrait son salaire diminué de 37 383,8 francs CFA. On attribue ainsi à la caractéristique "mig_1" une influence négative (et significative à un seuil inférieur à 1 pour mille) sur le salaire.

Pour retrouver le salaire d'une catégorie particulière, il suffit d'ajouter à la constante les coefficients correspondant à cette catégorie. Une femme migrante de la génération 1955-64 aurait un salaire de :

```
. display _coef[_cons]+_coef[femme]+_coef[mig_1]+_coef[gener3]
34571.336
```

Ce type de calcul est surtout utile lorsqu'on a pu sélectionner un petit nombre de variables indépendantes significatives. Les applications pratiques dans les disciplines comme la médecine, l'épidémiologie ou les assurances (actuariat), sont directes. Par exemple, il peut suffire à un praticien de connaître la valeur de deux ou trois variables pour évaluer les risques d'une maladie, les chances de réussite d'un traitement ou le montant souhaitable d'une police d'assurance.

Dans le domaine de la recherche en sciences sociales, ce type de calcul est plus rare : on ne s'intéresse généralement pas à évaluer précisément un risque pour telle ou telle catégorie de la population, mais plutôt à mesurer l'influence moyenne de telle ou telle variable. Dit autrement, on s'intéresse moins aux individus en particulier (ayant telle ou telle caractéristique) qu'aux groupes sociaux auxquels ils appartiennent. En sciences sociales, le nombre des variables testées est généralement plus important parce que la composition sociale d'une population est jugée complexe. L'individu est le produit de nombreuses appartenances. Le but de la modélisation n'est pas tant de trouver deux ou trois variables indépendantes pour expliquer la variable dépendante, que de trouver un système de causalité pour expliquer une situation sociale.

L'une et l'autre de ces approches (sélection d'un petit nombre de variables expliquant le maximum de variance, et sélection du maximum de variables significatives) ne sont cependant pas contradictoires, et on les trouvera souvent mêlées quel que soit le domaine scientifique.

Lorsque le nombre de variables explicatives n'est pas trop important, on peut calculer le salaire de chaque catégorie de travailleur, en croisant entre elles toutes les caractéristiques retenues dans le modèle. Cela revient à construire un tableau croisé de toutes les variables, et à indiquer dans chacune des cases, le montant du salaire :

```
. drop salpred
. predict salpred
. sort v304
. by v304 : tab migrant strate, summ(salpred) nost
-> v304=masculin
                    Means and Frequencies of salpred

           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+---------------------------------+----------
       oui |  81693.93  95549.195  55786.762 | 81613.246
           |        73         70         38 |       181
-----------+---------------------------------+----------
       non | 119077.73  132932.98  93170.555 | 117066.24
           |        67        115         82 |       264
-----------+---------------------------------+----------
     Total | 99584.747  118787.77  81332.354 | 102646.03
           |       140        185        120 |       445

-> v304=feminin
                    Means and Frequencies of salpred

           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+---------------------------------+----------
       oui |   60478.5  74333.766  34571.336 |  52907.34
           |         4          8         11 |        23
-----------+---------------------------------+----------
       non | 97862.297  111717.56  71955.133 | 86359.551
           |         3         10         20 |        33
-----------+---------------------------------+----------
     Total | 76500.127  95102.542  58689.915 |  72620.25
           |         7         18         31 |        56
```

En comparant ces tableaux avec les tableaux équivalents pour le salaire observé (voir plus haut), on voit immédiatement que la modélisation ne donne pas les mêmes résultats que l'analyse descriptive. Comme on l'a dit, le tableau descriptif correspond à l'évaluation d'une interaction entre les trois variables explicatives, tandis que le tableau issu de la régression correspond à un modèle additif où chaque variable joue en principe indépendamment des autres.

c) Le test des interactions entre variables indépendantes

Dans le modèle qui précède, on n'a pas testé les interactions entre variables. On peut penser que les salariés de différentes générations ont des niveaux de salaires différents selon qu'ils sont migrants ou non, de sexe masculin ou féminin.

Commençons par tester l'interaction entre la génération et le statut migratoire. Il faut alors créer deux séries de modalités pour la génération, l'une correspondant aux migrants, l'autre aux non migrants (en considérant par exemple comme modalité de référence, les migrants des générations 1930-44). Deux méthodes sont possibles pour créer ces modalités. La première consiste à conserver les variables définies pour les générations, et à créer une série de variables qui mesurent la différence entre migrant et non migrant pour une même génération :

```
. gen mstr1=cond(gener2==0 & gener3==0 & mig_1==1,1,0)

. gen mstr2=cond(gener2==1 & mig_1==1,1,0)

. gen mstr3=cond(gener3==1 & mig_1==1,1,0)
```

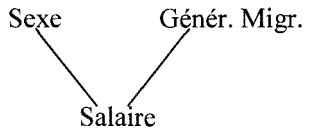
Par exemple, "mstr3" prend la valeur 1 si "gener3" prend la valeur 1 et "mig_1" prend la valeur 1 ; sinon "mstr3" prend la valeur 0. On pourrait écrire de la même façon :

```
. gen mstr1=gener2==0 & gener3==0 & mig_1==1

. gen mstr2=gener2==1 & mig_1==1

. gen mstr3=gener3==1 & mig_1==1
```

ou encore :

```
. gen mstr1=(1-gener2-gener3)*mig_1

. gen mstr2=gener2*mig_1

. gen mstr3=gener3*mig_1
```

La régression tenant compte de l'interaction entre statut migratoire et génération, en plus de l'effet propre du sexe peut être formalisée par la relation causale :



Elle est estimée par :

```
. fit salaire femme gener2 gener3 mstr1 mstr2 mstr3

  Source |       SS       df       MS                  Number of obs =     501
---------+------------------------------                F(  6,   494) =    5.35
   Model |  3.2900e+11     6  5.4834e+10                Prob > F      =  0.0000
Residual |  5.0597e+12   494  1.0242e+10                R-square      =  0.0611
---------+------------------------------                Adj R-square  =  0.0496
   Total |  5.3887e+12   500  1.0777e+10                Root MSE      =  1.0e+05

------------------------------------------------------------------------------
 salaire |      Coef.   Std. Err.       t     P>|t|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
   femme |  -21253.05   14660.78    -1.450   0.148      -50058.21    7552.117
  gener2 |   12227.14   15117.97     0.809   0.419      -17476.31    41930.59
  gener3 |  -28074.74   15867.54    -1.769   0.077      -59250.92    3101.45
   mstr1 |   -40117.7   16713.84    -2.400   0.017      -72956.69   -7278.71
   mstr2 |  -36869.46   14606.81    -2.524   0.012       -65568.6   -8170.329
   mstr3 |  -35098.89   17595.85    -1.995   0.047      -69670.83   -526.9538
   _cons |   120511.6   12112.52     9.949   0.000       96713.16     144310
------------------------------------------------------------------------------
```

À partir des résultats de la régression, le salaire pour un homme non migrant de la strate 2 (générations 1945-54) sera simplement calculé à l'aide du coefficient de la variable "gener2", qui mesure la différence moyenne entre son salaire et celui d'un non migrant de la strate 1 (générations 1930-44) :

```
. display _coef[_cons] + _coef[gener2]
132738.7
```

Pour l'homme migrant de la strate 2, le salaire sera égal à :

```
. display _coef[_cons] + _coef[gener2] + _coef[mstr2]
95869.236
```

Une autre seconde méthode consiste à produire directement les résultats pour chaque catégorie résultant du croisement des deux variables. On crée une nouvelle variable comme suit :

```
. gen strm=(mig_1 * 10) + strate
```

On peut vérifier le résultat obtenu par cette commande en faisant :

```
. sort migrant

. by migrant : tab strm strate

-> migrant=     oui
           | strate
      strm |    G30-44     G45-54     G55-64 |     Total
-----------+---------------------------------+----------
        11 |         77          0          0 |        77
        12 |          0         78          0 |        78
        13 |          0          0         49 |        49
-----------+---------------------------------+----------
     Total |         77         78         49 |       204


-> migrant=     non
           | strate
      strm |    G30-44     G45-54     G55-64 |     Total
-----------+---------------------------------+----------
         1 |         70          0          0 |        70
         2 |          0        125          0 |       125
         3 |          0          0        102 |       102
-----------+---------------------------------+----------
     Total |         70        125        102 |       297
```

La nouvelle variable "strm" servira à créer les variables dichotomiques comme suit :

```
. tab strm, gen(strm_)
      strm |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |          70       13.97       13.97
         2 |         125       24.95       38.92
         3 |         102       20.36       59.28
        11 |          77       15.37       74.65
        12 |          78       15.57       90.22
        13 |          49        9.78      100.00
-----------+-----------------------------------
     Total |         501      100.00
```

Comme pour les autres variables, on supprimera la catégorie de référence (non migrant de la génération 1930-44) :

```
. drop strm_1
```

Enfin, on pourra vérifier que les résultats produits par les deux méthodes sont équivalents, en calculant par exemple le différentiel de salaire d'un migrant de la génération 1945-54 :

```
. display _coef[gener2] + _coef[mstr2]
-24642.324
```

et en le comparant au coefficient ("strm_5") pour le nouveau modèle :

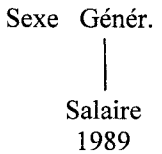
```
. fit salaire femme strm_*

  Source |       SS       df       MS                  Number of obs =     501
---------+------------------------------                F(  6,   494) =    5.35
   Model |  3.2900e+11     6  5.4834e+10                Prob > F      =  0.0000
Residual |  5.0597e+12   494  1.0242e+10                R-square      =  0.0611
---------+------------------------------                Adj R-square  =  0.0496
   Total |  5.3887e+12   500  1.0777e+10                Root MSE      =  1.0e+05

------------------------------------------------------------------------------
 salaire |      Coef.   Std. Err.       t     P>|t|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
   femme |  -21253.05   14660.78    -1.450   0.148      -50058.21    7552.117
  strm_2 |   12227.14   15117.97     0.809   0.419      -17476.31    41930.59
  strm_3 |  -28074.74   15867.54    -1.769   0.077      -59250.92    3101.45
  strm_4 |   -40117.7   16713.84    -2.400   0.017      -72956.69   -7278.71
  strm_5 |  -24642.32    16685.2    -1.477   0.140      -57425.02    8140.376
  strm_6 |  -63173.63   19037.73    -3.318   0.001      -100578.5   -25768.72
   _cons |   120511.6   12112.52     9.949   0.000       96713.16     144310
------------------------------------------------------------------------------
```

Pour juger l'adéquation du modèle, on calcule le montant du salaire ajusté pour chaque catégorie définie par le croisement des trois variables indépendantes :

```
. drop salpred

. predict salpred

. sort v304

. by v304 : tab migrant strate, summ(salpred) nost

-> v304=masculin
                Means and Frequencies of salpred
           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+------------------------------------+----------
       oui | 80393.859  95869.234   57337.93 |  81538.34
           |        73         70         38 |       181
-----------+------------------------------------+----------
       non | 120511.56   132738.7   92436.82 |  117117.59
           |        67        115         82 |       264
-----------+------------------------------------+----------
     Total | 99593.046  118788.09  81322.172 |  102646.03
           |       140        185        120 |       445

-> v304=feminin
                Means and Frequencies of salpred
           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+------------------------------------+----------
       oui | 59140.813  74616.188  36084.883 |  53496.803
           |         4          8         11 |        23
-----------+------------------------------------+----------
       non | 99258.508  111485.65  71183.773 |  85948.711
           |         3         10         20 |        33
-----------+------------------------------------+----------
     Total |  76334.11  95099.221  58729.328 |  72620.249
           |         7         18         31 |        56
```

```
------------+-----------------------------------
          1 |         67       13.37       13.37
          2 |        115       22.95       36.33
          3 |         82       16.37       52.69
         11 |         73       14.57       67.27
         12 |         70       13.97       81.24
         13 |         38        7.58       88.82
        101 |          3        0.60       89.42
        102 |         10        2.00       91.42
        103 |         20        3.99       95.41
        111 |          4        0.80       96.21
        112 |          8        1.60       97.80
        113 |         11        2.20      100.00
------------+-----------------------------------
      Total |        501      100.00
. drop strmf_1
```

Les hommes non migrants des générations 1930-44 constituent alors la catégorie de référence. Le modèle peut être formalisé par la relation causale :




```
. fit salaire strmf_*

  Source |       SS       df       MS                  Number of obs =     501
---------+------------------------------                F( 11,   489) =    3.10
   Model |  3.5167e+11     11  3.1970e+10                Prob > F      =  0.0005
Residual |  5.0370e+12    489  1.0301e+10                R-square      =  0.0653
---------+------------------------------                Adj R-square  =  0.0442
   Total |  5.3887e+12    500  1.0777e+10                Root MSE      =  1.0e+05
------------------------------------------------------------------------------
 salaire |      Coef.   Std. Err.       t     P>|t|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
 strmf_2 |   16435.99   15598.45     1.054   0.293     -14212.27    47084.25
 strmf_3 |  -30342.04   16714.02    -1.815   0.070      -63182.2    2498.115
 strmf_4 |  -38859.72   17171.07    -2.263   0.024     -72597.89   -5121.537
 strmf_5 |    -21850    17346.27    -1.260   0.208     -55932.41    12232.41
 strmf_6 |  -64102.35   20610.93    -3.110   0.002     -104599.3   -23605.44
 strmf_7 |   7591.294   59893.97     0.127   0.899       -110090    125272.6
 strmf_8 |  -41975.37   34406.47    -1.220   0.223     -109578.1    25627.38
 strmf_9 |  -33727.27   25860.65    -1.304   0.193     -84538.97    17084.42
strmf_10 |  -60532.37   52238.89    -1.159   0.247     -163172.7       42108
strmf_11 |  -58275.37   37964.73    -1.535   0.125     -132869.5    16318.76
strmf_12 |  -75711.74   33017.61    -2.293   0.022     -140585.6   -10837.84
   _cons |   119275.4   12399.23     9.620   0.000      94913.03    143637.7
------------------------------------------------------------------------------
```

Là encore, on peut comparer les salaires réels aux salaires ajustés, pour constater que cette fois-ci, il n'y a pas de différence :

```
. drop salpred

. predict salpred

. sort v304

. by v304 : tab migrant strate, summ(salpred) nostandard

-> v304=masculin
```

Là encore, on peut comparer les salaires réels aux salaires ajustés, pour constater que cette fois-ci, il n'y a pas de différence :

```
. drop salpred

. predict salpred

. sort v304

. by v304 : tab migrant strate, summ(salpred) nostandard

-> v304=masculin
                    Means and Frequencies of salpred
           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+------------------------------------+----------
       oui | 80415.656  97425.375  55173.027 | 81694.443
           |        73         70         38 |       181
-----------+------------------------------------+----------
       non | 119275.38  135711.36  88933.328 | 117010.57
           |        67        115         82 |       264
-----------+------------------------------------+----------
     Total | 99012.807  121224.77  78242.566 | 102646.03
           |       140        185        120 |       445

-> v304=feminin
                    Means and Frequencies of salpred
           | strate
   migrant |    G30-44     G45-54     G55-64      Total
-----------+------------------------------------+----------
       oui |     58743      61000  43563.637 | 52268.348
           |         4          8         11 |        23
-----------+------------------------------------+----------
       non | 126866.66      77300  85548.102 |  86804.91
           |         3         10         20 |        33
-----------+------------------------------------+----------
     Total | 87938.856  70055.556  70650.388 |  72620.25
           |         7         18         31 |        56
```

Il y a une correspondance exacte entre un tableau croisé de plusieurs variables et le modèle où interviennent les interactions de ces mêmes variables. Aboutir à un tel résultat n'a pas beaucoup d'intérêt du point de vue de l'analyse, puisque les variables n'ont plus d'effet propre. Elles n'agissent plus indépendamment les uns des autres mais en interaction les uns avec les autres.

Le modèle de régression est fondé sur l'**hypothèse d'indépendance des variables explicatives.** Pour se conformer à cette hypothèse, il faudrait vérifier que chaque variable introduite dans le modèle est bien indépendante des autres. Ce n'est pas toujours possible, notamment lorsque l'échantillon est de taille trop réduite pour constituer des catégories croisées. On imagine aisément que tester toutes les interactions est extrêmement fastidieux, et d'ailleurs, il est très rare de voir pratiquer ces tests, même dans les publications les plus sérieuses. Pourtant, **la seule manière de voir si les variables explicatives sont véritablement indépendantes entre elles, ce qui justifie la phrase rituelle "toutes choses égales par ailleurs", est de tester leurs interactions.**

Tester les interactions peut mener assez loin. Il n'y a pas en effet que les interactions de variables deux à deux, mais les interactions de rang supérieur à trois aussi. On a vu ce que donne l'évaluation de l'interaction de trois variables, mais on pourrait le faire pour quatre variables ou plus !

Par ailleurs, lorsqu'on teste les interactions, on le fait souvent seulement entre les variables sélectionnées à l'issue de plusieurs régressions. Or, une variable qui n'est pas significative peut avoir une interaction significative avec une autre variable. Théoriquement même, l'interaction entre deux variables peut très bien être significative même si aucune des deux n'a un effet significatif propre. C'est là que doit jouer la connaissance *a priori* du phénomène. Il ne suffit pas de faire confiance seulement à une pratique statistique routinière (tester systématiquement les variables, puis leur interactions deux à deux, etc.), car la meilleure sélection des variables à introduire dans un modèle reste celle que l'on fait en connaissance de cause. Après tout, une variable non significative peut très bien avoir été mal saisie lors de l'enquête, tandis qu'une variable significative peut très bien être une approximation d'une autre variable explicative, plus pertinente, qui n'a pas été saisie dans l'enquête. Comme on l'a dit au deuxième chapitre, **le calcul statistique ne peut se passer d'une évaluation qualitative**.

EXERCICE 1

Nous laissons au lecteur le soin de vérifier à partir du fichier des salariés, les différentes possibilités d'interactions. C'est un bon exercice d'essayer de trouver une explication à l'effet (ou à l'absence d'effet) des variables que l'on teste soit séparément, soit en interaction.

```
gen strf=(femme*10) + strate
tab strf, gen(strf_)
drop strf_1
fit salaire strf_* mig_1
drop salpred
predict salpred
sort v304
by v304 : tab migrant strate, summ(salpred) nostandard
gen mf=(femme*10) + mig_1
tab mf, gen(mf_)
drop mf_1
fit salaire mf_* gener2 gener3
drop salpred
predict salpred
sort v304
by v304 : tab migrant strate, summ(salpred) nostandard
```

Comparez les résultats ci-dessus avec les résultats commentés précédemment dans le manuel. Quel est, à votre avis, le modèle qui ajuste le mieux les salaires ? Pourquoi le salaire des femmes est-il moins bien ajusté ?

EXERCICE 2

On reste fasciné par notre capacité à trouver des explications à tout ce qu'on obtient d'un modèle statistique, même lorsqu'on s'est trompé. Il arrive, par exemple, qu'une variable ait été créée maladroitement, conduisant à des modalités erronées : or, dans ce cas, il est bien rare que l'on n'arrive pas quand même à trouver une explication de l'effet de cette "fausse" variable. Un test consiste à créer une variable dichotomique "bidon" (à partir d'une fonction aléatoire) qui ne correspond à rien dans les données et à la tester dans le modèle :

```
gen bidon=round(uniform(),1)
fit salaire gener2 gener3 femme mig_1 bidon
```

Répétez cette opération plusieurs fois, et remarquez comme les coefficients et leur degré de significativité peuvent varier. On serait bien en peine pourtant d'interpréter l'effet de cette variable purement aléatoire !

d) Approche exploratoire versus approche confirmatoire

Deux méthodes de sélection du modèle peuvent être envisagées, dont nous allons présenter les avantages et inconvénients. **L'approche confirmatoire correspond à l'approche hypothético-déductive et l'approche exploratoire à l'approche inductive**, dont nous avons parlé dans le premier chapitre. L'une possède les inconvénients de la théorisation, et l'autre de l'empirisme.

Pour tester de nombreuses variables avec de petits échantillons, comme c'est souvent le cas en sciences sociales, on peut utiliser une procédure dite exploratoire ou inductive. Les variables du modèle sont sélectionnées par élimination, de manière à obtenir une **forme réduite du modèle**, à la différence du modèle complet où toutes les variables indépendantes interviennent dans l'équation. La procédure de sélection est la suivante :

1. les quelques variables *a priori* les plus pertinentes sont d'abord testées ;

2. pour chaque variable, les modalités dont les coefficients ne sont pas significatifs sont supprimées, et d'autres variables sont introduites dans le modèle ; cette étape est réitérée plusieurs fois jusqu'à ce que toutes les variables soient testées au moins une fois ;

3. après plusieurs essais, on obtient une sélection des modalités des variables les plus pertinentes : pour être sûr de n'avoir pas omis une modalité éliminée dans une première étape de notre sélection, on teste à nouveau toutes les modalités éliminées, variable par variable, jusqu'à ne sélectionner que les modalités les plus pertinentes, dont les coefficients sont significatifs.

Notez que si le nombre de modalités des variables n'est pas trop important par rapport à l'effectif de l'échantillon, il n'est pas nécessaire de faire plusieurs essais pour la sélection des modalités significatives : il suffit alors d'introduire toutes les modalités possibles dans le modèle, de manière à tester directement le modèle complet. Dans l'exemple de ce chapitre, cela ne serait pas possible car la taille de l'échantillon n'est pas assez importante.

Cette procédure peut être qualifiée d'exploratoire, dans le sens où elle ne confronte pas un modèle particulier, avec des variables sélectionnées selon des hypothèses précises, avec un autre modèle constitué selon d'autres hypothèses. Cette deuxième procédure est appelée "confirmatoire" par opposition à la première. Les deux méthodes ont des inconvénients que l'on va passer en revue.

Le risque majeur avec la méthode exploratoire est de procéder en aveugle, sans vraiment se préoccuper des variables qui font sens du point de vue de la causalité (ou du système de causalité) que l'on essaie de mettre en lumière. À la limite, on peut sélectionner les modalités les plus significatives, sans même savoir si elles sont signifiantes. **Pour éviter cela, le chercheur doit d'abord sélectionner des variables à tester selon leur pertinence du point de vue d'un cadre théorique (déduction)**. Ces variables sont déjà considérées comme le résultat d'une sélection opérée depuis l'élaboration du questionnaire. Le chercheur, par une **procédure systématique**, veut vérifier la pertinence de cette sélection par le degré de significativité de chaque variable.

Les **procédures automatiques** "*stepwise*" (par étapes) ont été imaginées pour simplifier ce travail de sélection des variables. Elles se distinguent cependant très nettement des procédures systématiques telles que celle que nous avons exposées plus haut. Une procédure automatique de sélection n'est pas souhaitable dans la mesure où elle comporte une part importante d'arbitraire (qui permet justement l'automatisme). En effet, une procédure automatique ne permet pas de réfléchir après chaque sélection pour procéder éventuellement à des regroupements de modalités proches du point de vue de leur signification, ou aux tests d'interactions entre variables.

Selon la procédure confirmatoire, le cadre théorique est d'abord défini, et plusieurs modèles sont confrontés pour confirmer ou infirmer une hypothèse (approche hypothético-déductive). On n'introduit aucune sélection automatique des variables. Seulement, doit-on s'interdire toute exploration ? Dans tout cadre théorique, il y a des zones d'ombres, une part d'incertitude, que l'outil statistique peut aider à lever (induction). Le but est en définitive de sélectionner le meilleur modèle possible. Une procédure par essais et erreurs peut aboutir à un modèle qui explique mieux le phénomène que l'on étudie : que signifierait une recherche qui ne mène qu'à confirmer des résultats attendus selon un cadre pré-établi ? La recherche peut être fortement stimulée par des résultats inattendus : le cadre théorique doit

pouvoir intégrer de nouveaux éléments, et ce sont ces nouveaux éléments qui peuvent éventuellement le remettre en cause.

Les cadres théoriques sont construits à partir d'analyses qui nécessairement vieillissent : il y a souvent un décalage entre les cadres théoriques, qui évoluent par bonds qualitatifs, et les techniques d'analyses, qui elles, évoluent plus linéairement. Or, ces techniques peuvent susciter un renouvellement des approches et déclencher ainsi un bond dans le cadre théorique : dans ce cas, l'analyse exploratoire est fondamentale, car elle suscite de nouvelles interrogations, une nouvelle interprétation du réel.

En définitive, nous pensons qu'il y a toujours une part confirmatoire dans l'approche exploratoire, et *vice versa*. La différence entre les deux approches réside essentiellement dans la présentation des résultats finals (comparaison des modèles théoriques, ou présentation du modèle réduit optimal), mais dans la pratique de la recherche, il se produit plus souvent un va-et-vient continuel entre les deux approches.

Pour notre part, nous préférons utiliser, après avoir sélectionné *a priori* les variables les plus pertinentes d'un point de vue théorique, une approche exploratoire, essentiellement pour des raisons pratiques : si nous disposions de données numériquement importantes, nous pourrions à la limite introduire toutes les variables disponibles dans une seule régression et présenter les résultats tels quels. Or nous disposons la plupart du temps d'un échantillon de taille limité qui nous interdit une telle procédure radicale. Pour tester, "toutes choses égales par ailleurs", une des variables qui nous intéresse, il nous faut nécessairement aller par étapes en sélectionnant les autres variables les plus pertinentes. On procède ainsi "à l'économie" en utilisant la procédure décrite au début de cette section.

3. L'ajustement et son diagnostic

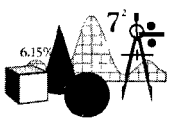
Pour simplifier l'exposé qui va suivre, seules les variables indépendantes sexe, statut migratoire et génération ont été retenues, sans interaction. On devrait bien entendu faire intervenir d'autres variables pour améliorer le modèle. On pense par exemple au niveau d'instruction, à l'origine ethnique ou géographique, au type d'emploi salarié, etc. Ces variables peuvent être aisément créées sous forme dichotomique ou polytomique, avec la commande « tabulate *nom_variable*, generate(*nouvelle_variable*) ». On peut aussi créer des variables indépendantes de type continu, telle que le nombre d'années d'ancienneté dans le salariat. En fin de chapitre, le lecteur trouvera quelques exercices qui l'aideront à tester sa compréhension des mécanismes de la régression statistique.

a) Les statistiques de diagnostic global

Supposons que nous ayons sélectionné au mieux les modalités des variables, c'est-à-dire celles qui sont non seulement pertinentes mais significatives aussi, Comment allons-nous faire pour prouver que le modèle est bon ?

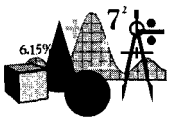
Que signifie d'abord sélectionner un bon modèle ? **Un modèle où toutes les variables sélectionnées sont significatives n'est pas nécessairement un bon modèle.** Il faut pour le vérifier, faire l'analyse de l'ajustement du modèle à nos données. Rappelons les résultats du modèle sans interaction :

```
. fit salaire gener2 gener3 femme mig_1

  Source |       SS       df       MS                  Number of obs =     501
---------+------------------------------                F(  4,   496) =    8.05
   Model |  3.2854e+11     4  8.2135e+10                Prob > F      =  0.0000
Residual |  5.0601e+12   496  1.0202e+10                R-square      =  0.0610
---------+------------------------------                Adj R-square  =  0.0534
   Total |  5.3887e+12   500  1.0777e+10                Root MSE      =  1.0e+05

------------------------------------------------------------------------------
 salaire |      Coef.   Std. Err.       t     P>|t|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   13855.27   11034.27     1.256   0.210      -7824.401    35534.93
  gener3 |  -25907.17   12082.96    -2.144   0.033      -49647.27   -2167.061
   femme |  -21215.43   14630.77    -1.450   0.148      -49961.36    7530.508
   mig_1 |   -37383.8   9310.886    -4.015   0.000      -55677.44   -19090.16
   _cons |   119077.7   9667.409    12.317   0.000       100083.6    138071.8
------------------------------------------------------------------------------
```



Une des manières les plus simples de voir l'adéquation du modèle aux données est d'analyser la différence entre la variance expliquée par le modèle et la variance résiduelle (qui n'est pas expliquée par le modèle). Un coefficient, le R^2 (*R-square* ou sa mesure ajustée qu'on lui préfère généralement), est particulièrement utile dans ce cas : il mesure le **pourcentage de la variance expliquée par le modèle.** Dans l'exemple ci-dessus, 5,34 % de la variance du salaire (*adjusted R-square*) est expliquée par le modèle.

En sciences sociales, la variance résiduelle (non expliquée par le modèle) est importante, et ce n'est pas à notre avantage : à considérer seulement ce résultat, on en viendrait rapidement à jeter à la poubelle toutes nos données. Heureusement, cette mesure synthétique n'est en fait pas le seul critère pour juger la qualité du modèle. L'importance des résidus peut très bien venir du fait que l'on a de nombreuses observations en comparaison du nombre de variables que l'on a pu tester. De plus, nos données peuvent contenir des erreurs qui ne peuvent être bien expliquées par le modèle et qui contribuent ainsi fortement à la variance totale. En somme, **une variance résiduelle importante ne signifie pas que les variables sélectionnées ne soient pas pertinentes.**



Généralement, on s'attache d'abord à **savoir si la variance expliquée par le modèle est suffisante au regard du nombre de variables que l'on a introduit dans le modèle**, à l'aide d'un test appelé ***F-test***. Ce test est souvent plus rassurant que le pourcentage de la variance expliquée. Dans notre exemple, on a une probabilité inférieure à 1 pour mille de se tromper (ce qui est indiqué par Prob>*F*=0,0000 dans le tableau de résultats) en disant que les variables du modèle sont significatives : en effet, la valeur du test pour 4 variables et 496 degrés de liberté est de $F(4;496)=8,05$, ce qui correspond à une probabilité inférieure à 0,00005. Mais on peut souvent deviner le résultat de ce test du simple fait que certaines modalités des variables indépendantes sont significatives.

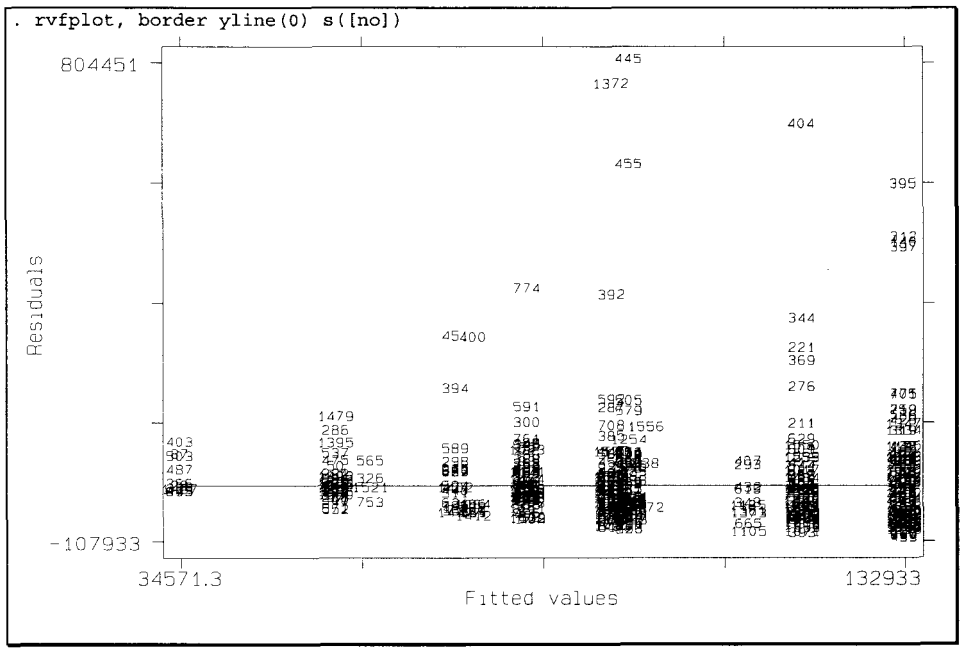
Les méthodes graphiques sont en fait plus utiles pour juger l'adéquation du modèle aux données. Dans la suite de ce chapitre, nous passerons d'abord en revue les différentes commandes graphiques que propose STATA, afin de donner au lecteur une approche intuitive du phénomène. Par la suite, les différents tests de dispersion seront expliqués.

b) Repérer les observations aberrantes : « rvfplot »

Il faut s'imaginer la régression comme la construction d'une droite au travers d'un nuage de points qui représentent chacun des individus de l'échantillon. Le calcul aboutit à minimiser la distance entre chacun des points et la droite.

Les observations aberrantes sont celles qui se situent loin de la droite de régression. Leur éloignement est mesuré par les **résidus**, c'est-à-dire par **l'écart entre les valeurs de la variable dépendante pour chaque observation de l'échantillon et les valeurs prédites par le modèle pour ces mêmes observations**. Ces résidus peuvent être dispersés aléatoirement sur l'ensemble de l'échantillon (auquel cas le modèle donne une assez bonne mesure, en moyenne, de l'effet de chaque variable) ou bien concentrés, de façon linéaire ou non, sur certaines observations : la valeur de la variable dépendante pour ces observations n'est donc pas bien expliquée par le modèle. On cherchera à repérer les observations les plus aberrantes (par rapport au modèle), c'est-à-dire celles qui concentrent en elles la plus grande part de la variance résiduelle.

La commande graphique « rvfplot » permet de repérer les observations aberrantes. Après la régression dont les résultats figurent dans la section précédente, on a exécuté cette commande :



À partir de ce graphique, on peut déjà constater que de nombreuses observations (1 372, 445, 404, etc.) se trouvent assez loin de la droite y=0, pour laquelle les résidus sont nuls. On voit aussi que les résidus sont plus souvent positifs que négatifs, et ceci particulièrement lorsque les valeurs ajustées (*fitted values*) augmentent.

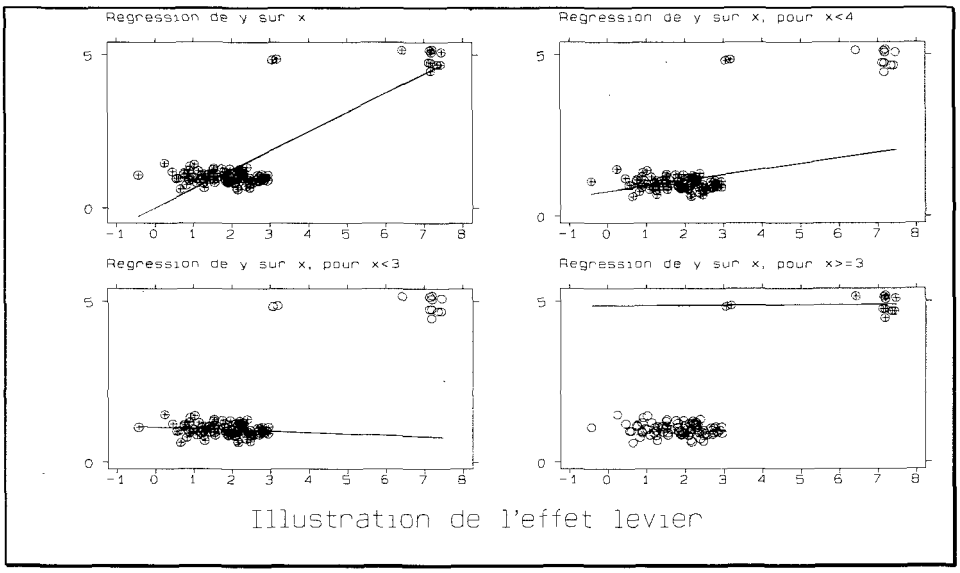
Ce résultat indique déjà, à propos de ce modèle très simple, que la troisième hypothèse de base de la régression linéaire (variance unique pour tout l'échantillon) n'est pas vérifiée ici. En effet, les résidus (la différence entre les valeurs observées et ajustées, c'est-à-dire l'erreur résultant de la régression) doivent être distribués aléatoirement autour de la droite de régression. Or, le graphique qui précède montre que ce n'est manifestement pas le cas : l'erreur est dite *hétéroscédastique* selon le salaire, c'est-à-dire que l'erreur croît avec la valeur ajustée du salaire. Dans cet exemple, le modèle de régression ne s'adapte pas aux salaires élevés, probablement parce que la génération, le sexe et le statut migratoire ne suffisent pas à expliquer le niveau de salaire.

c) Repérer les observations extrêmes : « lvr2plot »

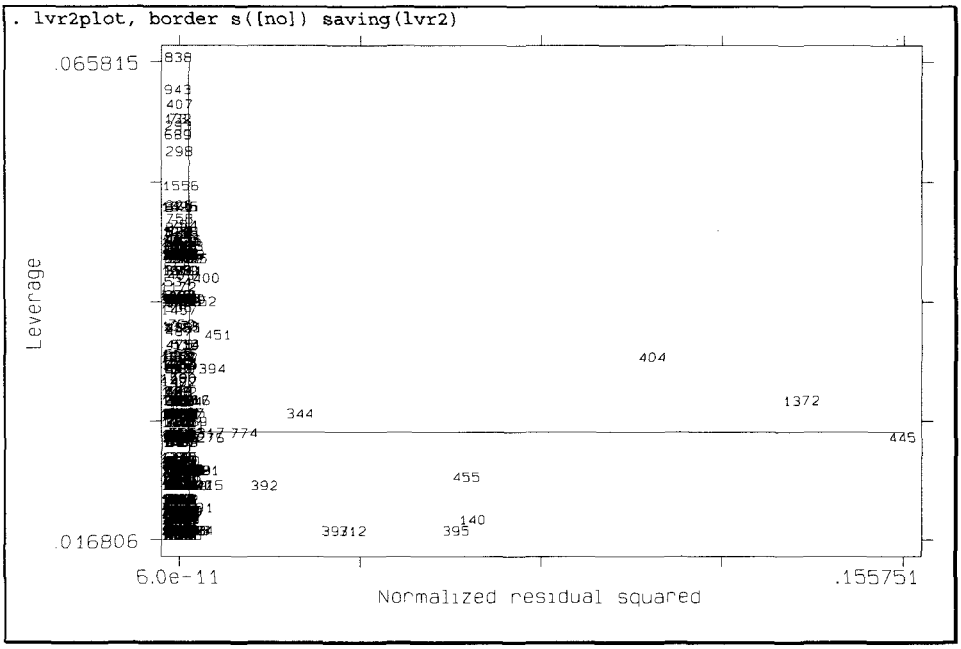
On appelle observations extrêmes celles qui sont éloignées de la masse des observations sans pour autant être éloignées de la droite de régression : ces observations peuvent avoir un "**effet levier**" (en anglais *leverage*), de sorte que leur

éloignement du reste des observations a une forte influence sur la pente de la droite de régression qui tendra à passer par ces points, même s'ils sont peu nombreux.

Ce phénomène est plus facilement compréhensible à l'aide d'un exemple artificiel sur deux variables x et y : un amas de points en bas à gauche du premier graphique ("Régression de y sur x"), et quelques points en haut à droite du même graphique, font passer la droite de régression à travers l'amas de points et les quelques points éloignés. Si on supprime les points éloignées en haut à droite ("Régression de y sur x, pour x<4"), la droite de régression n'aura plus la même allure. D'une manière générale, la droite de régression prendra la forme du nuage de points restant, comme on le voit en comparant les résultats des régressions pour une sélection différente des valeurs de x ("Régression de y sur x, pour x<3, et pour x>=3"). Cette illustration est aisée dans le cas où seulement deux variables sont en jeu, mais elle s'applique aussi avec des modèles plus complexes.



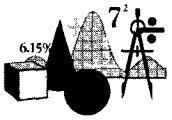
Pour une régression qui fait intervenir plusieurs variables simultanément, un coefficient peut être calculé pour mesurer l'importance globale de l'effet de levier. Comme les observations aberrantes peuvent aussi être repérées par un coefficient normalisé, on peut mettre en rapport les deux mesures, ce que la commande « lvr2plot » fait très bien pour nous :



Sur ce graphique, les points au-dessus de la ligne horizontale $y=0$ ont un effet levier plus important que la moyenne, tandis que les points éloignés de la verticale $x=0$ ont des résidus plus élevés que la moyenne. On voit clairement que ce ne sont pas les mêmes observations qui sont concernées (auquel cas on aurait des observations en haut à droite du graphique).

d) Tests statistiques : « fpredict », « outlev »

Les commandes graphiques « rvfplot » et « lvr2plot », et d'autres encore que nous laissons au lecteur le soin de découvrir dans le manuel de STATA (voir la section [5s] fit), sont basées sur le calcul préalable d'un certain nombre de statistiques. Le but étant de repérer certaines observations (aberrantes, extrêmes ou influentes), il n'est généralement pas nécessaire de conserver ces statistiques : il suffit de repérer l'identifiant de ces observations, et constater l'importance de l'écart entre les valeurs observées et ajustées.



Si les graphiques ne suffisent pas, il est toujours possible d'utiliser les options de la commande « fpredict » pour obtenir des résultats détaillés pour chaque individu de l'échantillon. En effet, pour résumer l'influence d'une observation, les options « dfits », « cooksd » et « welsch » calculent trois indices (DFITS, distance de Cook, distance de Welsh) qui sont affectés par l'importance des

résidus (valeurs aberrantes) en même temps que par l'importance de l'effet de levier (valeurs extrêmes). On peut en outre calculer un seuil pour chacun de ces trois indices qui permet de repérer systématiquement les observations influentes.

Une autre mesure de l'influence de chaque observation est calculée par l'option « covratio » de la commande « fpredict », qui compare les estimations de la matrice des covariances avec ou sans cette observation. De la même façon que pour les autres mesures d'influences, on peut fixer un seuil pour le repérage des observations les plus influentes.

Étant donné que le repérage des observations à partir de ces statistiques peut se faire de façon assez mécanique (on fait la liste des observations pour lesquelles la statistique dépasse un certain seuil), nous avons écrit une commande qui fait directement ces calculs « outlev *ident*, [dfbeta(*liste_variables*)] » où *ident* représente l'identifiant de l'individu. Les seuils pour repérer systématiquement les observations influentes ont été choisis selon les recommandations faîtes dans la section "[5s] fit" du manuel de STATA. L'option « dfbeta » produit des résultats plus longs, puisqu'elle permet de repérer l'influence de chaque observation (par comparaison des coefficients avec et sans l'observation) sur le calcul des coefficients pour les variables indépendantes choisies dans la liste des variables *(liste_variables)* :

```
. outlev no, dfbeta(gener2 gener3 femme mig_1)
```

(Listing omis)

4. Conclusions sur la régression statistique

Comme on vient de le voir, il est bien rare que l'ajustement du modèle soit satisfaisant : les observations extrêmes ne sont pas forcément un inconvénient (même s'il y a lieu de vérifier les données sur ces observations pour voir si elles ne recèlent pas des erreurs), mais en revanche, les observations aberrantes sont toujours problématiques, surtout si elles sont en même temps extrêmes.

Une première attitude consiste à condamner le modèle : les observations aberrantes sont la preuve que les hypothèses du modèle ne sont pas vérifiées. On peut très bien considérer que les données sont fiables (ou même exactes) et que le but de l'analyse est de vérifier un modèle spécifique, en l'occurrence, un modèle linéaire. De ce point de vue, l'inadéquation entre le modèle et la réalité (des données) suggère peut-être d'utiliser d'autres techniques d'ajustement, par exemple un modèle non linéaire.

Cependant, il est bien rare, en sciences sociales, que l'on puisse considérer les données comme absolument exactes. Il est bien rare aussi d'appliquer un modèle

seulement pour en vérifier la validité (hypothèse de linéarité, par exemple). La rigueur de l'analyse statistique ne nous dispense pas d'être souple, et plus modeste aussi. **Ce que nous cherchons généralement à obtenir, c'est une idée de l'effet d'une variable sur une autre, même pas forcément la mesure exacte de cet effet, mais simplement son importance, sa direction, relativement à d'autres variables.**

Par conséquent, il est souhaitable d'essayer plusieurs chaussures avant de choisir la bonne paire. La bonne paire, ici, sera une combinaison d'un bon échantillon et d'un bon modèle (ou du moins mauvais qui soit). On tentera de constituer un "bon" échantillon en réduisant l'influence des observations aberrantes sur l'estimation. On peut par exemple éliminer les observations aberrantes et relancer le modèle sur le reste des observations.

Choisir un bon modèle adapté à l'échantillon, est plus difficile. Ce manuel ayant pour objectif d'expliquer les techniques d'analyse des biographies, il ne sera pas fait ici d'exposé sur des techniques de régression complexes sur des échantillons où le temps n'intervient pas. Signalons cependant au lecteur que des techniques d'**estimations robustes** existent dans STATA : « rreg », pour "régression robuste", « bsqreg » pour "régression sur les quantiles" tels que la médiane (voir sections [5s] rreg et [5s] bstrap du manuel de STATA). Ces estimations sont sensibles aux variations autour de la moyenne (ou plus généralement près du centre de la distribution, là où se situent la plupart des observations) mais peu sensibles aux variations loin de cette moyenne (observations aberrantes).

EXERCICE 3

Le modèle utilisé dans cette section teste l'effet de trois variables seulement. Comment doit-on introduire dans le modèle le niveau d'instruction ? Sous la forme d'une variable continue ou discrète (polytomique) ?

Après avoir fait votre choix, testez le modèle et diagnostiquez-le avec les commandes graphiques pour identifier les observations les plus influentes. Testez à nouveau le modèle en excluant ces observations avec l'option « if no !=... & no !=... ».

Les procédures de régression robustes permettent d'accorder moins d'importance aux observations aberrantes. Testez le modèle avec les commandes « rreg » et « bsqreg ». Comparez les résultats avec ceux de la commande « fit » avec et sans les observations aberrantes.

CHAPITRE 4

LE MODÈLE DE RÉGRESSION LOGISTIQUE

L'utilisation des modèles logistiques est courante dans l'analyse des biographies, mais est en même temps foncièrement ambiguë. L'acquisition d'une caractéristique (événement qui survient au cours de la vie de l'individu) est souvent confondue avec la présence d'une caractéristique dans la population, de sorte que le modèle logistique est souvent utilisé abusivement pour l'analyse des biographies.

Néanmoins, le modèle logistique est utile pour comprendre les principes du modèle de régression à risques proportionnels, et en cela constitue une étape indispensable vers le modèle de Cox. Sa large diffusion dans les milieux scientifiques justifie aussi que l'on s'y attarde.

Le modèle logistique a aussi une variante fort intéressante, le modèle multinomial "logit". Lorsque la caractéristique est multiple, c'est-à-dire lorsqu'elle présente plusieurs modalités, on peut calculer les probabilités relatives de rencontrer l'une ou l'autre de ces modalités si elles sont exclusives entre elles.

1. L'utilisation d'une variable dépendante dichotomique en régression

L'idée à la base des modèles de type logistique est de faire l'analyse, non pas sur une variable dépendante continue (du type "salaire", comme on l'a vu à propos de la régression simple) mais sur une **variable dichotomique** : l'individu a ou n'a pas cette caractéristique (être salarié, par exemple). Cette variable est codée 1 ou 0 selon le cas, à la différence d'une variable continue qui peut prendre une infinité de valeurs.

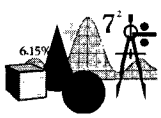
a) Le temps dans la régression logistique

Le modèle logistique s'applique parfaitement dans des situations où certains membres d'une population ont une caractéristique permanente que d'autres n'ont pas. La modélisation a pour but de déterminer les facteurs qui peuvent expliquer la présence de cette caractéristique.

Pourquoi parle-t-on de **caractéristique permanente** ? On entend par là toute caractéristique **acquise dès le moment où l'individu entre dans la population étudiée.** Ainsi, on peut étudier la fréquence des malformations génétiques chez les nouveau-nés, la proportion de ruraux parmi les immigrants, la fréquence des femmes instruites au moment de la consultation prénatale, etc. La caractéristique étudiée (malformation, origine, instruction, etc.) est définie précisément au moment de l'observation qui peut être le moment de l'enquête mais aussi un autre moment de la vie (naissance, immigration, consultation, etc.). En principe, le temps est considéré comme nul ou inopérant, comme dans un modèle de régression simple. **Le temps n'intervient pas dans la régression logistique : le moment d'observation est unique**.

b) L'équation et son interprétation

Le modèle logistique fait partie d'une classe de **modèles dits log-linéaires** qui ont tous pour but l'analyse des **ratios**, qu'ils soient exprimés sous forme logistique ou non. Les principes de calculs et la présentation des résultats sont communs à tous ces modèles, et aussi aux modèles d'analyse des biographies que nous aborderons plus tard.



La quantité qui va être modélisée par le modèle logistique n'est pas $E(Y_j) = p_j$ où p est la proportion de la population ayant la caractéristique étudiée. Dans le modèle logistique, la quantité modélisée est constituée du rapport de deux populations distinctes, celle qui a la caractéristique étudiée (en proportion p_j) et celle qui ne l'a pas $(1-p_j)$:

$$r_j = \frac{p_j}{1 - p_j}$$

Dans ce modèle, il est préférable de considérer comme variable dépendante celle qui caractérise le moins d'individus dans la population. En effet, si on considère la caractéristique la plus commune, on risque d'aboutir à un ratio tendant vers plus l'infini (dénominateur tendant vers 0) dans le cas d'une sous-catégorie de la population où la caractéristique étudiée est omni-présente. **Le modèle logistique est donc approprié pour les caractéristiques rares** dans la population. Dans ce cas, il se rapproche d'un simple modèle sur les proportions.

Il faut souligner à ce propos que le modèle logistique est souvent utilisé sans être confronté à d'autres modèles de la même classe. La comparaison des résultats de divers modèles pour un échantillon donné est la plupart du temps négligée. Le fait que le modèle logistique ait été approprié pour nombre de problèmes rencontrés jusqu'à présent ne constitue en rien une garantie scientifique, un gage de validité pour les questions à venir. Il est conseillé de tester plusieurs modèles sur les "ratios" dans la classe des modèles log-linéaires, pour chaque problème à traiter.

Un des modèles les plus proches est le modèle "probit", qui diffère simplement du modèle logistique par l'hypothèse de distribution des erreurs (*error term* : voir section [5s] logit du manuel de STATA). Mais on pense surtout aux modèles linéaires généralisés (voir section [5s] glm du manuel de STATA) dont les modèles logistique et "probit" ne sont en fait que des versions particulières. Cependant, pour expliquer les grands principes des modèles log-linéaires, nous nous en tiendrons ici au seul modèle logistique, et à une de ses variantes, le modèle multinomial "logit".

2. Les commandes « logistic » et « logit »

Le premier emploi des femmes est une application possible pour le modèle logistique. La question à laquelle on se propose de répondre est la suivante : quelle est la proportion de femmes qui, dans leur premier emploi, sont salariées ?

Notons d'abord que cette question est absolument équivalente à la question : quelle est la proportion de femmes qui sont indépendantes dans leur premier emploi ? Les deux questions sont équivalentes parce que les catégories "salariée" et "indépendante" sont exclusives et complémentaires l'une de l'autre. Ensemble, ces catégories forment les premiers emplois : si le premier emploi n'est pas un emploi salarié, c'est un emploi indépendant. Pour l'instant, nous nous intéressons aux seules femmes qui ont travaillé au moins une fois à Dakar. Nous verrons dans les chapitres consacrés aux tables de séjour comment inclure dans l'analyse les femmes qui n'avaient pas encore travaillé au moment de l'enquête.

a) La mise en forme du fichier

Avec cet exemple, nous analysons un fichier qui servira jusqu'à la fin de ce manuel : JOB.DTA.

```
. use job

. desc

Contains data from job.dta
  Obs :   964 (max=  7435)
 Vars :    38 (max=   581)
Width :    46 (max=   592)
  1. no            int     %8.0g
  2. v304          byte    %4.0g     v304      sexe
  3. v305          byte    %4.0g               mois de naissance
  4. v306          byte    %4.0g               annee de naissance
  5. v307          byte    %4.0g     v307      rang de naissance chez mere
  6. v308          byte    %4.0g     v308      rang naissance chez pere
  7. v309          byte    %4.0g     v309      lieu naissance pere
  8. v310          byte    %4.0g               region naissance pere
  9. v311          byte    %4.0g     v311      ethnie pere
 10. v319          byte    %4.0g               secteur activite pere
 11. v320          byte    %4.0g     v320      caste pere
 12. v321          byte    %4.0g     v321      lieu naissance mere
 13. v322          byte    %4.0g               region naissance mere
 14. v332          byte    %4.0g     v332      caste mere
 15. v347          byte    %4.0g     v347      conversation francais
 16. v349          byte    %4.0g     v349      religion
 17. v351          byte    %4.0g     v351      situation matrimoniale
 18. strate        byte    %4.0g     strat     No generation
 19. v601          byte    %4.0g               Numero periode
 20. v602          byte    %4.0g               Mois debut periode
 21. v603          byte    %4.0g               Annee debut periode
 22. v604          byte    %4.0g               Type activite
 23. v609          int     %8.0g               Profession
 24. v610          byte    %4.0g               Secteur
 25. v611          byte    %4.0g               Statut d'activite
 26. v614          byte    %4.0g     O_N       Comptabilite ecrite
 27. v615          byte    %4.0g               Lieu d'activite
 28. v620          byte    %4.0g     O_N       Au service d'un parent
 29. v627          byte    %4.0g     O_N       Fiche de paie
 30. v628          byte    %4.0g               Regularite de paiement
 31. v629          byte    %4.0g     O_N       Temps partiel
 32. v632          long    %9.0g               Salaire debut periode
 33. v633          long    %9.0g               Salaire fin periode
 34. v634          byte    %4.0g               Raison changement
 35. v637          byte    %4.0g               Mois fin periode
 36. v638          byte    %4.0g               Annee fin periode
 37. grprof        byte    %4.0g               Groupe professionnel
 38. instr         byte    %4.0g     instr     niveau d'instruction
Sorted by :  no  v601
```

Ce fichier contient toutes les informations sur les différentes activités professionnelles des enquêtés à partir de l'âge de 12 ans jusqu'à la date d'enquête (variables v601 à v638), ainsi que les informations sur les caractéristiques socio-démographiques des enquêtés (variables v304 à v351). Seules les femmes ont été sélectionnées et certaines variables d'intérêt secondaire ont été supprimées pour ne pas alourdir l'exposé, et pour que le fichier soit de taille raisonnable pour être utilisé sur n'importe quelle machine, quelle que soit la version de STATA.

On notera que la variable v603 marque l'année de début d'une activité professionnelle, et la variable v638 l'année de fin. Pour l'instant, les variables v604

(type d'activité) et v611 (statut dans l'activité), suffiront à qualifier la période du point de vue de l'activité professionnelle.

```
. by no : list no v306 v603 v604 v611 v638, nolab

-> no=       3
     v306  v603  v604  v611  v638
  1.   40    52     4     0    84
  2.   40    84     1     3    89

-> no=       4
     v306  v603  v604  v611  v638
  3.   58    70     4     0    89

(etc)
```

Étant donné que pour la régression logistique, on ne prendra en compte que le premier emploi rémunéré à Dakar (comme salarié ou indépendant ; les périodes d'apprentissage ne sont pas considérées comme des périodes d'emploi rémunéré), il faut sélectionner pour chaque individu l'enregistrement qui contient les informations sur cet emploi. Les périodes d'emplois sont repérées par la variable v611, qui prend alors la valeur 2 (salarié) ou 3 (indépendant). La seule difficulté consiste à repérer la première période d'emploi vécue par l'individu.

b) La création de la variable dépendante

Les observations sont classées selon le numéro d'identifiant de l'individu (no) et par numéro de périodes (v601) : les périodes traversées par l'individu se succèdent dans le fichier dans l'ordre chronologique. Une première période d'emploi est donc une période d'emploi qui ne succède à aucune autre.

On calculera, pour chaque individu, le nombre d'emplois occupés jusqu'à la période courante (variable "jobcum"), en s'assurant d'abord que les enregistrements sont bien classés par ordre chronologique :

```
. sort no v601

. by no : gen byte jobcum=sum(v611==2 | v611==3)

-> no=       3
-> no=       4
-> no=      11
-> no=      13
-> no=      15
-> no=      20
-> no=      28
-> no=      29
(etc)
```

Pour éviter l'affichage à l'écran des numéros d'identifiant, on exécutera plutôt la commande précédée de « quietly » :

```
. quietly by no : gen byte jobcum=sum(v611==2 | v611==3)
```

Il ne suffit pas de repérer les périodes pour lesquelles la variable "jobcum" prend la valeur 1, car les périodes qui succèdent à une première période d'emploi seront codées 1 s'il s'agit de période hors emploi (formation, chômage, maladie...). Il faut donc repérer lorsque la variable "jobcum" passe de 0 ("n'a pas encore connu d'emploi") à 1 ("a connu une première période d'emploi"). Ce repérage est possible grâce à une fonction d'adressage propre au langage de STATA qui permet de lire la valeur d'une variable pour un enregistrement précédant ou suivant : pour une variable "var" l'expression "var[_n]" est équivalente à "var", tandis que "var[_n-1]" correspond à la valeur de "var" pour l'enregistrement précédent, et "var[_n+1]" à la valeur de la variable pour l'enregistrement suivant. Notez au passage que l'on peut faire référence à n'importe quel autre enregistrement : le premier du fichier "var[1]", le dernier "var[_N]" (_N est le nombre total d'enregistrements dans le fichier), ou un enregistrement d'autre rang "var[123]", "var[_n-24]", "var[_n+325]", "var[_N-1]", etc.

Lorsqu'on utilise l'option « by liste_*variables* : generate ... », la valeur de _n est définie pour chaque valeur de la variable indiquée dans l'option « by ». Dans notre exemple, l'identifiant "no" sera utilisé, de sorte que "var[1]" correspond au premier enregistrement de l'individu courant, et "var[_N]" au dernier enregistrement pour ce même individu.

On peut donc calculer la valeur d'une variable de l'enregistrement courant en fonction de la valeur d'une variable de l'enregistrement précédent, et ceci pour chaque individu. Un tableau logique est utile pour envisager tous les cas possibles :

Tableau 1. Valeur de la variable "job1" "premier emploi" pour un individu :

	Enregistrement courant [_n]		
Enregistrement précédent [_n-1]	**jobcum=.**	**jobcum=0**	**jobcum>=1**
jobcum[_n-1]=.	job1=.	job1=0	job1=1
jobcum[_n-1]=0	job1=.	job1=0	job1=1
jobcum[_n-1]=1	job1=.	job1=0	job1=0

En définitive, la variable "job1", qui est calculée quand la valeur de "jobcum" est définie (non manquante), prend la valeur 1 lorsque "jobcum" est égale à 1 et "jobcum[_n-1]" est différente de 1. Sinon "job1" prend la valeur 0 :

```
. quietly by job : gen byte job1=(jobcum==1 & jobcum[_n-1]!=1) if jobcum!=.

. by no : list v601 v611 jobcum job1

-> no=       3
      v601  v611    jobcum      job1
  1.     1     0         0         0
  2.     2     3         1         1

-> no=       4
  3.     1     0         0         0

-> no=      11
  4.     1     0         0         0
  5.     2     3         1         1
(etc)
-> no=      20
      v601  v611    jobcum      job1
 12.     1     2         1         1
 13.     2     0         1         0
 14.     3     0         1         0
(etc)
-> no=      75
      v601  v611    jobcum      job1
 42.     1     3         1         0
 43.     2     3         2         0

-> no=      78
      v601  v611    jobcum      job1
 44.     1     0         0         0
 45.     2     2         1         1
 46.     3     2         2         0
 47.     4     4         2         0
 48.     5     0         2         0
 49.     6     0         2         0
 50.     7     3         3         0
 51.     8     0         3         0
(etc)
```

Un programme « firstocc » (pour *first occurrence*), dont une copie figure sur la disquette fournie avec ce manuel, et qui a été diffusé dans le numéro 22 du STATA, permet de faire ce calcul, sans créer de variable intermédiaire, en ne tapant qu'une ligne de commande :

```
. drop job1

. firstocc no v601 =(v611==2 | v611==3), gen(job1)
```

Maintenant que l'enregistrement qui contient les informations sur le premier emploi est repéré, il reste à créer la variable dépendante. Comme on l'a dit plus haut, le modèle logistique est plus approprié pour l'analyse d'une caractéristique rare dans la population. Chez les femmes de notre échantillon, le salariat est moins fréquent que le travail indépendant : on va donc créer une variable codée 1 pour une période de salariat et 0 pour une période de travail indépendant, dans le cas où il s'agit bien de la période de premier emploi.

Pour ne pas se tromper, et pour faciliter les vérifications et les procédures de diagnostic, il est préférable de constituer d'abord un nouveau fichier JOB1.DTA où ne figureront que les périodes de premier emploi :

```
. keep if job1==1
(714 observations deleted)

. save job1, replace
file job1.dta saved
```

D'ailleurs, la commande « firstocc » avec l'option « saving() » crée directement ce fichier. À la place des trois dernières lignes de commandes, on aurait pu simplement taper :

```
. save job, replace
file job.dta saved

. firstocc no v601 =(v611==2 | v611==3), sav(job1)
```

Notez qu'avant d'utiliser la commande « firstocc », on a pris bien soin d'enregistrer les modifications éventuelles du fichier original : cette précaution est valable pour toute exécution d'une commande créant un nouveau fichier. Enfin, il ne reste plus qu'à créer la variable dépendante :

```
. gen byte salarie=v611==2
```

c) L'interprétation des résultats

La commande « logistic » se présente comme toutes les commandes de régression linéaire dans STATA. À la suite du nom de la variable dépendante, figurent les noms des variables indépendantes. Les variables indépendantes discrètes sont construites comme pour le modèle de régression simple du chapitre précédent, c'est-à-dire par une série de variables dichotomiques (codées 0/1), d'où on élimine la catégorie de référence :

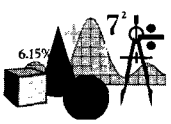
```
. tab strate, gen(gener)
         No|
 generation|      Freq.     Percent        Cum.
-----------+-----------------------------------
   Gen 1930 |         98       39.20       39.20
   Gen 1945 |         90       36.00       75.20
   Gen 1955 |         62       24.80      100.00
-----------+-----------------------------------
      Total |        250      100.00

. drop gener1
```

On veut tester le modèle suivant :

Génération
|
Salariée
1989

On utilise la commande « logistic » de la façon suivante :

```
. logistic sal gener*

Logit Estimates                                   Number of obs =     250
                                                  chi2(2)       =   47.55
                                                  Prob > chi2   =  0.0000
Log Likelihood = -146.90831                       Pseudo R2     =  0.1393

------------------------------------------------------------------------------
 salarie | Odds Ratio   Std. Err.      z    P>|z|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   3.804031   1.260629    4.032   0.000     1.986836    7.283267
  gener3 |   11.00619   4.210276    6.270   0.000     5.200211    23.29449
------------------------------------------------------------------------------
```



Le chiffre correspondant à chaque variable indépendante est appelé "**rapport de chances**" (*odds ratio*). Par exemple, une femme du groupe de générations 1945-54 ("gener2") avait près de 4 fois plus de chances d'obtenir un emploi salarié qu'une femme du groupe de générations 1930-44, qui forme la catégorie de référence. Cela correspond au modèle,

$$r_j = r_0 \exp\left(\sum_i b_i X_{ij}\right)$$

où b_i et X_i représentent respectivement les coefficients et les variables indicatrices. On remarque que le modèle est exprimé ici sous forme multiplicative.

Le rapport de chances sera noté o_j et est calculé en faisant le rapport entre les deux ratios r_j et r_0 :

$$o_j = \frac{r_j}{r_0}$$

Pour obtenir les résultats sous forme de coefficients additifs après avoir obtenu les résultats du modèle « logistique », il suffit d'exécuter la commande « logit » sans argument :

```
. logit

Logit Estimates                                   Number of obs =    250
                                                  chi2(2)       =  47.55
                                                  Prob > chi2   = 0.0000
Log Likelihood = -146.90831                       Pseudo R2     = 0.1393

------------------------------------------------------------------------------
 salarie |      Coef.   Std. Err.       z     P>|z|      [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   1.336061   .3313929      4.032   0.000       .6865432     1.98558
  gener3 |   2.398458   .3825371      6.270   0.000       1.648699    3.148217
   _cons |  -1.425009   .2555168     -5.577   0.000      -1.925813    -.9242051
------------------------------------------------------------------------------
```

Les coefficients du tableau sont simplement égaux aux logarithmes des rapports de chances :

$$\sum_i b_i X_{ij} = \log(o_j) = \log\left(\frac{r_j}{r_0}\right)$$

On remarque que, dans cette présentation-là, une constante est calculée. Cette constante correspond à la catégorie de référence. C'est d'ailleurs sous la forme additive que le modèle est réellement estimé :

$$\log(r_j) = \log(r_0) + \sum_i b_i X_{ij}$$

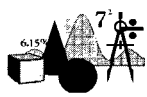
où $\log(r_0)$ est une constante que l'on peut noter c :

$$\log(r_i) = c + \sum_i b_i X_{ij}$$

On peut obtenir le risque r_0 pour la catégorie de référence par transformation exponentielle :

```
. display exp(_coef[_cons])
.24050633
```

Rappelons qu'il ne s'agit pas d'une proportion mais d'un rapport de proportions. On peut déduire la proportion elle-même en transformant la fonction puisque :



$$\exp(c) = \frac{p_0}{1-p_0} \Rightarrow p_0 = \frac{\exp(c)}{1+\exp(c)}$$

On aura donc :

```
. display exp(_coef[_cons]) / (1 + exp(_coef[_cons]))
.19387755
```

Cette probabilité peut être calculée plus simplement pour chaque catégorie de générations avec la commande « predict » pour le modèle logistique :

```
. lpredict pred
```

On peut aisément lire cette probabilité pour chaque individu de l'échantillon,

```
. list no salarie pred in 1/10

          no   salarie       pred
  1.       3         0   .1938775
  2.      11         0   .7258065
  3.      13         0   .4777778
  4.      20         1   .7258065
  5.      29         1   .7258065
  6.      33         0   .1938775
  7.      44         0   .1938775
  8.      63         1   .7258065
  9.      65         0   .4777778
 10.      73         1   .7258065
```

ou pour les différentes modalités de la variable indépendante (remarquez l'usage de la variable "strate" telle qu'elle est codée avant la polytomisation) :

```
. tab strate, summ(pred)
         No|        Summary of Pr(salarie)
  generation|        Mean   Std. Dev.       Freq.
------------+------------------------------------
   Gen 1930 |   .19387755           0          98
   Gen 1945 |   .47777778           0          90
   Gen 1955 |   .72580647           0          62
------------+------------------------------------
      Total |   .42800001   .21107308         250
```

Les écarts de probabilité entre les trois catégories peuvent se mesurer en faisant le rapport entre d'une part la proportion pour la génération 1930-44 (catégorie de référence) et d'autre part la proportion pour la génération 1945-54 :

```
. display .47777778/.19387755
2.4643275
```

ou la proportion pour la génération 1955-64 :

```
. display .72580647/.19387755
3.7436334
```

On appellera ces chiffres des "**rapports de proportions**", que l'on notera R_i et qu'on obtient facilement à partir des ratios par la formule :

$$R_i = \frac{p_i}{p_0} = \frac{r_i + r_0 r_i}{r_0 + r_0 r_i}$$

où r_0 est la constante exp*(c)*.

On calcule rapidement ces rapports de proportions pour tous les individus de l'échantillon :

```
. sort strate

. gen rapprop=pred/pred[1]

. tab strate, summ(rapprop)

          No|          Summary of rapprop
  generation|       Mean   Std. Dev.       Freq.
------------+------------------------------------
   Gen 1930 |          1           0          98
   Gen 1945 |  2.4643276           .          90
   Gen 1955 |  3.7436335           .          62
------------+------------------------------------
      Total |   2.207579   1.0886928         250
```

On voit que les rapports de proportions ne sont pas du tout égaux aux rapports de chances calculés précédemment avec la commande « logistic » (voir page 65).

Remarquons d'ailleurs que les rapports de chances ont des valeurs comprises entre 0 et plus l'infini, tandis que **les rapports de proportions ont une limite supérieure finie** (dans la mesure où la proportion pour la catégorie de référence existe), égale à $\frac{1}{p_0}$. Dans notre exemple, le rapport de proportion maximum est de :

```
. display 1/.19387755
5.1578948
```

Cela signifie que par rapport à une femme de la génération 1930-44, une femme d'une autre génération ne peut voir ses chances d'être salariée lors de son premier emploi multipliées par plus de 5,2, ce qui correspondrait à une proportion de 100 % de femmes salariées.

On voit que les rapports de proportions ont une signification plus concrète et pratique que les rapports de chances, bien que ce soient ces derniers qui sont le plus souvent présentés dans les articles scientifiques. Pour des applications pratiques du modèle logistique (estimation des risques de maladie, des chances de rémission, etc., pour telle ou telle catégorie de la population), ce sont les rapports de proportions que l'on va utiliser.

d) La présentation des résultats

Présenter les résultats en termes de rapports de chances (ou de rapport de proportions) est généralement considéré comme plus simple que la présentation en termes de coefficients, bien que ce soit le modèle additif qui soit réellement estimé[6] . Le choix des rapport ou des coefficients est une question de préférence. Les résultats doivent être interprétés de la même façon, que ce soit avec la commande « logistic » ou « logit ».

L'avantage des coefficients additifs est qu'ils sont situés sur une échelle allant de moins l'infini à plus l'infini, le point de symétrie étant zéro. Mais, lors de l'interprétation, il est difficile d'utiliser tels quels ces coefficients. L'utilisateur ne se rendra pas bien compte, en termes de probabilité, de la différence entre, par exemple, un coefficient de 1,1 et un autre de 1,5 dans un modèle additif.

Par conséquent, la transformation exponentielle, c'est-à-dire le passage au rapport de chances, est nécessaire et d'ailleurs très courant dans la littérature. Or, comme on l'a dit plus haut, le rapport de chances est placé sur une échelle allant de 0 à plus l'infini, avec pour point de référence la valeur 1. Le lecteur comprendra facilement, en termes de probabilité, la signification d'un rapport de chances passant de 3 à 4,5 par exemple, mais devra effectuer une petite gymnastique intellectuelle pour évaluer l'écart entre un rapport de chances de 0,5 et un autre de 0,2 qui correspondent respectivement à une réduction de 50 % et à une réduction de 80 % du risque de posséder la caractéristique étudiée.

En fait, il est plus facile pour le lecteur, lorsque le rapport de chances est inférieur à 1 de raisonner en terme de division des chances. Le point de référence reste toujours la valeur 1, mais les valeurs citées vont de 1 à plus l'infini dans les deux cas : seul change le terme "multiplié" ou "divisé".

Le tableau 2 résume les différentes formes de présentation de résultats. La première ligne correspond au coefficient b_i effectivement calculé dans le modèle additif par la commande « logit ». En comparant cette ligne et la dernière, on voit tout de suite l'avantage de la symétrie. Nous suggérons au lecteur d'utiliser dans le corps d'un texte ou lors d'une présentation orale, la dernière ligne de présentation. Dans les tableaux statistiques, il peut choisir la présentation des coefficients (obtenus par la commande « logit »), mais ce sont les rapports de chances (obtenus par la commande « logistic ») qui sont le plus souvent présentés dans la littérature statistique. Si le lecteur veut faire mieux encore, il pourra recalculer les rapports de proportions en s'aidant de la formule de la section précédente.

---

[6] La transformation par les logarithmes est nécessaire pour les estimations d'un modèle linéaire.

Tableau 2. Présentation des résultats d'un modèle logistique

	Limite inférieure	Exemple de coefficient négatif	Point de référence	Exemple de coefficient positif	Limite supérieure
Modèle sous forme additive « logit » : b_i	$-\infty$	-1,5	0	+1,5	$+\infty$
Modèle sous forme multiplicative « logistic » : =exp(b_i)	0	0,22	1	4,5	$+\infty$
Interprétation immédiate	Réduction de :			Multiplication par :	
	100 %	78 %	0 % / 1	4,5	$+\infty$
Interprétation symétrique	Division par :			Multiplication par :	
	$+\infty$	4,5	1	4,5	$+\infty$

Le modèle permet en principe d'évaluer certains coefficients à plus ou moins l'infini. Comment sont représentés de tels coefficients dans la régression ? Nous allons l'illustrer avec l'effet de l'instruction :

```
. tab instr, gen(ins_)

      niveau|
d'instructio|
           n|      Freq.     Percent        Cum.
------------+-----------------------------------
      nsplf |         149       59.60       59.60
        slf |          14        5.60       65.20
    CP / CE |          11        4.40       69.60
         CM |          24        9.60       79.20
    CES inc |          10        4.00       83.20
      3eme  |          19        7.60       90.80
    Lyc inc |           8        3.20       94.00
      Techn |           5        2.00       96.00
    Termin. |           5        2.00       98.00
        Sup |           5        2.00      100.00
------------+-----------------------------------
      Total |         250      100.00

. drop ins_1

. logistic salarie ins_*

Note : ins_9~=0 predicts success perfectly
       ins_9 dropped and 5 obs not used

Note : ins_10~=0 predicts success perfectly
       ins_10 dropped and 5 obs not used
```

.../...

```
Logit Estimates                                   Number of obs =     240
                                                  chi2(7)       =   41.90
                                                  Prob > chi2   =  0.0000
Log Likelihood = -140.97137                       Pseudo R2     =  0.1294

------------------------------------------------------------------------------
 salarie | Odds Ratio   Std. Err.      z    P>|z|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
   ins_2 |   2.820513   1.596635    1.832   0.067      .9299941    8.554133
   ins_3 |   3.384615   2.144358    1.924   0.054      .9777366    11.71647
   ins_4 |   2.820513   1.265774    2.311   0.021       1.17039    6.797131
   ins_5 |   6.581197   4.704163    2.636   0.008      1.621347    26.71369
   ins_6 |   15.04273   9.870322    4.132   0.000      4.157279    54.43077
   ins_7 |   8.461538     7.0865    2.550   0.011      1.638974    43.68443
   ins_8 |   11.28205   12.78773    2.138   0.033      1.223452    104.0373
------------------------------------------------------------------------------
```

Deux catégories ("ins_9" et "ins_10") ont été exclues du modèle de même que les femmes appartenant à ces catégories (niveau de la classe de terminale ou études supérieures). Un message nous informe que toutes ces femmes ont obtenu un emploi salarié, et que par conséquent le modèle prédit pour elles des chances totales de "succès". Dans le cas de chances nulles de "succès", les individus concernés et les modalités correspondantes seraient aussi éliminés et le message serait "var_name~=0 predicts failure perfectly".

On assiste là à un véritable paradoxe : les individus pour lesquels le modèle s'ajuste parfaitement sont éliminés des calculs. Pour obtenir des coefficients tendant vers plus ou moins l'infini qui correspondent à des chances totales ou, au contraire, nulles de succès, on utilisera la commande « mlogit », avec une option marquant une contrainte de convergence « ltol(0,000001) ». La commande « mlogit » est normalement utilisée pour estimer le modèle multinomial logit qui sera explicité dans la section suivante.

```
. mlogit salarie ins_*, ltol(0.000001) rrr

Iteration 0 :  Log Likelihood =-170.68576
(etc)
Iteration 10 :  Log Likelihood = -140.9715

Multinomial regression                            Number of obs =     250
(Log Likelihood tolerance 1.00e-06)               chi2(9)       =   59.43
                                                  Prob > chi2   =  0.0000
Log Likelihood = -140.97142                       Pseudo R2     =  0.1741

------------------------------------------------------------------------------
 salarie |        RRR   Std. Err.      z    P>|z|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
1        |
   ins_2 |   2.820513   1.596635    1.832   0.067       .929994    8.554133
   ins_3 |   3.384615   2.144358    1.924   0.054      .9777366    11.71647
   ins_4 |   2.820513   1.265774    2.311   0.021       1.17039    6.797132
   ins_5 |   6.581197   4.704163    2.636   0.008      1.621347    26.71369
   ins_6 |   15.04274   9.870663    4.131   0.000      4.157094    54.43318
   ins_7 |   8.461538     7.0865    2.550   0.011      1.638973    43.68444
   ins_8 |   11.28205   12.78776    2.138   0.033      1.223447    104.0378
   ins_9 |   583894.5   1.19e+08    0.065   0.948     3.67e-168   9.30e+178
  ins_10 |   583894.5   1.19e+08    0.065   0.948     3.67e-168   9.30e+178
------------------------------------------------------------------------------
(Outcome salarie==0 is the comparison group)
```

Ce nouveau tableau de résultats montre que le maximum de vraisemblance estimé (*Log Likelihood*) est resté identique, de même que les coefficients pour les niveaux d'instruction "ins_2" à "ins_8". Par contre, les niveaux d'instruction "ins_9" et "ins_10" sont conservés dans la régression, mais leurs coefficients sont anormalement élevés, car ils correspondent à des valeurs positives infinies, de même que les écarts-types (*standard errors*). On remarque que les degrés de significativité pour ces coefficients sont proches de 1. Cette valeur, qui indiquerait dans le cas d'un coefficient normal, une non-significativité absolue, doit être interprétée à l'opposé. Le degré de ***significativité*** n'est pas calculable pour un coefficient tendant vers l'infini, mais le coefficient en lui-même est très ***signifiant*** puisqu'il indique une prédiction totale. On voit encore une fois l'importance de l'interprétation qualitative des données statistiques.

Le calcul du modèle « mlogit » a été rendu possible par l'option « ltolerance(0,000001) » qui fixe un seuil de tolérance entre deux itérations successives pour interrompre le calcul. Si le rapport entre le maximum de vraisemblance à l'itération *n* et le maximum de vraisemblance à l'itération *n*-1 est inférieur à 0,000001, les itérations sont interrompues et les résultats affichés.

L'avantage de la commande « mlogit » est que l'on peut estimer le gain de vraisemblance véritable, compte tenu des coefficients tendant vers l'infini : dans l'exemple, ce gain mesuré par le *Pseudo R2* est de 17,41 % alors que la commande « logistic » l'estimait seulement à 12,94 %.

Les résultats équivalents à la commande « logit » peuvent être simplement obtenus sans l'option « rrr » (*relative risk ratios*) :

```
. mlogit salarie ins_*, ltol(0.000001)

Iteration 0 :  Log Likelihood =-170.68576
(etc)
Iteration 10 :  Log Likelihood = -140.9715

Multinomial regression                          Number of obs =     250
(Log Likelihood tolerance 1.00e-06)             chi2(9)       =   59.43
                                                Prob > chi2   =  0.0000
Log Likelihood = -140.97142                     Pseudo R2     =  0.1741

------------------------------------------------------------------------------
     salarie |      Coef.   Std. Err.       z    P>|z|     [95 % Conf. Interval]
-------------+----------------------------------------------------------------
1            |
       ins_2 |   1.036919   .5660797     1.832   0.067    -.0725771    2.146415
       ins_3 |    1.21924   .6335603     1.924   0.054     -.022515    2.460996
       ins_4 |   1.036919   .4487746     2.311   0.021     .1573368    1.916501
       ins_5 |   1.884217   .7147884     2.636   0.008     .4832571    3.285176
       ins_6 |   2.710895   .6561747     4.131   0.000     1.424816    3.996974
       ins_7 |   2.135531   .8374954     2.550   0.011     .4940701    3.776992
       ins_8 |   2.423213    1.13346     2.138   0.033      .201672    4.644754
       ins_9 |   13.27748   203.4794     0.065   0.948    -385.5348    412.0898
      ins_10 |   13.27748   203.4794     0.065   0.948    -385.5348    412.0898
       _cons |  -1.036919   .1863651    -5.564   0.000    -1.402188   -.6716499
------------------------------------------------------------------------------
(Outcome salarie==0 is the comparison group)
```

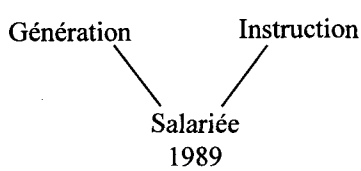
Hormis la prise en compte des coefficients tendant vers l'infini, les résultats des commandes « mlogit » et « logistic » sont absolument équivalents : les coefficients pour les autres modalités restent inchangés. Il est donc préférable d'utiliser la commande « mlogit » qui estime le même modèle que la commande « logistic » mais dont les résultats sont plus informatifs et plus proches de l'interprétation que l'on veut donner au modèle. Cependant, de nombreux diagnostics de la régression sont disponibles après la commande « logistic » ce qui nous obligera quand même à l'utiliser en temps voulu.

3. L'ajustement et son diagnostic

Quelle que soit la présentation des tableaux de résultats, le chercheur doit formuler son interprétation le plus intelligiblement possible. Pour cela, il doit évaluer la pertinence du modèle, en s'aidant d'un diagnostic des résultats.

Comme on l'a dit à propos de la régression simple, un modèle n'a d'intérêt que s'il comporte plusieurs variables. Nous illustrerons donc le diagnostic du modèle logistique en ajoutant l'effet de la génération à celui de l'instruction.

La régression peut être formalisée par la relation causale :



Elle peut être calculée par :

```
. mlogit salarie gener2 gener3 ins_2-ins_10, rrr ltol(0.000001)

Iteration 0 :  Log Likelihood =-170.68576
(etc)
Iteration 10 :  Log Likelihood =-128.93896

Multinomial regression                          Number of obs =     250
(Log Likelihood tolerance 1.00e-06)             chi2(11)      =   83.49
                                                Prob > chi2   =  0.0000
Log Likelihood = -128.93887                     Pseudo R2     =  0.2446
```

.../...

salarie	RRR	Std. Err.	z	P>|z|	[95 % Conf.	Interval]
1						
gener2	2.836626	1.035768	2.855	0.004	1.386736	5.802437
gener3	7.378917	3.149012	4.683	0.000	3.196947	17.03138
ins_2	2.914907	1.775821	1.756	0.079	.8831924	9.620419
ins_3	1.969819	1.342558	0.995	0.320	.5179378	7.491605
ins_4	1.509047	.735096	0.845	0.398	.5808452	3.920533
ins_5	6.037051	4.53247	2.395	0.017	1.386001	26.29579
ins_6	8.342833	5.686612	3.112	0.002	2.193414	31.73266
ins_7	4.910448	4.340554	1.800	0.072	.868376	27.76735
ins_8	8.217896	9.770157	1.772	0.076	.7994243	84.47805
ins_9	438555.1	8.50e+07	0.067	0.947	4.40e-160	4.37e+170
ins_10	260839.3	5.20e+07	0.063	0.950	5.71e-165	1.19e+175

(Outcome salarie==0 is the comparison group)

Pour obtenir les rapports de proportions après la commande « mlogit », on effectue les calculs suivants :

```
. sort strate instr
. drop pred
. predict pred, outcome(1)
. drop rapprop
. gen rapprop=pred/pred[1]
. tab instr strate, summ(rapprop) nost
```

Means and Frequencies of rapprop

niveau d'instructi on	No generation Gen 1930	Gen 1945	Gen 1955	Total
nsplf	1	2.2407467	3.8359601	1.8077299
	79	49	21	149
slf	2.2821503	4.0286565	5.4185033	3.4532277
	7	4	3	14
CP / CE	1.7272716	3.357444	4.910882	3.7671572
	2	4	5	11
CM	1.4054564	2.9020927	4.512742	3.4532274
	3	10	11	24
CES inc	3.4909868	5.135272	6.0979328	4.8345186
	3	5	2	10
3eme	4.0436687	5.5270176	6.3018284	5.8159622
	1	9	9	19
Lyc inc	3.1352584	4.8499126	5.9384336	5.1798414
	1	3	4	8
Techn	4.0183458	5.5102854	6.2934518	5.525164
	1	2	2	5
Termin.	6.9063621	6.9064221	6.9064422	6.9064181
	1	2	2	5
Sup	.	6.9063997	6.9064341	6.9064203
	0	2	3	5
Total	1.3390057	3.2997488	5.0127482	2.9559614
	98	90	62	250

On peut les comparer aux résultats des calculs obtenus après la commande « logistic » :

```
. logistic salarie gener2 gener3 ins_2-ins_10

Note : ins_9~=0 predicts success perfectly
       ins_9 dropped and 5 obs not used

Note : ins_10~=0 predicts success perfectly
       ins_10 dropped and 5 obs not used

Logit Estimates                                   Number of obs =    240
                                                  chi2(9)       =  65.96
                                                  Prob > chi2   = 0.0000
Log Likelihood = -128.93881                       Pseudo R2     = 0.2037

------------------------------------------------------------------------------
 salarie | Odds Ratio   Std. Err.      z    P>|z|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   2.836626   1.035768    2.855   0.004     1.386736    5.802439
  gener3 |   7.378917   3.149014    4.683   0.000     3.196946    17.03138
   ins_2 |   2.914907   1.775821    1.756   0.079     .8831924    9.620419
   ins_3 |   1.969819   1.342558    0.995   0.320     .5179378    7.491606
   ins_4 |   1.509047    .735096    0.845   0.398     .5808451    3.920533
   ins_5 |   6.037051    4.53247    2.395   0.017     1.386001    26.29579
   ins_6 |   8.342833   5.686612    3.112   0.002     2.193414    31.73266
   ins_7 |   4.910448   4.340554    1.800   0.072      .868376    27.76735
   ins_8 |   8.217896   9.770157    1.772   0.076     .7994243    84.47805
------------------------------------------------------------------------------
```

Pour obtenir les rapports de proportions qui tiennent compte des prédictions parfaites, on calcule d'abord :

```
. sort strate instr

. drop pred

. lpredict pred
```

Ensuite, la prédiction est remplacée par la valeur 0 dans le cas de prédiction d'échec total (ce qu'on n'a pas rencontré dans notre exemple) et par la valeur 1 dans le cas de prédiction de succès total (ici pour les niveaux d'instruction "terminale" et "supérieur") :

```
. replace pred=1 if ins_9==1 | ins_10==1

. drop rapprop

. gen rapprop=pred/pred[1]
```

a) Tests statistiques : « lfit », « lstat », « lroc »

Comme pour les simples modèles de régression, il est important de voir l'adéquation du modèle aux données. Outre le gain de vraisemblance (*Pseudo R2*), on peut comparer pour chaque catégorie de la population, la proportion de réponses positives observées (ici, le nombre de femmes salariées) et la proportion de réponses positives prédites par le modèle.

Les proportions de salariées figurent dans le tableau croisant niveau d'instruction et groupe de générations :

```
. tab instr strate, summ(salarie) nost

           Means and Frequencies of salarie

    niveau| No generation
d'instructi|
        on|  Gen 1930   Gen 1945   Gen 1955      Total
----------+---------------------------------+----------
    nsplf | .15189873  .30612245  .57142857 | .26174497
          |        79          49         21 |       149
----------+---------------------------------+----------
      slf |  .14285714        .75          1 |        .5
          |         7          4          3 |        14
----------+---------------------------------+----------
  CP / CE |          0        .5         .8 | .54545455
          |         2          4          5 |        11
----------+---------------------------------+----------
       CM |  .33333333        .4  .63636364 |        .5
          |         3         10         11 |        24
----------+---------------------------------+----------
  CES inc |  .66666667        .6          1 |        .7
          |         3          5          2 |        10
----------+---------------------------------+----------
     3eme |          1  .88888889  .77777778 | .84210526
          |         1          9          9 |        19
----------+---------------------------------+----------
  Lyc inc |          0  .66666667          1 |       .75
          |         1          3          4 |         8
----------+---------------------------------+----------
    Techn |          1          1         .5 |        .8
          |         1          2          2 |         5
----------+---------------------------------+----------
  Termin. |          1          1          1 |         1
          |         1          2          2 |         5
----------+---------------------------------+----------
      Sup |          .          1          1 |         1
          |         0          2          3 |         5
----------+---------------------------------+----------
    Total | .19387755  .47777778  .72580645 |      .428
          |        98         90         62 |       250
```

Contrairement au modèle à une seule variable, les proportions observées et prédites par le modèle ne sont pas les mêmes :

```
. tab instr strate, summ(pred)

                Means and Frequencies of Pr(salarie)
    niveau| No generation
d'instructi|
        on|  Gen 1930   Gen 1945   Gen 1955      Total
----------+------------------------------------+----------
     nsplf | .14479208  .32444239  .55541664  | .26174497
           |        79         49         21  |       149
----------+------------------------------------+----------
       slf | .33043727  .58331758  .78455633  | .50000001
           |         7          4          3  |        14
----------+------------------------------------+----------
   CP / CE | .25009525  .48613131  .71105683  | .54545453
           |         2          4          5  |        11
----------+------------------------------------+----------
        CM | .20349896  .42020005   .6534093  | .49999999
           |         3         10         11  |        24
----------+------------------------------------+----------
   CES inc | .50546724  .74354672  .88293236  | .70000001
           |         3          5          2  |        10
----------+------------------------------------+----------
      3eme | .58549124  .80026835  .91245484  | .84210526
           |         1          9          9  |        19
----------+------------------------------------+----------
   Lyc inc |  .4539606  .70222896  .85983813  |       .75
           |         1          3          4  |         8
----------+------------------------------------+----------
     Techn | .58182466  .79784566  .91124201  |        .8
           |         1          2          2  |         5
----------+------------------------------------+----------
   Termin. |         1          1          1  |         1
           |         1          2          2  |         5
----------+------------------------------------+----------
       Sup |         .          1          1  |         1
           |         0          2          3  |         5
----------+------------------------------------+----------
     Total | .19387756  .47777778  .72580645  |      .428
           |        98         90         62  |       250
```

On remarque au passage que les proportions prédites pour les niveaux "terminale" et "supérieur" sont de 100 %.

Pour le diagnostic, la commande « logistic » sera préférée à la commande « logit » ou « mlogit », dans la mesure où l'on garde à l'esprit que les individus pour lesquels le modèle prédit parfaitement le résultat sont éliminés. Les commandes de diagnostic « lfit », « lstat » et « lroc » fournies par STATA ne prévoient pas ces cas de prédiction parfaite, ce qui est paradoxal puisque cela devrait être pris en compte dans l'évaluation du modèle. C'est pourquoi nous avons modifié ces commandes de diagnostic, pour tenir compte des prédictions parfaites dans les calculs, sous la forme de nouveaux programmes figurant sur la disquette accompagnant ce manuel : « lfit2 », « lstat2 » et « lroc2 ».

Le test du $\chi 2$ de Pearson (*Pearson's Chi2*) évalue la différence entre les deux séries de proportions observées et ajustées :

```
. lfit
Logistic estimates for salarie, goodness-of-fit test

          no. of observations =         240
    no. of covariate patterns =          24
              Pearson chi2(12) =       14.95
                    P>chi2 =          0.2442
```

Avec le programme « lfit2 », on obtient les calculs corrigés :

```
. lfit2
Logistic estimates for salarie, goodness-of-fit test

          no. of observations =         250
    no. of covariate patterns =          29
              Pearson chi2(17) =       14.95
                    P>chi2 =          0.5992
```

Les catégories de la population (*covariate patterns*) résultent du croisement des variables introduites dans le modèle. La statistique de test ($P>\chi 2$) indique, contrairement au test sur le gain de vraisemblance, le risque que l'on a de se tromper en disant que le modèle n'explique pas la proportion de salariées. On voit que l'ajustement est nettement meilleur lorsqu'on tient compte des prédictions parfaites : le risque est de 24,42 % de se tromper en rejetant le modèle dans le premier cas et de 59,92 % dans le second, ce qui indique que le modèle est plutôt bon.

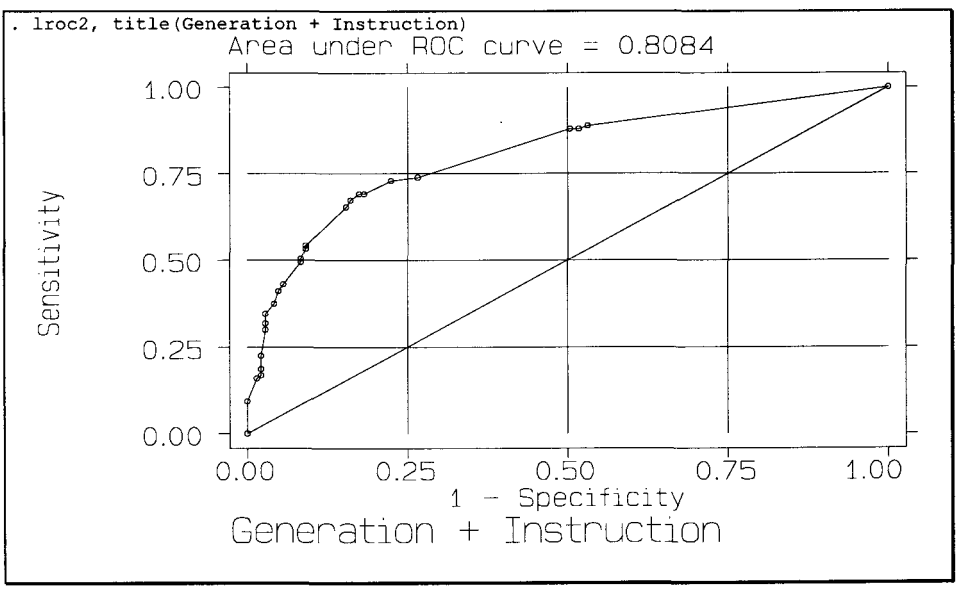
Une autre manière d'évaluer le modèle est de classer les individus selon leurs probabilités prédites et la réponse observée. Tout d'abord, on fixe un seuil (le plus souvent 0,5) de probabilité au-delà duquel on considère que la réponse est positive. Ensuite, on compare la sensibilité (*sensitivity*), c'est-à-dire la proportion de réponses positives observées (*Observed Positive*) qui sont correctement classées comme positives par le modèle (*Classified Positive*), et la spécificité (*specificity*), c'est-à-dire la proportion de réponses négatives observées (*Observed Negative*) qui sont correctement classées comme négatives (*Classified Negative*). De la même façon qu'avec « lfit2 », on utilisera le programme « lstat2 » :

```
. lstat2
Logistic estimates for salarie  (Positive = p>=.5)
(Statistics for different sample than estimation)

                      ----- Classified -----
           Observed |  Negative      Positive  |       Total
           ---------+--------------------------+-----------
           Negative |       120            23  |         143
           Positive |        35            72  |         107
           ---------+--------------------------+-----------
           Total    |       155            95  |         250
-------------------------------------------------------------------------------
Model sensitivity          67.29 %   |   Model specificity             83.92 %
False positive rate        24.21 %   |   False negative rate           22.58 %
Positive predictive value  75.79 %   |   Negative predictive value     77.42 %
-------------------------------------------------------------------------------
```

La sensibilité est donc égale à 72/107 soit 67,29 %, et la spécificité à 120/143 soit 83,29 %. Si le modèle prédisait parfaitement les données, les individus seraient répartis sur la diagonale du tableau croisé des réponses observées et classées. Le message "*Statistics for different sample than estimation*" apparaît lorsqu'on utilise le programme « lstat2 », car le diagnostic est étendu aux observations qui ont été éliminées par la commande « logistic ».

La commande « lstat » (ou « lstat2 ») permet de classer les observations selon les résultats du modèle. Ce classement est surtout utile en médecine, en biostatistiques ou en actuariat, où les besoins d'applications pratiques de la modélisation sont importants : établir un diagnostic médical sur un ensemble de patients, par exemple. En sciences sociales, la classification est rarement le but de la modélisation, mais elle permet au moins un diagnostic du modèle.

Une bonne manière de visualiser l'ajustement des données au modèle est de tracer la courbe ROC[7], avec la commande « lroc2 » pour tenir compte des prédictions parfaites :



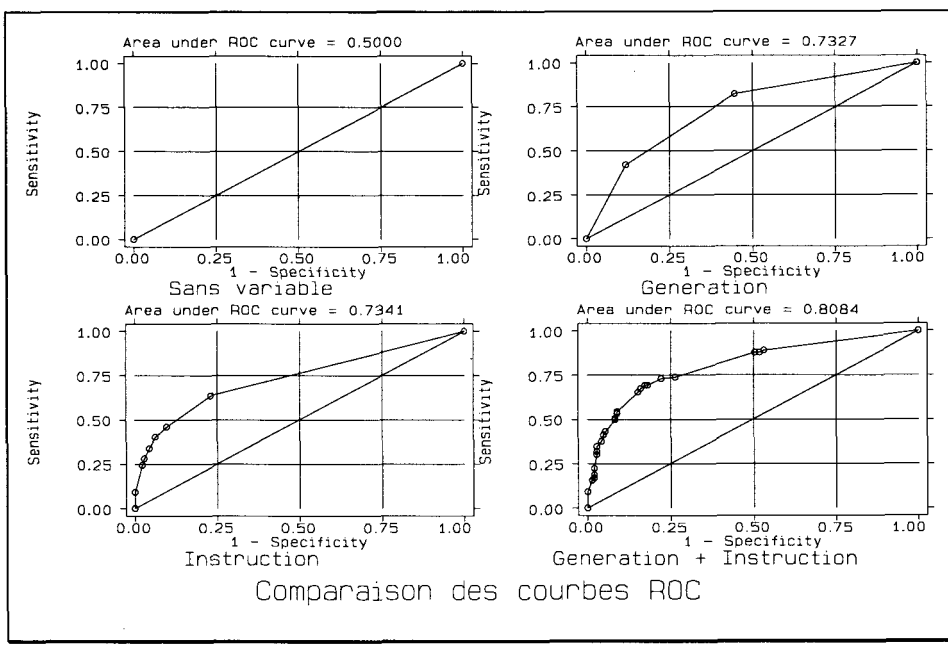
Plus le tracé est courbe vers le coin en haut à gauche du graphique, meilleure est la prédiction. Un modèle sans pouvoir de prédiction aurait une courbe tracée à 45° qui se confondrait alors avec la diagonale sur le graphique. La surface au-dessous de la courbe serait de 50 %. Un modèle avec un pouvoir total de prédiction

[7] *Receiver Operating Characteristic* : le terme, qui provient de la théorie de détection des signaux, est utilisé par convention.

aurait pour courbe ROC une droite verticale le long de l'axe de sensibilité[8]. La surface au-dessous de la courbe serait alors de 100 %.

Dans notre exemple, la surface au-dessous de la courbe est de 80,84 %, comme indiqué sur le graphique. La limite inférieure étant de 50 % et la limite supérieure de 100 %, il est préférable d'évaluer le pouvoir de prédiction du modèle en tenant compte de ces limites : par rapport à un modèle sans variable explicative (pouvoir de prédiction nul), le chiffre de 80,84 % correspond à un pouvoir de prédiction de (0,8084-0,50)/0,50 soit 61,68 % supérieur.

On peut ainsi utiliser les surfaces au-dessous de la courbe ROC pour comparer les résultats de différents modèles entre eux, comme dans le prochain graphique. La première figure correspond à un pouvoir de prédiction nul, la deuxième (variable indépendante : génération) à un pouvoir de prédiction de (0,7327-0,5)/0,5 soit 46,54 %, la troisième (variable indépendante : instruction) à un pouvoir de prédiction de (0,7341-0,5)/0,5 soit 46,82 %, et la quatrième (variables indépendantes : génération et instruction ensemble) à un pouvoir de prédiction de (0,8084-0,5)/0,5 soit 61,68 %.



[8] Rappelons cependant qu'un modèle qui prédit parfaitement les données ne peut être estimé sous forme logistique.

b) Représentation graphique des résidus : « lpredict »

Le modèle logistique présente des possibilités de diagnostic graphique intéressantes après quelques calculs en s'aidant de la commande « lpredict ». Pour calculer les résidus (standardisés), l'effet de levier (*leverage* : voir le chapitre sur la régression simple), et l'effet de chaque catégorie sur les coefficients de la régression (*dbeta de Pregibon*), on procédera comme suit :

```
. lpredict levier, hat

. lpredict resid, rstandard

. lpredict dbeta, dbeta
```

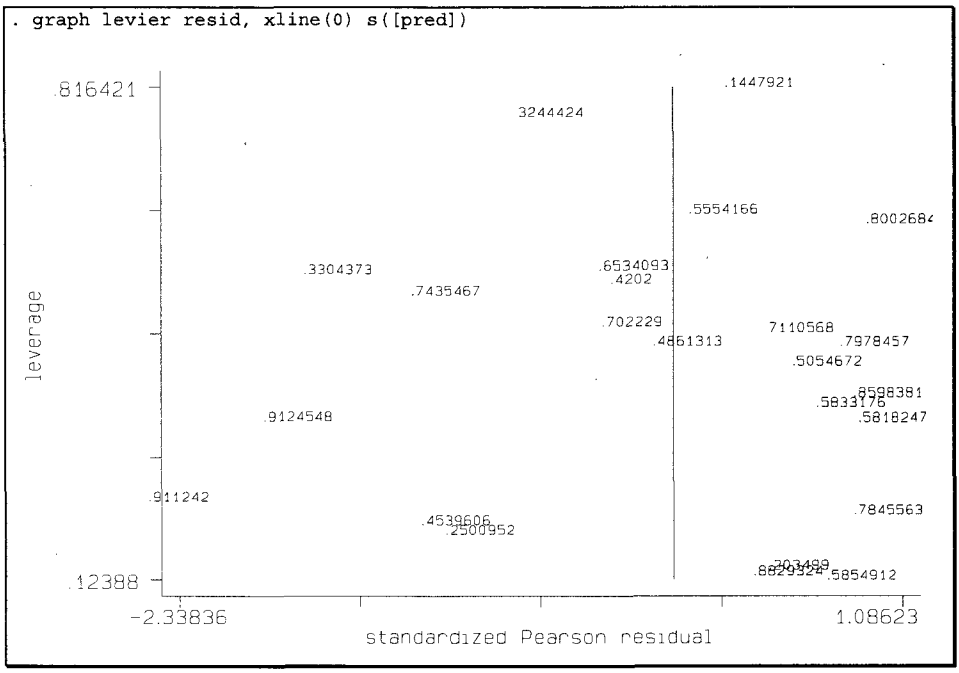
Ces calculs ne sont faits que sur l'échantillon qui a servi aux estimations, c'est-à-dire sans tenir compte des prédictions parfaites, pour lesquels les individus concernés ont été éliminés. Rappelons que pour ces individus les résidus sont nuls. Les autres statistiques (effet levier, dbeta...) étant basées essentiellement sur ces résidus, elles ne peuvent en fait être calculées pour les catégories formées par les prédictions parfaites. Les diagnostics valent donc seulement pour les autres catégories.

Ces statistiques sont données pour les catégories résultant du croisement de toutes les variables prises en compte dans le modèle. Les calculs ne se rapportent donc pas à chaque individu en particulier mais aux catégories auxquelles ils appartiennent. Dans notre exemple, il s'agit des catégories formées par le croisement des variables "groupe de générations" et "instruction". On peut obtenir simplement le nombre de catégories en exécutant la commande :

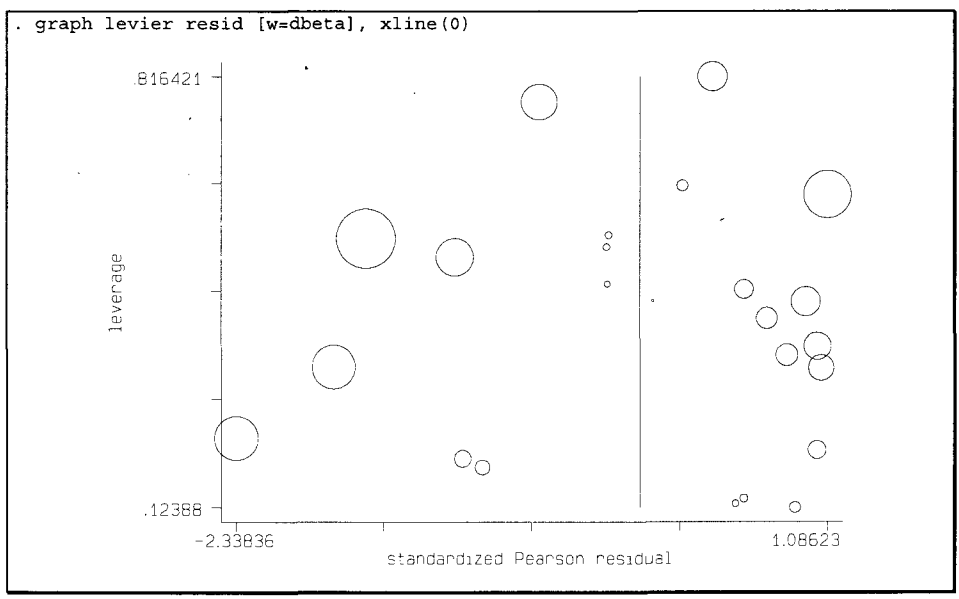
```
. lpredict categ, number

. summarize categ

Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
   categ |     240    10.30833    7.751146          1         24
```

Les statistiques seront donc calculées pour 24 catégories dans l'échantillon des 240 femmes pour lesquelles la prédiction n'est pas parfaite.

Pour visualiser directement les résidus et l'effet levier, on peut faire le graphique suivant :



Chacune des 24 catégories est représentée par sa proportion correspondante. Pour représenter les catégories par des cercles proportionnels aux indices *dbeta* de *Pregibon*, on donnera l'instruction suivante :



Les indices *dbeta* mesurent le changement dans le vecteur des coefficients que causerait l'absence de chacune des catégories. On voit maintenant que la catégorie qui présente à la fois le plus de résidus, un effet levier important et un indice *dbeta* élevé est la catégorie où la proportion prédite de salariée est de 33 %. Il s'agit en fait des femmes non scolarisées parlant français de la génération 1930-44.

4. Le modèle multinomial logit : commande « mlogit »

Le modèle logistique est adapté aux situations où la caractéristique étudiée est unique et homogène : on mesure ainsi l'effet de certaines variables explicatives sur la présence ou l'absence de telle ou telle caractéristique dans l'échantillon.

Dans certaines situations, la caractéristique étudiée n'est pas homogène, elle est constituée de plusieurs modalités. Plusieurs cas d'hétérogénéité peuvent se présenter :

- la caractéristique est évaluée par degré, c'est-à-dire selon une échelle ordinale, comme par exemple un niveau de satisfaction, allant par pallier de "mauvais" à "excellent". La caractéristique est à la fois multinomiale et ordonnée, sans toutefois être continue comme dans la régression simple. Dans ce cas, on utilisera un **modèle logit ordonné** (ordered logit model), que STATA estime avec la commande « ologit ». Le modèle logit ordonné ne sera pas commenté ici, car il n'existe pas actuellement de modèle équivalent pour l'analyse des biographies. Ajoutons que les variables dépendantes ordonnées sont rares dans les données rétrospectives en sciences sociales ;
- *la caractéristique est multinomiale mais n'est pas ordonnée*. Les modalités de la variable dépendante ne peuvent être classées selon un ordre de valeur. On sait simplement qu'elles sont exclusives l'une de l'autre, c'est-à-dire que les individus ne peuvent être classés à la fois selon l'une et l'autre de ces modalités. On utilisera dans ce cas un **modèle multinomial logit**, estimé avec la commande « mlogit » ;
- le **modèle logit conditionnel** est aussi un cas intéressant. Ce modèle est adapté aux situations où ce n'est plus la variable dépendante qui est multinomiale, mais l'échantillon qui est réparti selon des *strates exclusives l'une de l'autre*. Le modèle logit conditionnel est souvent utilisé en médecine ou en biologie pour comparer des échantillons assortis par paire (*Matched case-control data*). Mais, comme pour le modèle logit ordonné, il n'y a pas actuellement d'équivalent du modèle logit conditionnel pour l'analyse des biographies. De plus, les échantillons sont peu souvent dans une telle logique en sciences sociales.

a) La présentation des résultats

Nous avons déjà vu que le modèle multinomial logit produit les mêmes résultats que le modèle logistique (hormis pour les prédictions parfaites) dans le cas où la variable dépendante indique la présence ou l'absence d'une seule caractéristique (codée 0 ou 1). Les coefficients mesurent les chances relatives de rencontrer cette caractéristique selon telle ou telle catégorie définie par les variables indépendantes.

Dans le modèle multinomial logit, ce n'est pas la présence d'une seule caractéristique dont on mesure les chances, mais l'alternative entre plusieurs caractéristiques. Considérons non plus simplement le fait pour une femme d'être salariée ou non, mais le fait d'être salariée du secteur moderne (avec fiche de paie), salariée du secteur informel (sans fiche de paie), commerçante indépendante ou indépendante dans une autre branche :

```
. gen emploi=1 if salarie==1 & v627==1
(194 missing values generated)

. replace emploi=2 if salarie==1 & v627==2
(51 real changes made)

. replace emploi=3 if salarie==0 & v610==4
(107 real changes made)

. replace emploi=4 if salarie==0 & v610 !=4
(36 real changes made)

. lab def emploi 1 "SalForm" 2 "SalInfl" 3 "IndComm" 4 "IndAutre"

. lab val emploi emploi

. tab emploi
     emploi|      Freq.     Percent        Cum.
-----------+-----------------------------------
    SalForm |         56       22.40       22.40
    SalInfl |         51       20.40       42.80
    IndComm |        107       42.80       85.60
   IndAutre |         36       14.40      100.00
-----------+-----------------------------------
      Total |        250      100.00
```

On estime le modèle multinomial logit avec les variables indépendantes reflétant le groupe de générations et le niveau d'instruction :

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001)

Iteration 0 :  Log Likelihood =-325.42306
(etc)
Iteration 12 :  Log Likelihood =-251.26818
Multinomial regression                          Number of obs =     250
(Log Likelihood tolerance 1.00e-07)             chi2(33)      =  148.31
                                                Prob > chi2   =  0.0000
Log Likelihood = -251.26817                     Pseudo R2     =  0.2279
------------------------------------------------------------------------------
      emploi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
SalForm      |
     gener2 |   .3972223   .5556149     0.715   0.475    -.691763    1.486208
     gener3 |   1.791126   .6429739     2.786   0.005     .5309206   3.051332
      ins_2 |   2.297029   .9030164     2.544   0.011     .5271497   4.066909
      ins_3 |   2.272653   1.012711     2.244   0.025     .2877752    4.25753
      ins_4 |   1.086041   .7475807     1.453   0.146    -.3791904   2.551272
      ins_5 |   4.025768   1.150131     3.500   0.000     1.771553   6.279983
      ins_6 |   4.417761   1.122512     3.936   0.000     2.217677   6.617844
      ins_7 |    2.48996   .9840618     2.530   0.011     .5612344   4.418686
      ins_8 |   3.359705   1.214394     2.767   0.006     .9795366   5.739873
      ins_9 |   17.81895   1184.642     0.015   0.988    -2304.036   2339.674
     ins_10 |   17.45469   1189.671     0.015   0.988    -2314.258   2349.167
      _cons |  -2.679517   .4541988    -5.899   0.000    -3.569731  -1.789304
------------+----------------------------------------------------------------
SalInfl      |
     gener2 |   1.215703   .4621725     2.630   0.009     .3098612   2.121544
     gener3 |   2.602818   .5591844     4.655   0.000     1.506837   3.698799
      ins_2 |   1.400715   .8348915     1.678   0.093    -.2356427   3.037072
      ins_3 |   .8811903   .9811347     0.898   0.369    -1.041798   2.804179
      ins_4 |   .3595641   .5890075     0.610   0.542    -.7948693   1.513997
      ins_5 |   .9198975   1.452188     0.633   0.526    -1.926339   3.766134
      ins_6 |   1.510702   1.197013     1.262   0.207    -.8353996   3.856803
      ins_7 |   .3893076   1.080967     0.360   0.719    -1.729348   2.507963
      ins_8 |  -13.65369   1247.216    -0.011   0.991    -2458.152   2430.845
      ins_9 |  -.0659891   2107.338    -0.000   1.000    -4130.373   4130.241
     ins_10 |  -.6073151   2153.879    -0.000   1.000    -4222.132   4220.917
      _cons |  -2.027458   .3695397    -5.486   0.000    -2.751742  -1.303173
------------+----------------------------------------------------------------
IndAutre     |
     gener2 |   -.488996   .4865033    -1.005   0.315    -1.442525    .4645329
     gener3 |   .7756905   .6002216     1.292   0.196    -.4007222   1.952103
      ins_2 |    1.71382   .8050016     2.129   0.033     .1360461   3.291594
      ins_3 |   1.757784   .9696229     1.813   0.070    -.1426418    3.65821
      ins_4 |   .6388595   .6998169     0.913   0.361    -.7327564   2.010475
      ins_5 |   2.234429   1.257701     1.777   0.076    -.2306195   4.699477
      ins_6 |   2.137776   1.277392     1.674   0.094    -.3658672   4.641419
      ins_7 |  -12.58135   738.8459    -0.017   0.986    -1460.693    1435.53
      ins_8 |  -13.27682   1542.523    -0.009   0.993    -3036.567   3010.013
      ins_9 |   .2841112   2421.949     0.000   1.000    -4746.649   4747.218
     ins_10 |   .1573964   2471.544     0.000   1.000     -4843.98   4844.295
      _cons |   -1.39156   .2878738    -4.834   0.000    -1.955783   -.8273381
------------------------------------------------------------------------------
(Outcome emploi==IndComm is the comparison group)
```

Les résultats de commande « mlogit », dans le cas où la variable dépendante est multinomiale (codée de 0 ou 1 jusqu'à *n*), se présentent comme ceux de la commande « logit », sous forme de coefficients. On remarque cependant que la catégorie de référence (*comparison group*) pour la variable dépendante est "indépendante dans le commerce" (*emploi==IndComm*) : cette catégorie a été choisie pour référence parce qu'elle regroupe le plus grand nombre d'individus dans l'échantillon. Les coefficients sont donnés par rapport à cette catégorie. Pour obtenir

les coefficients par rapport à une autre catégorie, par exemple les salariées du secteur moderne (catégorie codée 1), on peut utiliser l'option « basecategory(1) » ou bien « basecategory(SalForm) » selon le code ou le label utilisé pour la variable dépendante "emploi".

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001) basecategory(1)
(listing omis)
```

Pour obtenir toutes les combinaisons de coefficients de chaque catégorie par rapport aux autres, on exécutera la commande « mcross » immédiatement après la commande « mlogit » :

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001)
(listing omis)
. mcross
```

emploi	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
SalInfl-SalForm						
gener2	.8184803	.6432916	1.272	0.203	-.4423481	2.079309
gener3	.8116916	.6644809	1.222	0.222	-.490667	2.11405
ins_2	-.8963147	.8604659	-1.042	0.298	-2.582797	.7901675
ins_3	-1.391462	.9198102	-1.513	0.130	-3.194257	.4113326
ins_4	-.7264767	.7486157	-0.970	0.332	-2.193737	.7407831
ins_5	-3.10587	1.155985	-2.687	0.007	-5.371559	-.8401821
ins_6	-2.907059	.772565	-3.763	0.000	-4.421258	-1.392859
ins_7	-2.100652	.9688082	-2.168	0.030	-3.999482	-.2018231
ins_8	-17.01339	1247.216	-0.014	0.989	-2461.511	2427.484
ins_9	-17.88494	1742.842	-0.010	0.992	-3433.792	3398.023
ins_10	-18.06201	1795.516	-0.010	0.992	-3537.208	3501.084
_cons	.6520596	.5467073	1.193	0.233	-.419467	1.723586
IndAutre-SalForm						
gener2	-.8862183	.6388955	-1.387	0.165	-2.13843	.3659938
gener3	-1.015436	.6774909	-1.499	0.134	-2.343293	.312422
ins_2	-.5832091	.8734144	-0.668	0.504	-2.29507	1.128652
ins_3	-.5148686	.9405921	-0.547	0.584	-2.358395	1.328658
ins_4	-.4471813	.8513035	-0.525	0.599	-2.115706	1.221343
ins_5	-1.791339	.9267477	-1.933	0.053	-3.607731	.0250531
ins_6	-2.279985	.8955531	-2.546	0.011	-4.035237	-.5247329
ins_7	-15.07131	738.8458	-0.020	0.984	-1463.183	1433.04
ins_8	-16.63652	1542.523	-0.011	0.991	-3039.926	3006.653
ins_9	-17.53483	2112.454	-0.008	0.993	-4157.869	4122.8
ins_10	-17.29729	2166.383	-0.008	0.994	-4263.329	4228.734
_cons	1.287957	.4969501	2.592	0.010	.3139528	2.261961
IndAutre-SalInfl						
gener2	-1.704699	.589517	-2.892	0.004	-2.860131	-.5492666
gener3	-1.827127	.6245075	-2.926	0.003	-3.05114	-.6031151
ins_2	.3131056	.8034074	0.390	0.697	-1.261544	1.887755
ins_3	.8765937	.9069907	0.966	0.334	-.9010755	2.654263
ins_4	.2792954	.7186064	0.389	0.698	-1.129147	1.687738
ins_5	1.314531	1.289501	1.019	0.308	-1.212844	3.841907
ins_6	.6270741	.9890586	0.634	0.526	-1.311445	2.565593
ins_7	-12.97066	738.846	-0.018	0.986	-1461.082	1435.141
ins_8	.3768704	1983.664	0.000	1.000	-3887.534	3888.287
ins_9	.3501003	2738.606	0.000	1.000	-5367.218	5367.919
ins_10	.7647115	2813.732	0.000	1.000	-5514.049	5515.579
_cons	.6358975	.4260222	1.493	0.136	-.1990907	1.470886

On peut donc estimer l'effet des variables "groupe de générations" et "instruction" sur le fait d'être salariée dans le secteur informel plutôt que dans le secteur formel, d'être indépendante hors commerce plutôt que salariée dans le secteur formel, et d'être indépendante hors commerce plutôt que salariée dans le secteur informel. Les commandes « mlogit » et « mcross » permettent de calculer les coefficients du modèle multinomial logit pour chaque paire de modalités de la variable dépendante.

Pour obtenir les résultats selon la présentation du modèle logistique, il suffit d'ajouter aux commandes « mlogit » et « mcross » l'option « rrr » :

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001) rrr
(listing omis)
. mcross, rrr
(listing omis)
```

b) L'interprétation des coefficients

Les résultats du modèle multinomial logit sont plus difficiles à interpréter que ceux du modèle logistique. Il faut non seulement raisonner avec plusieurs variables indépendantes (explicatives), elles-mêmes le plus souvent multinomiales, mais il faut aussi ajouter une dimension supplémentaire, la variable dépendante étant scindée en plusieurs catégories.

Une des premières difficultés consiste à repérer les coefficients qui identifient une prédiction totale (d'échec ou de succès). Lorsqu'on avait utilisé la commande « mlogit » pour contourner le problème des prédictions totales dans le cas du modèle logistique, les coefficients tendant vers l'infini étaient facilement repérés par leur valeur anormalement élevée, à laquelle correspondaient un écart-type lui aussi élevé et un degré de significativité proche de l'unité. Observons ce qu'il en est maintenant de l'effet des niveaux d'instruction "ins_9" (terminale) et "ins_10" (supérieur) :

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001)
(listing partiel : voir plus haut le listing complet)
------------------------------------------------------------------------------
      emploi |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
SalForm      |
(etc)
       ins_9 |    17.81895   1184.642     0.015   0.988    -2304.036    2339.674
      ins_10 |    17.45469   1189.671     0.015   0.988    -2314.258    2349.167
(etc)
-------------+----------------------------------------------------------------
SalInfl      |
(etc)
       ins_9 |   -.0659891   2107.338    -0.000   1.000    -4130.373    4130.241
      ins_10 |   -.6073151   2153.879    -0.000   1.000    -4222.132    4220.917
(etc)
-------------+----------------------------------------------------------------
IndAutre     |
(etc)
       ins_9 |    .2841112   2421.949     0.000   1.000    -4746.649    4747.218
      ins_10 |    .1573964   2471.544     0.000   1.000     -4843.98    4844.295
(etc)
------------------------------------------------------------------------------
(Outcome emploi==IndComm is the comparison group)
```

Les coefficients pour la catégorie "SalForm", calculés par rapport à la catégorie de référence "IndComm", tendent vers plus l'infini. En fait, aucune femme indépendante dans le commerce n'a un niveau d'instruction aussi élevé. Par contre, certaines femmes salariées dans le secteur formel ont ce niveau d'instruction : la probabilité pour ces femmes instruites d'être salariées dans le secteur formel plutôt que commerçantes est donc égale à 1.

Les coefficients pour la catégorie "SalInfl", calculés par rapport à la catégorie de référence "IndComm", sont au contraire très faibles, bien que leur écart-type soit très élevé, et leur degré de significativité proche de l'unité. Rappelons qu'aucune femme commerçante n'a un niveau d'instruction "ins_9" ou "ins_10" : pour ces niveaux d'instruction le risque est nul d'être commerçante. Les coefficients pour "SalInfl" doivent donc s'interpréter comme une mesure de la différence avec un risque nul : si cette différence n'est pas significative, cela veut dire que le risque d'être salariée dans le secteur informel est aussi nul pour les femmes d'instruction élevée.

Le même raisonnement s'applique pour les indépendantes hors commerce qu'on ne rencontre jamais chez les femmes de niveau d'instruction élevé. En fait, les femmes ayant poursuivi leurs études jusqu'en classe de terminale ou dans l'enseignement supérieur se rencontrent exclusivement chez les salariées du secteur moderne.

c) Aide à l'interprétation

La difficulté de l'interprétation provient du fait qu'on est en présence d'un modèle à plusieurs équations. Chaque équation du modèle multinomial logit est interprétable, mais contrairement au modèle logistique, l'effet global d'une variable explicative n'apparaît pas immédiatement.

La méthode des "prédictions recyclées" (*recycled predictions method*) permet de dégager l'effet d'une variable indépendante sur la distribution de l'échantillon entre les différentes catégories de la variable dépendante. Cette méthode est décrite pas à pas dans la section [5s] mlogit du manuel de STATA, et nous l'avons systématisée dans une nouvelle commande « recpred ». Il s'agit de répartir l'échantillon entre les catégories de la variable dépendante, selon les probabilités associées à chacune des modalités d'une variable indépendante, en maintenant l'effet des autres variables indépendantes constant. La syntaxe de la nouvelle commande est simplement « *recpred liste_variables* » où la liste est une série de modalités d'une variable polytomique.

Un exemple aidera à comprendre l'intérêt de la commande « recpred ». Pour évaluer l'effet du groupe de générations après avoir obtenu les résultats de la commande « mlogit », on donnera l'instruction suivante :

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001)
(listing omis)

. recpred gener3 gener2
Method of recycled predictions after mlogit.
------------------------------------------------------------------------------
UNADJUSTED                    Dependent variable  : emploi
------------------------------------------------------------------------------
(% in row)    Obs  SalForm  SalInfl  IndComm IndAutre
  gener3 |     62    38.71    33.87    14.52    12.90
  gener2 |     90    24.44    23.33    42.22    10.00
  [Ref.] |     98    10.20     9.18    61.22    19.39
------------------------------------------------------------------------------
ADJUSTED                      Dependent variable  : emploi
------------------------------------------------------------------------------
(% in row)          SalForm  SalInfl  IndComm IndAutre
  gener3 |            27.38    39.59    19.17    13.87
  gener2 |            21.57    24.09    44.36     9.98
  [Ref.] |            19.08     8.59    52.73    19.60
```

Le premier tableau des valeurs non ajustées donne simplement la répartition de la variable dépendante dans chacun des groupes de générations. Parmi les 98 femmes de la génération 1930-44, par exemple, 10,20 % sont salariées dans le secteur formel. Le deuxième tableau montre que ce pourcentage serait de 19,08 % si toutes les autres caractéristiques de l'échantillon (hormis la génération) étaient maintenues constantes (égales à la moyenne dans l'échantillon total), c'est-à-dire toutes choses égales par ailleurs.

Les distributions des valeurs ajustées sont comparables d'une ligne à l'autre, et les différences entre ces distributions sont attribuables à l'effet de génération. On voit par exemple que la part du salariat est plus importante chez les jeunes femmes (ce qu'on pouvait déjà observer sur les distributions non ajustées), mais que c'est surtout le salariat informel qui prend de l'importance chez les jeunes. En effet d'après les données ajustées, le pourcentage de salariées dans le secteur formel est multiplié par 27,38/19,08=1,4 entre le premier groupe de générations et le troisième (contre 3,8 dans les données brutes), alors que le pourcentage de salariées dans le secteur informel est multiplié par 39,59/8,59=4,6 (contre 3,7 dans les données brutes). Malgré ces fortes variations, la répartition entre salariat formel et informel n'est pas significativement différente entre les deux générations comme le rappellent les résultats de la régression :

```
. mcross
(listing partiel : voir plus haut le listing complet)
------------------------------------------------------------------------------
  emploi |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
SalInfl-SalForm

  gener3 |   .8116916   .6644809     1.222   0.222      -.490667     2.11405
(etc)
```

En fait, c'est plutôt le statut d'indépendante par opposition au statut de salariée qui fait la différence entre la génération la plus âgée et la génération la plus jeune :

```
. mlogit
(listing partiel : voir plus haut le listing complet)
Multinomial regression                          Number of obs =     250
(Log Likelihood tolerance 1.00e-07)             chi2(33)      =  148.31
                                                Prob > chi2   =  0.0000
Log Likelihood = -251.26817                     Pseudo R2     =  0.2279
------------------------------------------------------------------------------
  emploi |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
SalForm  |
  gener3 |   1.791126   .6429739      2.786   0.005       .5309206    3.051332
(etc)
---------+--------------------------------------------------------------------
SalInfl  |
  gener3 |   2.602818   .5591844      4.655   0.000       1.506837    3.698799
(etc)
---------+--------------------------------------------------------------------
IndAutre |
  gener3 |   .7756905   .6002216      1.292   0.196      -.4007222    1.952103
(etc)
```

L'effet de la modalité "gener3" n'est pas significatif entre les catégories des indépendantes commerçantes et non commerçantes.

L'interprétation des résultats du modèle multinomial logit est, comme on le constate, assez complexe. Elle l'est d'autant plus qu'il n'y a pas actuellement de diagnostic statistique ou graphique comme pour les modèles de régression simple ou logistique. La difficulté du diagnostic provient toujours du fait que plusieurs équations interviennent dans le modèle. Le développement des techniques de diagnostic est donc très attendu.

5. Conclusions sur le modèle logistique

Le modèle logistique est certainement le modèle le plus utilisé pour analyser des ratios bien que la fonction logistique soit particulièrement séduisante. Même si des précautions doivent être prises avant son utilisation, ce modèle reste incontournable.

Malgré les difficultés d'interprétation, le modèle multinomial logit nous a permis d'introduire la notion de risques multiples. Ce modèle est une bonne introduction aux risques concurrents dans l'analyse des biographies.

En utilisant un modèle de la classe des modèles log-linéaires, que ce soit un modèle de type "logit" ou un de ses dérivés, il faut bien se rappeler, au moment de la conceptualisation du problème, que la caractéristique étudiée doit être considérée comme fixe et non comme une caractéristique acquise : le temps ne doit pas intervenir dans l'analyse. Par exemple, pour l'étude de la proportion de salariées, le temps mis pour obtenir un emploi n'a pas été pris en compte ; une unique observation a été faite à un moment donné. On a simplement mesuré la présence des salariées parmi les femmes occupant leur premier emploi, et non le risque qu'une femme devienne salariée. Pour mesurer ce risque il aurait fallu prendre en compte les femmes qui n'ont pas encore obtenu un emploi, c'est-à-dire qu'il aurait fallu situer le moment de l'acquisition du premier emploi dans la vie des femmes. Dans le chapitre suivant, nous introduisons cette dimension fondamentale du temps.

EXERCICE 4

Comment doit-on interpréter le fait que les coefficients pour "ins_9" et "ins_10" tendent vers moins l'infini pour les catégories "SalInfl-SalForm" et "IndAutre-SalForm" dans les résultats donnés par la commande « mcross » ? Pourquoi les coefficients de la catégorie "IndAutr-SalInfl" ne tendent-ils pas aussi vers moins l'infini ?

L'option « basecategory() » permet de choisir une catégorie de référence qui ne soit pas forcément la catégorie majoritaire. Exécutez la commande :

```
. mlogit emploi gener2 gener3 ins_2-ins_10, ltol(0.000001) basecategory(1)
```

Comparez avec les résultats sans cette option. Comment interprétez-vous les coefficients pour la catégorie "IndComm" ? Quelle différence voyez-vous entre d'une part les coefficients pour les catégories "SalInfl" et "IndAutre", et d'autre part les coefficients pour les catégories "SalInfl-SalForm" et "IndAutre-SalForm" obtenus précédemment avec la commande « mcross » ?

Pour vous aidez à interpréter l'effet de l'instruction, exécutez la commande « recpred » :

```
. recpred ins_2-ins_10
```

À partir de quels niveaux d'instruction l'accès au salariat formel est-il significativement différent de l'accès à d'autres emplois ? Les probabilités d'accès à tel ou tel emploi sont-elles significativement différentes pour les femmes de niveau d'instruction "ins_4" (cours moyen) par rapport aux femmes sans instruction (niveau de référence) ?

CHAPITRE 5

LES TABLES DE SÉJOUR

Le modèle de régression constitue le premier élément à la base des modèles de régression à risques proportionnels. La table de séjour constitue le second élément à la base de ces modèles. Avec cet élément, le temps est introduit dans l'analyse, ce qui nous rapproche au plus près des relations causales. Ce chapitre explique les concepts fondamentaux de risque dans le temps, de troncature et d'événements concurrents.

1. La description du temps et de l'événement

Comme on l'a souligné au début de ce manuel, le temps est à l'origine de bien des confusions et erreurs en statistique. Cette dimension est souvent prise en compte implicitement dans l'analyse, mais le chercheur "oublie" parfois d'en préciser les limites : le lecteur doit faire tout un travail de réflexion (sur des données qu'il ne connaît généralement pas) pour rétablir explicitement les hypothèses du modèle, relativement au temps. Pour soi-même autant que pour son lecteur, lorsqu'on traite ses propres données, il faut être particulièrement précis concernant les définitions de l'événement et du temps écoulé avant cet événement.

Prenons un exemple : la proportion de salariées parmi les actives occupées. Si les jeunes sont plus souvent salariées que les travailleuses âgées, cela veut-il dire que le salariat se généralise dans les générations récentes, ou bien que les travailleuses passent d'abord par une période de salariat avant de se mettre à leur compte, ou bien encore que les entreprises n'emploient que des jeunes, les autres n'ayant d'autre choix que de se mettre à leur compte ? La difficulté de l'interprétation de la variation de la proportion de salariées provient clairement du temps. L'état de salariée n'est pas un état permanent depuis l'entrée sur le marché du travail, du fait de la mobilité professionnelle. L'analyse sur la proportion de salariées ne peut être valable qu'au moment de l'observation : on ne peut rien en déduire sur

l'évolution de cette proportion dans le temps, même si l'on tient compte des générations.

Une telle analyse sera cependant plus correcte si on considère, comme on l'a fait au chapitre précédent, non pas le moment de l'enquête (toutes les femmes ne sont pas observées au même moment de leur itinéraire professionnel), mais le moment de l'entrée sur le marché du travail, qui est défini pour toutes les générations.

En somme, à chaque fois que l'on a à évaluer un problème, il faut **se poser la question du moment d'observation et du rapport au temps de la caractéristique étudiée dans la population**. Chaque fois qu'une caractéristique variant dans le temps est analysée par un modèle qui, lui, ne tient pas compte du temps, il faut questionner la validité des résultats qui sont présentés.

a) Groupe à risque et période d'observation

Dans le modèle logistique ou dans tout autre modèle sur les ratios, le risque est généralement conjugué au présent : on mesure la probabilité pour que tel individu (ou groupe) ait telle caractéristique au moment de l'observation de l'ensemble de l'échantillon. Le temps est concentré en un point donné, comme par exemple pour le calcul de la proportion de salariées parmi les actives occupées.

Cependant rien n'oblige à limiter le temps d'observation à un instant unique (le moment de l'enquête, ou l'entrée dans la vie active, ou le recrutement dans l'usine, etc). En fait, on peut très bien considérer un événement qui advient après une période de temps plus ou moins longue.

Dans ce cas-là, la caractéristique est **acquise** par une partie de la population au cours d'un intervalle de temps : au début de l'intervalle, aucun individu n'a connu l'événement et en fin d'intervalle, un certain nombre d'individus l'ont connu, les autres non. La population en début d'intervalle est appelée **groupe à risque** (de connaître l'événement). La population en fin d'intervalle est modifiée par l'événement, de sorte que les individus qui n'ont pas encore connu l'événement en fin d'intervalle constituent le groupe à risque au début de l'intervalle de temps suivant, et ainsi de suite.

Étudier une caractéristique acquise (un événement qui survient dans une population) a quelques conséquences sur la collecte des données : il faudra situer les événements dans le temps et non pas simplement demander à chaque individu s'il a connu ou non l'événement dans sa vie. Cela est indispensable pour constituer une population à risque homogène dans le temps.

Considérons par exemple une population où le mariage se produit avec une fréquence constante dans chaque génération, et le plus souvent entre 20 et 40 ans. Les conclusions seront très différentes si l'on interroge une population de 15 à 60 ans et si l'on interroge seulement les plus de 50 ans. Dans le premier cas, on interroge certains individus qui n'ont pas eu encore le temps de se marier, et dans le second, des individus qui ont largement eu le temps de le faire. Dans le premier cas, le processus est en cours d'évolution, et dans le second, le processus est quasiment terminé et presque tout le monde est marié. La mesure de la nuptialité ne sera pas équivalente dans les deux cas, quand bien même toute la population se marierait avant 50 ans.

Cela provient du fait que la durée d'observation (la vie, ou plus précisément la période de soumission au "risque" de se marier) n'est pas la même d'un individu à l'autre, selon son âge au moment de l'enquête. Cela signifie-t-il que l'on ne peut observer l'intensité d'un phénomène que lorsque plus aucun individu n'est soumis au risque (à 50 ans, dans l'exemple du mariage, ou plus généralement à son décés) ? Heureusement non, il y a des solutions.

b) Le diagramme de Lexis

Lorsque l'on veut étudier des biographies, le point minimal de comparaison est généralement la génération. Telle cohorte s'est mariée plus tôt que telle autre ; les vieilles générations entraient en activité plus tard que les jeunes, etc. Scinder l'échantillon en **cohortes** ou en générations est une manière de constituer des groupes homogènes du point de vue du temps : les individus qui composent chaque cohorte sont supposés avoir connu les mêmes conditions de vie, le même environnement, aux mêmes âges.

Pour bien se représenter ce que cela signifie, rien ne vaut un diagramme de Lexis. Ce diagramme permet de lire facilement quels sont les intervalles de temps couverts, exprimés en âge ou en année civile, pour chaque génération ou cohorte. Voyons à la figure 1 ce qu'il en est de l'enquête rétrospective IFAN-ORSTOM de Dakar. Les trois groupes de générations (qui constituent les strates de notre échantillon) forment des "couloirs" parallèles et obliques dans le diagramme. On les suit depuis leur date de naissance (axe horizontal) selon une ligne oblique (une "ligne de vie" en quelque sorte) jusqu'à la date d'enquête (axe vertical).

Ainsi les trois cohortes peuvent être comparées à l'âge de 15 ans entre les années 1945 et 1979. À l'âge de 25 ans, on peut encore les comparer sur la période 1955-1989, mais à 35 ans, on pourra seulement comparer la situation des générations 1930-44 et 1945-54 sur la période 1965-1989.

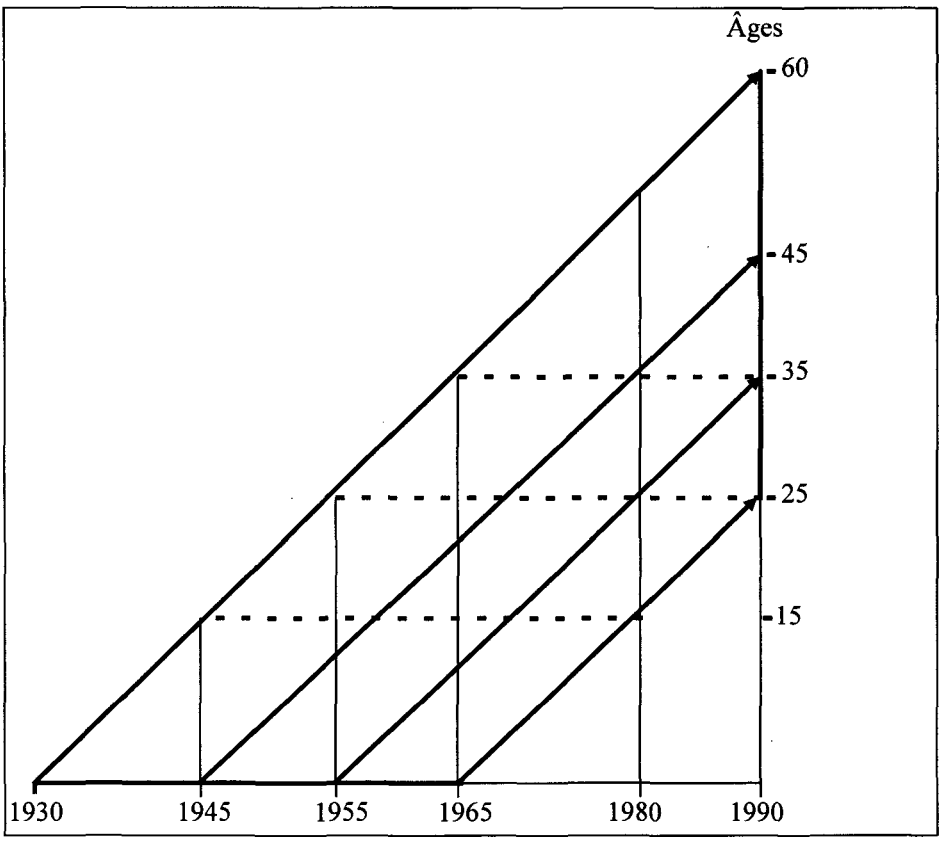


Figure 1. Diagramme de Lexis figurant les trois générations de l'enquête IFAN-ORSTOM

c) La notion de troncature à droite

Lorsqu'on travaille avec des données d'enquête, les individus ne sont pas observés sur toute leur vie, mais seulement jusqu'au moment de l'enquête. Lorsque l'individu n'a pas encore connu l'événement étudié, on appelle cette interruption de l'observation une "**troncature à droite**", parce qu'elle se situe à droite sur l'échelle du temps (qui, par convention, se déroule de gauche à droite). L'individu sort alors de l'observation à la date d'enquête ; il n'est plus soumis au risque. C'est le cas pour une bonne partie de notre population des 15-60 ans, en ce qui concerne le mariage.

Pour **résoudre le problème de l'hétérogénéité de la population soumise au risque**, il faut calculer les risques sur des intervalles plus courts que l'ensemble de la période de soumission au risque (la vie). Considérons trois individus âgés de 50, 30 et 25 ans sur le diagramme de Lexis de la figure 2 :

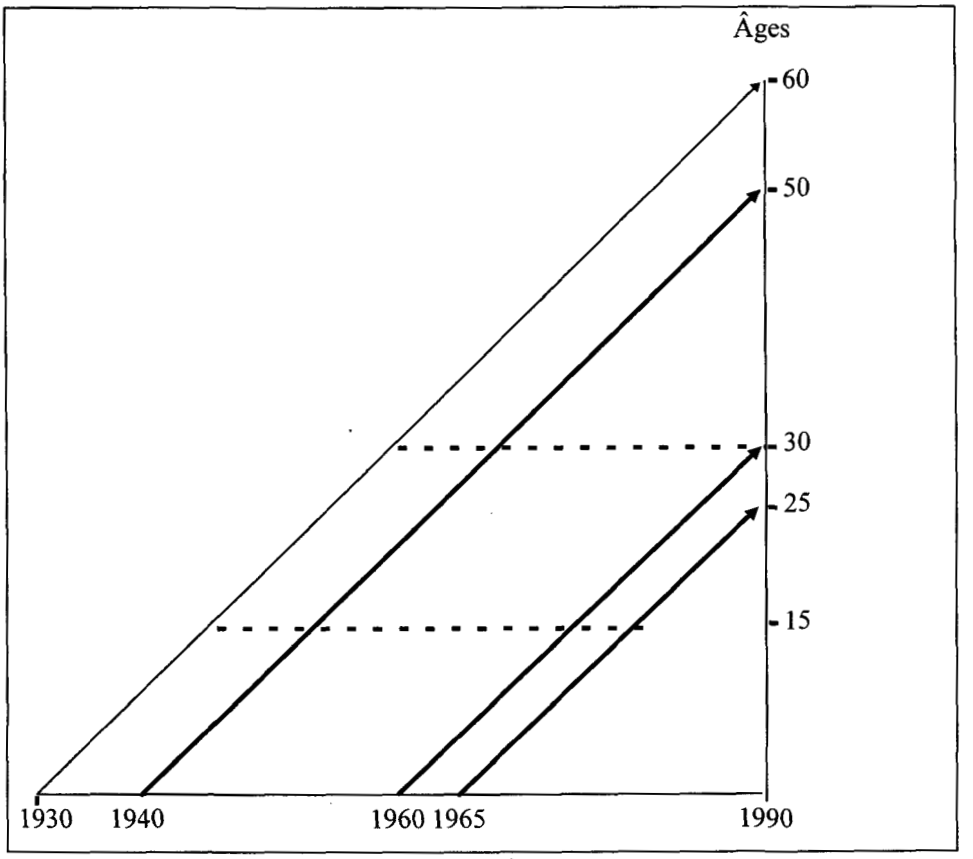
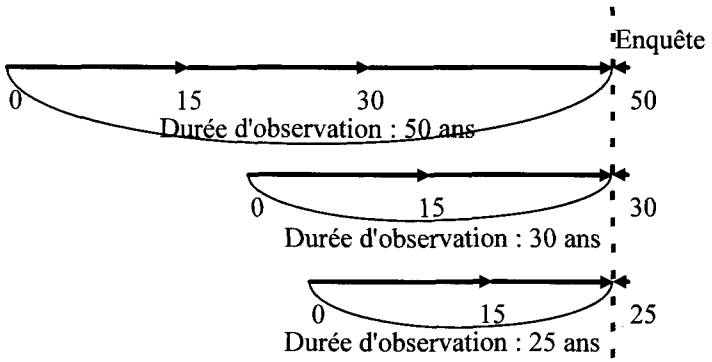
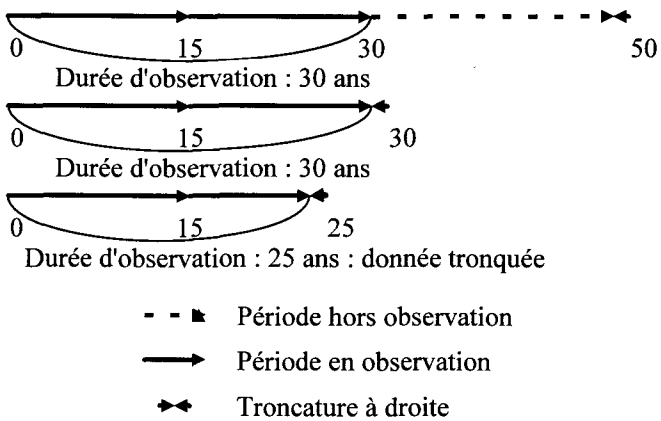


Figure 2. Diagramme de Lexis. Représentation de trois "lignes de vie"

Les périodes d'observation de ces individus peuvent être représentées par des lignes horizontales :



Pour ces individus, la date d'enquête représente une troncature à droite et leur période de soumission au risque (leur vie) varie de 25 à 50 ans. Si l'on réduit le temps de soumission au risque, on réduira ces troncatures à droite : ainsi en ne considérant que la période entre 15 et 30 ans, toutes les personnes âgées de 30 ans et plus constituent une population homogène par rapport au temps de soumission au risque :

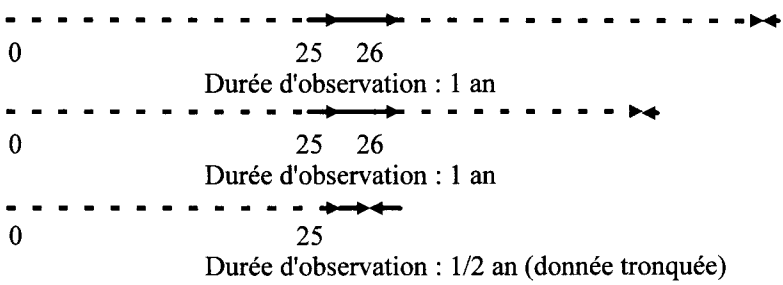


La période de soumission au risque de l'individu âgé de 25 ans reste tronquée à droite. On peut construire des intervalles de temps plus courts pour établir une série de probabilités de connaître l'événement dans chacun de ces intervalles : ainsi la probabilité de se marier entre 15 et 24 ans, entre 25 et 35 ans, etc.

En réduisant la largeur des intervalles de temps, on réduit du même coup, pour chaque intervalle, la proportion de troncatures dues au moment de l'enquête parmi les individus soumis au risque. En fait, on pourra même inclure dans la population soumise au risque les individus "tronqués" (qui sortent de l'observation), mais seulement pendant la moitié de cet intervalle. On fait alors l'hypothèse que les troncatures dues à la date d'enquête se répartissent uniformément au cours de l'intervalle. Il suffira alors de choisir la largeur de l'intervalle de façon à ce que cette

hypothèse d'uniformité sur l'intervalle de temps reste raisonnable quel que soit l'intervalle.

Prenons une année d'intervalle, comme on le fait le plus souvent en sciences sociales. On observe les individus entre 25 et 26 ans :

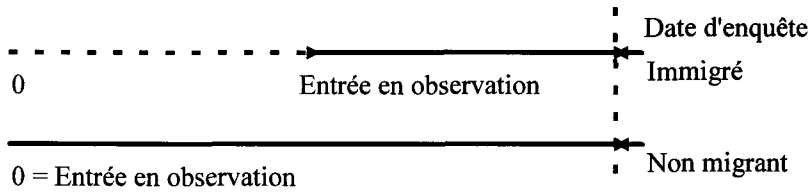


L'observation correspondant à l'individu qui avait 25 ans et demi au moment de l'enquête va se trouver tronqué, une demi-année après son anniversaire.

Si les événements ont lieu sur quelques mois, l'année n'est pas un intervalle raisonnable : ce serait le cas, par exemple, de la durée du chômage avant la reprise d'une activité rémunérée : la plupart des chômeurs retrouvent un emploi au cours des deux premières années de chômage. Mais dans bien des cas, l'année est un intervalle suffisamment court, par exemple pour les événements suivants : trouver un premier emploi, se marier, avoir un premier enfant, etc. La plus grande précision est souhaitable, mais son coût et même la possibilité d'obtenir l'information auprès des enquêtés (surtout par le recueil rétrospectif), empêchent généralement de faire des analyses plus précisément qu'à l'année près. Il n'est d'ailleurs pas forcément nécessaire d'être plus précis, pourvu que l'ordre de succession des événements soit conservé.

d) La notion de troncature à gauche

Hormis le problème, relativement facile à résoudre, des troncatures à droite généralement provoquées par le moment de l'enquête, il se pose le problème des troncatures à gauche : de la même façon qu'un individu peut sortir de l'observation avant les autres, un individu peut rentrer dans le champ d'observation au cours de l'intervalle de temps considéré. Ainsi, un immigré rentrera dans la population après sa naissance. L'immigré n'est soumis au risque de connaître l'événement qu'à partir du moment de son immigration dans la population étudiée (population représentée par l'échantillon), alors qu'il était soumis au risque dans une autre population (non étudiée) avant son immigration :



Le traitement de ces troncatures n'est pas seulement une question technique : il ne suffit pas, par exemple, de tenir compte des observations tronquées à gauche pendant la moitié de l'intervalle de temps (comme on l'a fait pour les troncatures à droite). En effet, le problème vient du fait qu'on ne peut pas considérer *a priori* pour toutes les analyses que l'ensemble des individus de l'échantillon, dont le temps d'observation est tronqué à gauche ou non, forment une population homogène.

En effet, on ne peut mesurer le risque de connaître un événement qu'à partir du moment où l'individu entre dans la population soumise au risque. Or, pour que les observations tronquées à gauche (les immigrés) soient parvenus dans la population soumise au risque, il faut qu'ils n'aient pas subi l'événement dans leur population d'origine avant le moment de l'immigration. On introduit donc un **biais de sélection par la migration** si l'on tient compte des immigrants au même titre que les autres individus de l'échantillon. On peut de plus imaginer (c'est en fait un cas fréquent en sciences sociales) que la troncature à gauche soit un événement (l'immigration) lui-même corrélé avec l'événement étudié. La prise en compte des troncatures à gauche n'est pas impossible, mais il n'y a pas de solution unique parce que cela touche au problème plus général de l'hétérogénéité de la population soumise au risque par rapport à une définition de l'événement que l'on veut précise.

En somme, on se trouve confronté à deux problèmes majeurs en traitant des données d'enquête : d'abord, les enquêtés peuvent **sortir par émigration** de la population soumise au risque pour un temps plus ou moins long, et pour une raison plus ou moins liée au phénomène que l'on étudie ; ensuite, les enquêtés peuvent **entrer par immigration** dans le lieu d'enquête après le début d'observation et connaître ainsi l'événement étudié.

La figure 3 schématise les différentes possibilités d'entrée et de sortie de la population soumise au risque, et la population finale qui en résulte. Parmi les enquêtés, on peut distinguer différents groupes selon le statut migratoire :

- les **sédentaires** forment une population relativement homogène du point de vue de l'unité de lieu et de l'unité de temps (ils n'ont pas migré). La date d'enquête, qui n'est pas en principe corrélée avec l'événement étudié, est la troncature à droite par excellence ;
- les **émigrants** peuvent aussi être considérés dans la population soumise au risque dans la mesure où ils ne se distinguaient pas des sédentaires jusqu'au moment de la migration. L'émigration est alors traitée comme une

troncature à droite. Cependant, l'émigration peut aussi être corrélée avec l'événement étudié. De plus, il faut être conscient qu'un **biais de sélection** s'opère au moment de l'enquête : parmi les émigrés, on n'interroge que les migrants de retour. Une partie des émigrés est perdue au moment de l'enquête, surtout parmi les émigrés de fraîche date qui ont peu de temps pour éventuellement revenir au lieu d'enquête ;

- une partie de l'échantillon a pu venir au lieu d'enquête après la date fixée pour le début de l'observation. Ces **immigrants** (comme d'ailleurs les migrants de retour) ont pu connaître l'événement en concurrence avec les sédentaires.

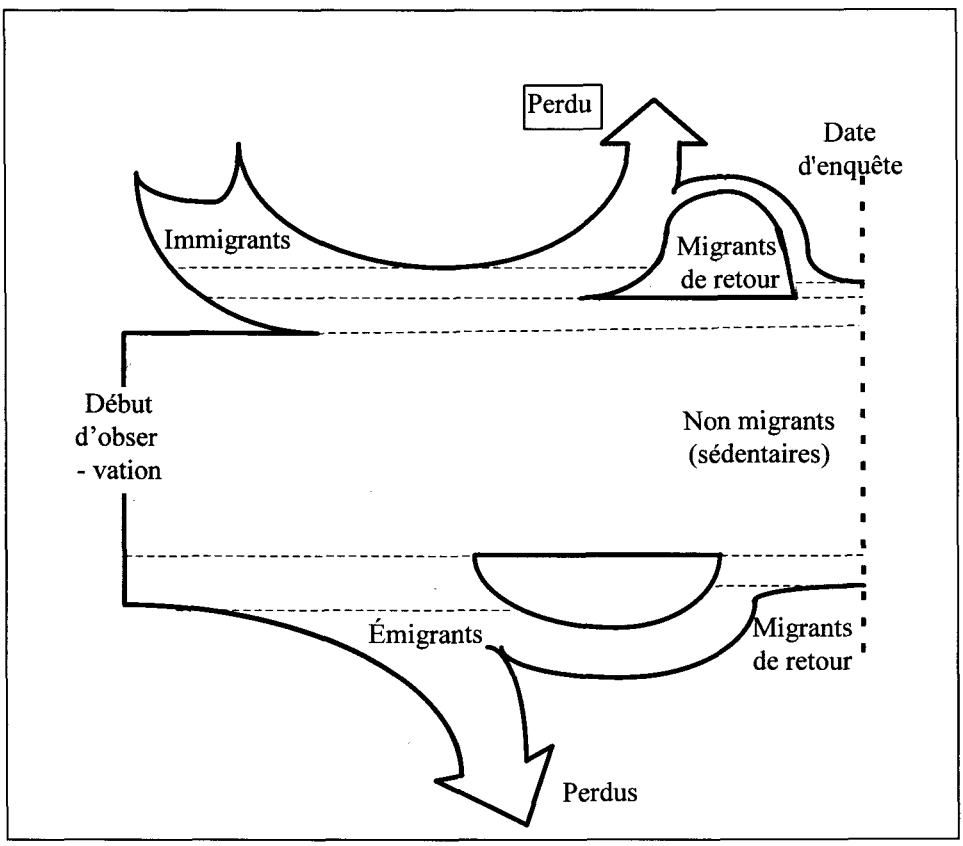


Figure 3. Entrées dans une population soumise au risque et sorties de cette population

2. La conceptualisation d'un problème et son application : l'entrée dans la vie active

Jusqu'alors, on a fait comme si les définitions des paramètres de l'analyse allaient de soi. Or, chacune de ces définitions doit être précisée pour justifier la pertinence du modèle. De nombreuses erreurs d'interprétation proviennent souvent de la négligence de cette étape essentielle de conceptualisation. On ne saurait trop insister sur ce point : **l'essentiel de l'analyse statistique consiste à réfléchir à l'événement, à la durée avant l'événement (c'est-à-dire au temps d'observation) et aux variables indépendantes.** Le calcul informatique d'estimation est un aspect relativement secondaire dans cette démarche. D'ailleurs, les programmes d'estimations ne sont généralement pas accessibles à l'utilisateur du logiciel qui ne maîtrise pas cette étape.

Dans la suite de ce chapitre, la démarche de conceptualisation de l'entrée dans la vie active sera systématisée, afin d'aboutir à des principes suffisamment généraux pour s'appliquer au plus grand nombre de problèmes des sciences sociales. Une série de questions servira à guider le lecteur pour qu'il soit en mesure de concevoir un modèle acceptable et de le présenter à la communauté scientifique, en anticipant les critiques.

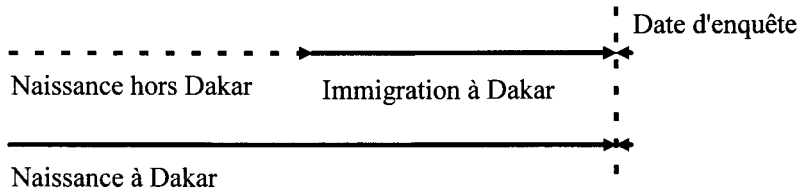
a) L'événement et la population soumise au risque : l'entrée en observation

On a vu plus haut que pour l'analyse d'un événement, il est nécessaire de constituer des cohortes homogènes selon le temps. Les risques doivent être calculés sur une population qui est soumise à ces risques pendant le même intervalle de temps. Or, cette homogénéité n'est pas si facile à obtenir en sciences sociales. Quelques exemples tirés de l'enquête IFAN-ORSTOM vont permettre de le montrer.

Le début d'observation d'une biographie n'est pas forcément la naissance de l'individu. Tout dépend de la constitution de l'échantillon et de l'événement étudié. L'étendue de l'échantillon (la population qu'il est sensé représenter) constitue l'espace dans lequel vont s'inscrire les événements. On n'étudie pas de la même façon la biographie des individus hors de Dakar et à Dakar. Dans le premier cas, les événements se sont produits dans une population que l'échantillon ne représente pas (la population hors de Dakar) : il ne sera donc pas possible, par exemple, d'étudier l'entrée en activité hors de Dakar, même si l'échantillon est composé en partie d'individus qui ont eu leur première activité hors de Dakar. Dans le cas de la

biographie à Dakar, tous les événements qui se sont produits dans cette ville pourront être rapportés, en principe, à la population de Dakar, représentée par l'échantillon.

En ce qui concerne l'immigration, on a le schéma suivant :



On doit distinguer les individus soumis au risque pendant toute leur vie jusqu'à obtenir leur premier emploi, et ceux qui sont entrés dans la population dakaroise "en cours de route", c'est-à-dire les individus qui ont immigré à Dakar entre leur naissance et leur premier emploi à Dakar. D'ailleurs, la définition du premier emploi à Dakar doit être prise au sens strict : pour les migrants il ne s'agit pas forcément du premier emploi de leur vie, ils ont pu travailler hors de Dakar avant de venir dans cette ville. **Les définitions temporelles et spatiales sont fortement liées entre elles** : l'événement (le premier emploi) ne peut se définir sans l'espace (Dakar) dans lequel il se produit, c'est-à-dire la population soumise au risque à laquelle il est rapporté.

La durée d'observation avant le premier emploi aura donc une signification fort différente selon la population de référence. Dans le cas des Dakarois de naissance, le premier emploi à Dakar correspond à leur entrée dans la vie active : ils pourront être suivis pendant toute leur formation. Pour les migrants, la durée prise en compte sera celle qui sépare le moment de la première migration à Dakar et l'obtention d'un emploi dans cette ville.

Une partie de ces migrants seront venus pour compléter leur formation à Dakar, les autres pour chercher directement un emploi : dans ce dernier cas, on mesure la durée de la recherche d'emploi (le chômage) des migrants et non pas leur entrée dans la vie active, même si pour certains, la migration correspond effectivement à la recherche d'un premier emploi. La migration à Dakar pour recherche d'emploi ne pourrait être étudiée en tant qu'événement de la biographie migratoire que si nous pouvions la rapporter à l'ensemble de la population soumise au risque de connaître ce type d'événement, c'est-à-dire à l'ensemble de la population du pays. Dans ce cas, le problème de la population soumise au risque se trouverait reporté sur les immigrants venus des autres pays.

On voit ici que les définitions sont inextricablement liées et renvoient à des **questions qui dépassent très largement le calcul statistique** : par exemple, il s'agit de définir le marché de l'emploi, le rôle de la formation sur ce marché, la signification de la migration de main-d'œuvre, etc. Il est bien rare en sciences

sociales que l'on puisse se soustraire à ce type de questions, et en fait, l'analyse statistique oblige à plus de rigueur dans la définition des concepts, qu'ils soient économiques, sociologiques, psychologiques, géographiques, historiques, etc.

En guise de garde-fou, le lecteur pourra répondre à une série de questions à chaque fois qu'il cherche à analyser un événement dans le temps. Dans le cas de l'entrée dans la vie active (accès au premier emploi), on aura la série de questions-réponses suivante :

Où l'événement a-t-il pu se produire ? (dans quel espace ?)

- Dans les limites de l'agglomération dakaroise.

À partir de quel moment de leur vie les individus de l'échantillon ont-ils pu connaître l'événement ? (Y a-t-il un événement initial qui doit être réalisé pour que l'événement étudié puisse avoir lieu ?)

- Les individus sont soumis au risque à partir de la naissance ou bien de l'arrivée à Dakar si cette arrivée a eu lieu avant l'entrée dans la vie active. Les individus doivent avoir connu une période de formation à Dakar ou bien avoir été présents à Dakar depuis leur naissance.

Quelle est la population qui peut effectivement avoir connu l'événement ?

- Les individus qui n'avaient encore jamais travaillé de leur vie.

Est-ce que l'événement est défini de la même façon pour tout l'échantillon ?

- L'entrée dans la vie active est définie de manière ambiguë pour les migrants qui n'avaient jamais travaillé avant de venir à Dakar. Leur inclusion dans l'échantillon se discute en fonction du type de séjour avant le premier emploi (formation ou chômage) et du temps de séjour.

b) L'événement et la sortie d'observation

Diverses raisons ont pu mener à la sortie de la population soumise au risque :

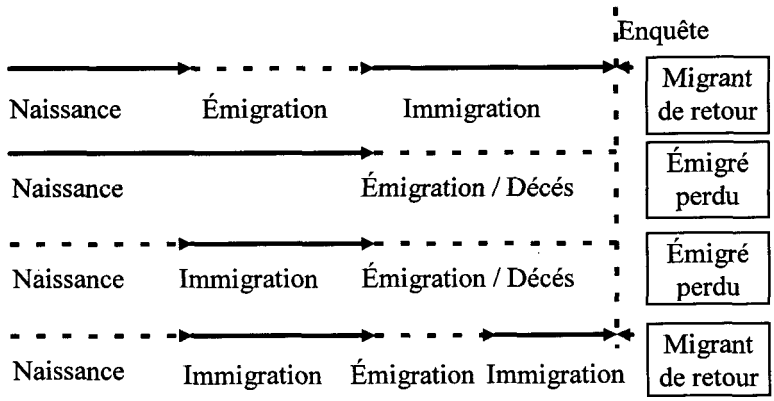
1. d'abord, l'**événement** lui-même (après avoir obtenu un premier emploi, on ne peut en connaître qu'un second !),

2. ensuite, des cas d'incapacité de connaître l'événement malgré la présence physique de l'individu dans l'échantillon (par exemple une maladie qui contraint le patient à l'inactivité) : il s'agit en principe de **sorties temporaires**,

3. et enfin, l'émigration, c'est-à-dire la sortie physique de la population soumise au risque, que l'on considère comme une **sortie définitive**.

Certains cas de sortie temporaire peuvent à la limite être considérés comme des sorties définitives : un individu peut être malade pour un temps, mais il peut aussi devenir invalide à vie. On verra plus loin que l'on peut traiter les cas d'une sortie temporaire, à l'aide d'une variable qui définira un risque nul. Mais le problème se pose plus sérieusement dans les cas de sorties définitives, c'est-à-dire de sorties qui interdisent le retour dans la population soumise au risque.

Si les cas d'incapacité définitive ne posent pas trop de problèmes (il suffit de considérer les individus concernés comme physiquement hors de la population soumise au risque), l'émigration est un phénomène plus délicat, car elle peut être suivie d'un retour dans la population soumise au risque. En fait, l'échantillon n'est en général constitué que de personnes présentes au moment de l'enquête, et donc, parmi les émigrés, de personnes ayant effectué une migration de retour. Par définition, les émigrés qui ne sont pas retournés au lieu d'enquête ne figurent pas dans l'échantillon, qu'ils aient connu l'événement étudié ou non, comme l'illustre le schéma suivant :



En ce sens, tout échantillon d'enquête rétrospective est nécessairement biaisé. Deux catégories de population ne peuvent être interrogées dans ce type d'enquête : les morts et les émigrants qui ne sont pas revenus au lieu d'enquête. Cela semble trivial, mais il est utile de le rappeler : **aucune enquête rétrospective ne peut se soustraire à des biais de sélection du fait de la survie ou de la sédentarisation** des individus ; c'est une limite inhérente à ces enquêtes. On peut seulement s'assurer que la procédure de sondage donne une image représentative des biographies des personnes présentes au moment de l'enquête. On peut espérer, dans un accès d'optimisme, que les itinéraires ne sont pas considérablement affectés par la mortalité (l'itinéraire des vivants représenterait assez bien celui des morts), mais il est en revanche très fréquent que l'itinéraire migratoire soit inextricablement lié au phénomène étudié.

Dans le cas des enquêtes prospectives longitudinales (de suivi, ou à passages répétés) ou bien dans le cas des registres de population, on évite ces biais de sélection du fait de la survie, puisque l'on dispose en principe des informations sur les individus sortis de l'échantillon. Le lecteur qui a la chance de disposer de données prospectives ou de registres, se trouve ainsi libéré d'un problème important en analyse des biographies. C'est d'ailleurs le principal avantage de ce type de données, mais le coût de ces opérations de collecte est en général assez élevé. Nous nous en tiendrons donc aux problèmes que posent les enquêtes rétrospectives à un seul passage, plus communes.

Dans les enquêtes rétrospectives, non seulement l'observation des immigrants et des émigrants est sélective, mais la migration de retour (émigration suivie d'un retour avant la date d'enquête) est aussi sélective. Dès lors, l'enquête ne peut donner une mesure exacte ni de l'immigration vers, ni de l'émigration hors de la ville, et par conséquent **il est impossible d'analyser la migration en tant qu'événement biographique, même si le statut migratoire** (défini au moment de l'entrée dans la population soumise au risque) **peut être introduit dans l'analyse comme variable explicative.**

Il faut en particulier être très vigilant pour **traiter les cas où l'émigration est motivée par un événement du même type que l'événement étudié.** Considérons par exemple l'émigration suite à une offre d'emploi à l'extérieur : comment traiter ce type d'événement ? Comme une entrée dans la vie active ou comme une émigration ? Pour savoir que répondre, il nous faut mettre en rapport la définition de la population soumise au risque et celle de l'événement. Par exemple, si l'offre concerne un emploi à l'extérieur, donc un marché de l'emploi plus vaste que celui de Dakar, l'émigration constitue bien une sortie d'observation. Mais si l'offre concernait un emploi à Dakar, et qu'immédiatement après l'employé a été affecté hors de Dakar, l'émigration n'est plus considérée comme une sortie d'observation puisqu'elle a eu lieu après l'événement (l'embauche). On conviendra cependant que ces cas sont très difficiles à distinguer les uns des autres : il faudrait disposer d'informations très précises sur le moment et les conditions de l'événement.

Les **cas d'émigrations suivies de retour** peuvent rester ambigus, même dans le cas où l'émigration n'a pas été motivée par un événement du même type que l'événement étudié. Prenons l'exemple d'un individu qui va faire ses études à l'extérieur de Dakar, pour revenir ensuite chercher du travail dans cette ville. Il a été soumis au risque d'entrer dans la vie active à l'extérieur de Dakar pendant la durée de son émigration, mais il est à nouveau soumis au risque à Dakar depuis son retour. Doit-on l'inclure dans le sous-échantillon de Dakarois de naissance ou bien dans le sous-échantillon de migrants venus chercher un emploi ?

C'est dire si **le problème de la migration ne peut être évacué dans la plupart des analyses de données biographiques : au contraire, il doit faire partie intégrante du raisonnement**. Le fait qu'on ne puisse mesurer la migration vers et hors de la population étudiée ne doit pas nous empêcher d'en souligner l'importance. Deux questions peuvent aider à mieux préciser les définitions quant aux sorties d'observation (les réponses sont relatives au premier emploi à Dakar).

Les troncatures sont-elles définies uniquement par la date d'enquête, ou bien aussi par la sortie physique de la population soumise au risque ?

- L'émigration représente une sortie possible de la population soumise au risque.

Les sorties d'observation sont-elles définitives ? (Y a-t-il des cas de sorties temporaires de la population soumise au risque ?)

- Certains individus sont revenus à Dakar sans avoir trouvé un emploi à l'extérieur. Ils peuvent donc à nouveau être soumis au risque de connaître leur premier emploi à Dakar.

3. L'analyse descriptive de l'événement

Les principes de la description d'un événement sont maintenant rassemblés : on a défini le groupe à risque, constitué des cohortes homogènes, défini les troncatures à droite et à gauche, choisi un intervalle de temps. Il s'agit maintenant de décrire l'événement dans le temps, à l'aide des tables de séjour et de leur représentation graphique.

Du point de vue de la relation causale, le seul élément explicatif dans cette phase de description est l'entrée de l'individu dans la population soumise au risque. Il faut être entré en observation (par naissance ou par immigration) pour connaître l'événement. L'entrée en observation n'est pas à proprement parler une cause au sens de variable explicative, mais elle peut être tout de même symbolisée sous la forme simple d'une cause unique,

$$D \longrightarrow E$$

où D est le début d'observation et E l'événement "premier emploi".

a) La présentation et la construction des fichiers de données

Tout ce dont on a besoin pour décrire un événement est la date de début d'observation de l'individu, ainsi que la date d'occurrence de l'événement ou la date de troncature, c'est-à-dire la date à laquelle l'individu sort de l'observation (soit par émigration, soit du simple fait de l'enquête). Du point de vue de la préparation des données, on a seulement besoin d'une durée d'observation de l'individu et d'un indice de troncature, qui par convention, sera égal à 1 si l'événement a eu lieu avant la date de sortie d'observation, et à 0 sinon (date d'enquête ou d'émigration dans le cas d'une troncature).

La création de la variable de durée avant l'événement

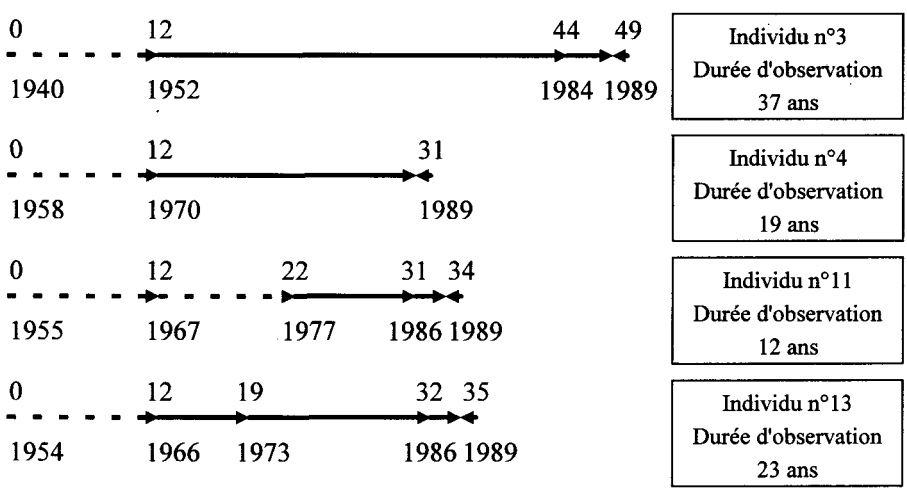
La durée doit être exprimée par une variable continue, dans la plus petite unité de temps nécessaire pour l'analyse. Il faudra donc convertir les dates en jour/mois/année dans une échelle commune, soit en jours, soit en mois, soit en années. Il est toujours préférable de constituer la plus petite échelle possible, car certaines options des commandes de STATA permettent de faire les calculs à une échelle supérieure sans pour autant modifier le fichier de données.

En sciences sociales, il est bien rare que l'on ne puisse pas situer un événement à l'année près. Le mois, et *a fortiori* le jour, sont beaucoup plus difficiles à obtenir, surtout dans une enquête rétrospective. Pourtant, le recueil sur le terrain de la datation au mois près a certains avantages, même si l'analyse se fait en définitive sur l'année : elle permet d'ordonner des événements de natures différentes (mariage, déménagement, emploi, naissance...) qui se seraient produits la même année. L'importance de cette remarque deviendra évidente lorsqu'on abordera l'interaction entre événements.

Pour l'instant, on va décrire un événement sur une échelle de temps annuelle : l'entrée dans la vie active des femmes à Dakar. Le fichier est constitué de plusieurs enregistrements par individu, chacun correspondant à une période d'étude, d'activité, de chômage ou d'inactivité. Il s'agit donc d'abord de sélectionner les périodes utiles, pour ensuite construire une variable qualifiant la durée. Les variables recueillies dans le questionnaire sont décrites comme suit pour les premiers individus du fichier (la variable "v306" donne l'année de naissance, la variable "v603" l'année de début de la période d'étude, d'activité, de chômage ou d'inactivité, et la variable "v638" l'année de fin de cette période) :

```
. use job

. sort no v601

. by no : list no v306 v603 v604 v611 v638, nolab

-> no=       3
          no  v306  v603  v604  v611  v638
  1.       3    40    52     4     0    84
  2.       3    40    84     1     3    89

-> no=       4
          no  v306  v603  v604  v611  v638
  3.       4    58    70     4     0    89

-> no=      11
          no  v306  v603  v604  v611  v638
  4.      11    55    77     4     0    86
  5.      11    55    86     1     3    89

-> no=      13
          no  v306  v603  v604  v611  v638
  6.      13    54    66     4     0    73
  7.      13    54    73     4     0    86
  8.      13    54    86     1     3    89

-> no=      15
(etc)
```

Ces observations correspondent aux "lignes de vie" suivantes :



La correspondance entre les enregistrements du fichier et la "ligne de vie" se fait comme suit pour l'individu n°13 :

```
-> no=      13
          no   v306   v603   v604   v611   v638

                              (avant 12 ans)      - - - - - ->
                                                  1954      1966

  6.      13    54     66     4      0      73          ------->
                                                    1966    1973

  7.      13    54     73     4      0      86      ----------->
                                                 1973       1986

  8.      13    54     86     1      3      89                -><-
                                                       1986 1989
```

Dans le fichier, chaque période d'emploi à Dakar correspond à une ligne d'observation. Sur les schémas ces périodes correspondent aux lignes pleines, les lignes en pointillé représentant les périodes hors de Dakar, ou bien la période avant l'âge de 12 ans, âge minimum à partir duquel les périodes d'emploi à Dakar sont prises en compte. En effet, le parcours professionnel n'est décrit, dans notre enquête, qu'à partir de l'âge de 12 ans. Comme on l'a dit précédemment, on ne peut prendre en compte les immigrés venus à Dakar après 12 ans parce qu'ils n'ont pas fait partie de la population soumise au risque depuis l'âge de 12 ans.

Toutes les commandes de STATA pour l'analyse de survie considèrent la durée comme le temps écoulé depuis le début d'observation, fixé par convention à l'instant 0. Cela signifie que si la durée est exprimée en années écoulées depuis la naissance, et qu'on souhaite ne décrire l'événement que depuis l'âge de 12 ans, il faudra alors faire les modifications nécessaires pour que l'origine soit fixée à 12 ans.

Pour l'instant, les variables "v604" (type d'activité) et "v611" (statut d'activité) suffiront à qualifier la période du point de vue de l'activité. La date de début de chaque période ("v603") est utile pour vérifier la cohérence des datations, mais c'est en fait la date de fin de période ("v638") qui servira à créer une variable de durée d'observation jusqu'à l'événement ou la troncature. La première étape consiste à calculer la variable de durée depuis l'âge de 12 ans. Il suffit pour cela d'exécuter la commande suivante :

```
. generate byte duree=v638-(v306+12)
```

Pour sélectionner le sous-échantillon des individus présents à Dakar depuis l'âge de 12 ans (appellés "Dakarois" pour simplifier), on crée une variable

indicatrice du statut migratoire. Construire une telle variable peut se faire aisément en utilisant une variable interne "_n" créée par STATA, comme on l'a vu dans le chapitre consacré au modèle logistique. On crée la variable statut migratoire comme suit :

```
. by no : gen byte immigree=v603[1]!=v306[1]+12
-> no=       3
-> no=       4
-> no=      11
-> no=      13
-> no=      15
(etc)
```

Pour éviter l'affichage à l'écran des numéros d'identifiant, on exécutera plutôt la commande précédée de "quietly" :

```
. quietly by no : gen byte immigree=v603[1]!=v306[1]+12
```

où "v603[1]" est la valeur prise par la variable "v603" pour _n==1, c'est-à-dire pour la première période d'observation de chaque individu. Le calcul se fait par individu (dont le numéro d'identifiant est contenu dans la variable "no"), parce que la commande est précédée de l'option « by *no* : ». Ainsi, pour l'individu n° 3

```
. no  v306  v603  v638
  3    40    52    84
  3    40    84    89
```

la variable "immigree" pour la première ligne d'observation prendra la valeur 0, car le début d'observation se situe en 1952 ("v603[1]==52") c'est-à-dire 12 ans après la naissance ("v306[1]+12==52"). La variable "immigree" pour la deuxième ligne d'observation prendra aussi la valeur 0, car pour chaque individu, le calcul se fait en fonction de la valeur de "v603" et de "v306" pour la première ligne d'observation (repérée par [1]).

La variable "immigree" prend la valeur 1 si la condition logique à droite du signe d'affectation est vérifiée ("v603[1]!=v306[1]+12", ce qui signifie que l'âge de début d'observation est différent de 12 ans), et la valeur 0 sinon. On obtiendrait le même résultat avec la commande :

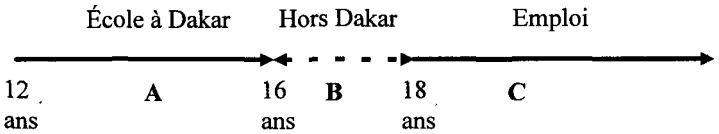
```
. quietly by no : gen byte immigree=cond(v603[1]!=v306[1]+12,1,0)
```

Pour les premiers individus du fichier le résultat de ce calcul est le suivant :

```
. by no : list no v603 v638 v611 duree immigree, nolab
-> no=       3
          no  v603  v638  v611     duree  immigree
  1.       3    52    84     0        32         0
  2.       3    84    89     3        37         0
-> no=       4
          no  v603  v638  v611     duree  immigree
  3.       4    70    89     0        19         0
-> no=      11
          no  v603  v638  v611     duree  immigree
  4.      11    77    86     0        19         1
  5.      11    86    89     3        22         1
-> no=      13
          no  v603  v638  v611     duree  immigree
  6.      13    66    73     0         7         0
  7.      13    73    86     0        20         0
  8.      13    86    89     3        23         0
(etc)
```

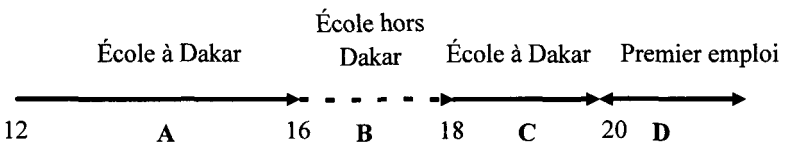
La création de l'indice de troncature

L'indice de troncature aura pour valeur "1" dans le cas où la fin de la période débouche sur un premier emploi et "0" sinon. Pour cela, il faut s'assurer que la période de soumission au risque n'a pas été interrompue par une émigration. En effet, si l'individu émigre hors de Dakar, il est soumis au risque dans une autre population au cours du temps qu'il passe hors de Dakar. Lorsqu'il revient (éventuellement) à Dakar, il est devenu un immigrant, qu'il faudra traiter comme les autres immigrants. Prenons le cas de l'itinéraire suivant :



Dans la période A, l'individu est soumis au risque d'obtenir d'un premier emploi dans la population des Dakarois. L'émigration à 16 ans constitue une troncature à droite (à traiter de la même façon que la date d'enquête). À 18 ans, l'individu rentre à nouveau dans la population soumise au risque, mais cette fois dans le sous-échantillon des immigrants.

Un autre cas peut se présenter :



Dans ce cas-là, il faut choisir si l'individu doit ou ne doit pas réintégrer la population des Dakarois, après une interruption pour la période B, dans la mesure où il n'a pas obtenu un premier emploi hors de Dakar. Il est possible de traiter un tel cas avec une approche modélisatrice. Ce traitement, assez complexe, ne sera pas évoqué dans ce manuel. Par contre, **l'analyse descriptive exclut le traitement des cas de sortie momentanée de la population soumise au risque**.

Pour que l'émigration soit considérée comme une troncature, on détermine si le séjour à Dakar a été interrompu par une période hors Dakar. On repère l'enregistrement qui suit immédiatement une période hors Dakar en comparant les dates de fin ("v638") et de début ("v603") de deux périodes consécutives pour un même individu. La commande « firstocc » (voir chapitre précédent) fait rapidement ce calcul :

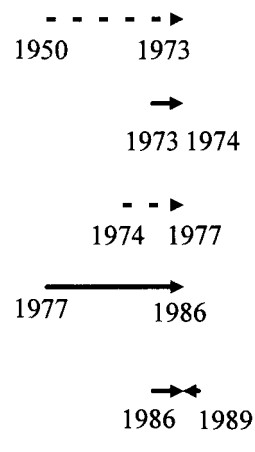
```
. firstocc no v601 =v603!=v638[_n-1] & _n>1, gen(emig)
```

Les périodes qui suivent l'émigration (périodes de retour à Dakar après avoir émigré hors de la ville) ne seront pas prises en compte pour la détermination de la population soumise au risque d'obtention d'un premier emploi à Dakar. On vérifie le résultat des calculs :

```
. by no : list no v603 v638 duree immigree emig, nolab noo nod
-> no=       3
      no  v603  v638  v611     duree  immigree      emig
       3    52    84     0        32         0         0
       3    84    89     3        37         0         0
-> no=       4
      no  v603  v638  v611     duree  immigree      emig
       4    70    89     0        19         0         0
-> no=      11
      no  v603  v638  v611     duree  immigree      emig
      11    77    86     0        19         1         0
      11    86    89     3        22         1         0
-> no=      13
      no  v603  v638  v611     duree  immigree      emig
      13    66    73     0         7         0         0
      13    73    86     0        20         0         0
      13    86    89     3        23         0         0
(etc)
-> no=      63
      no  v603  v638  v611     duree  immigree      emig
      63    67    71     0         4         0         0
      63    71    78     2        11         0         0
      63    78    89     0        22         0         0
-> no=      65
      no  v603  v638  v611     duree  immigree      emig
      65    73    74     0        12         1         0
      65    77    86     0        24         1         1
      65    86    89     3        27         1         0
-> no=      68
      no  v603  v638  v611     duree  immigree      emig
      68    71    89     0        24         1         0
-> no=      69
      no  v603  v638  v611     duree  immigree      emig
      69    63    66     0        16         1         0
      69    88    89     0        39         1         1
-> no=      73
(etc)
```

On remarque par exemple que l'individu n°65, arrivé en 1973 à 23 ans, a effectué une migration en 1974 et est revenu à Dakar en 1977. La période qui suit (de 1977 à 1986) est donc la période vécue à Dakar après la première émigration :

```
-> no=       65
no   v603   v638   v611   duree   immigree   emig

                (période avant l'arrivée à Dakar)

65     73     74      0      12          1      0

                (période d'émigration hors Dakar)

65     77     86      0      24          1      1

65     86     89      3      27          1      0
```



Il en est de même pour l'individu n°69 qui a émigré en 1966 pour retourner à Dakar en 1988, jusqu'à la date d'enquête (1989). Par contre, les individus n°63 et n°68 n'ont pas émigré.

On dispose maintenant de toutes les informations nécessaires pour délimiter une population homogène dans le temps : le statut migratoire et un indicateur d'émigration. On va donc créer un indice de troncature, que l'on va appeler "job1", égal à 1 si la période courante (avant le premier emploi) débouche sur une période d'emploi, c'est-à-dire si la période suivante est une période de premier emploi, et à 0 sinon :

```
. quietly by no  : gen nbjob=sum(v611==2 | v611==3)

. quietly by no  : gen nbemig=sum(emig)

. quietly by no  : gen byte job1=nbjob[_n+1]==1 & nbemig[_n+1]==0 if nbjob==0 &
nbemig==0 & immigree==0
```

Comme on le voit, la procédure est plutôt complexe, car pour calculer l'indice de troncature pour une observation donnée, il faut tenir compte de plusieurs valeurs de l'observation suivante. On a dû d'abord calculer le nombre cumulé de périodes passées hors Dakar ("nbemig"), ainsi que le nombre cumulé de périodes d'emplois (pour les personnes ayant déjà eu un premier emploi à Dakar : "nbjob") : ces nombres servent à la fois pour indiquer si la période suivante peut être prise en compte comme une période d'emploi (dans la mesure où la personne n'a encore effectué aucune émigration), et pour ne plus calculer d'indice de troncature si l'un

des deux événements (emploi ou émigration) a déjà eu lieu : on ne poursuit le calcul que si "nbemig==0 & nbjob==0". De plus, on a fait le calcul seulement pour les Dakaroises ("immigree==0").

Un problème se pose aussi dans le cas où la personne a débuté un premier emploi dès l'âge de 12 ans : dans ce cas, l'indice de troncature doit faire référence au début de la période (12 ans) et non pas à la fin de la période d'emploi :

```
. quietly by no  : replace job1=1 if nbjob==1 & _n==1 & immigree==0
```

De la même façon, on calculera la durée avant le premier emploi pour tenir compte de ce cas particulier :

```
. quietly by no  : gen Tjob1=duree if job1!=. & immigree==0

. quietly by no  : replace Tjob1=0 if nbjob==1 & _n==1 & immigree==0

. drop nbjob nbemig
```

Cette série bien longue de sept commandes peut heureusement être exécutée à l'aide d'une seule commande « censor », dont le programme est fourni avec ce manuel.

On exécutera donc la commande suivante :

```
. censor no duree v601 =v611==2 | v611==3 if immigree==0, gen(job1)
before(emig==1)
```

Dans cette ligne de commande, la première variable "no" est le numéro identifiant l'individu, et la seconde la durée jusqu'à l'événement, c'est-à-dire la durée en fin de période ("duree"). Remarquez l'usage d'une troisième variable ("v601") pour indiquer l'ordre des périodes : cette variable est en principe redondante avec la variable de durée (qui suffit à classer les périodes dans un ordre croissant). Cependant, il peut arriver que deux événements se produisent la même année : dans ce cas l'année n'est pas une unité de temps assez précise pour connaître l'ordre des événements.

La commande « censor » a créé directement les variables de troncature "job1" et de durée "Tjob1", sans calcul des variables intermédiaires "nbjob" et "nbemig". On remarque l'usage de l'option « before(*expression*) » pour ne considérer que les périodes ayant eu lieu avant un événement indépendant (ici l'émigration) : dans ce cas l'indice de troncature sera fixé à zéro. Pour toutes les périodes qui suivent l'événement, l'indice de troncature et la durée avant l'événement sont codés manquants (*missing value* : ".") :

```
. by no : list v611 duree immigree emig job1 Tjob1, noo nod
-> no=       3
v611     duree  immigree      emig      job1     Tjob1
   0        32         0         0         1        32
   3        37         0         0         .         .
-> no=       4
v611     duree  immigree      emig      job1     Tjob1
   0        19         0         0         0        19
-> no=      11
v611     duree  immigree      emig      job1     Tjob1
   0        19         1         0         .         .
   3        22         1         0         .         .
-> no=      13
v611     duree  immigree      emig      job1     Tjob1
   0         7         0         0         0         7
   0        20         0         0         1        20
   3        23         0         0         .         .
-> no=      15
(etc)
-> no=      63
v611     duree  immigree      emig      job1     Tjob1
   0         4         0         0         1         4
   2        11         0         0         .         .
   0        22         0         0         .         .
-> no=      65
v611     duree  immigree      emig      job1     Tjob1
   0        12         1         0         .         .
   0        24         1         1         .         .
   3        27         1         0         .         .
-> no=      68
v611     duree  immigree      emig      job1     Tjob1
   0        24         1         0         .         .
-> no=      69
v611     duree  immigree      emig      job1     Tjob1
   0        16         1         0         .         .
   0        39         1         1         .         .
(etc)
-> no=     113
v611     duree  immigree      emig      job1     Tjob1
   0         9         0         0         0         9
   0        17         0         1         .         .
-> no=     114
v611     duree  immigree      emig      job1     Tjob1
   3        36         1         0         .         .
-> no=     116
v611     duree  immigree      emig      job1     Tjob1
   0        19         0         0         1        19
   3        43         0         0         .         .
-> no=     121
(etc)
```

Le travail de codification de la durée avant l'événement et de la troncature peut paraître complexe, mais il suffit à tous les traitements pour la suite de ce manuel. Pour résumer, seulement deux commandes sont nécessaires, celle qui crée la variable de durée avant l'événement :

```
. generate byte duree=v638-(v306+12)
```

et celle qui crée la variable de troncature :

```
. censor no duree v601 =v611==2 | v611==3 if v603[1]!=v306[1]+12, gen(job1)
before(v603!=v638[_n-1] & _n>1)
```

Comme on le voit, la commande « censor » permet de calculer l'indice de troncature pour une sélection d'individus non migrants (« if *expression* ») et pour des périodes avant l'émigration éventuelle (« before(*expression*) »), sans même calculer les variables "immigree" et "emig". Les lecteurs qui ont l'expérience d'autres logiciels statistiques se rendront compte que la ligne de commande « censor » est particulièrement concise. Cependant, aussi courte l'étape de codification soit-elle, la réflexion sur la constitution d'une population soumise à un risque homogène ne doit pas être négligée : l'étape de conceptualisation est un préalable nécessaire quels que soient la composition de l'échantillon et le support informatique des données.

b) Les procédures « kapmeier », « ltable » et « survsum »

On est donc maintenant en mesure d'analyser l'entrée en activité. Rappelons que "Tjob1" et "job1" sont codées manquantes pour les immigrées. Pour cette raison, les observations correspondantes à ces données manquantes ne seront pas prises en compte dans les calculs : on ne se soucie plus d'ajouter les conditions ("immigree==0 & nbemig==0") dans les commandes qui vont suivre. Pour l'analyse descriptive, on dispose maintenant d'une durée, d'un indice de troncature et éventuellement d'une série de variables indicatrices du groupe auquel appartient l'individu (ex : la variable "strate", définie selon l'appartenance à un groupe de générations).

Un des outils les plus efficaces de l'analyse exploratoire des histoires de vie est certainement l'**estimateur de la fonction de séjour**, dit de Kaplan-Meier[9], qui permet de tenir compte des données tronquées à droite. On calcule une probabilité de connaître l'événement dans chaque intervalle de temps, et on obtient ainsi une courbe qui s'interprète simplement comme la proportion de "survivants" pour chaque durée de séjour dans un état donné.

Cette proportion a une signification probabiliste. **La courbe décrit le comportement hypothétique d'une cohorte qui aurait connu les mêmes conditions de vie pour que l'événement étudié, éventuellement, se réalise.** Pour que cette courbe corresponde effectivement au comportement d'une cohorte, il faudrait pouvoir suivre les individus jusqu'à leur décès (lorqu'ils sortent effectivement de la population soumise au risque) et que cette cohorte soit homogène selon toutes les caractéristiques dont pourrait dépendre la réalisation de l'événement. Cela veut dire que **le calcul suppose que la seule hétérogénéité est introduite par l'âge auquel chaque individu connaît l'événement.**

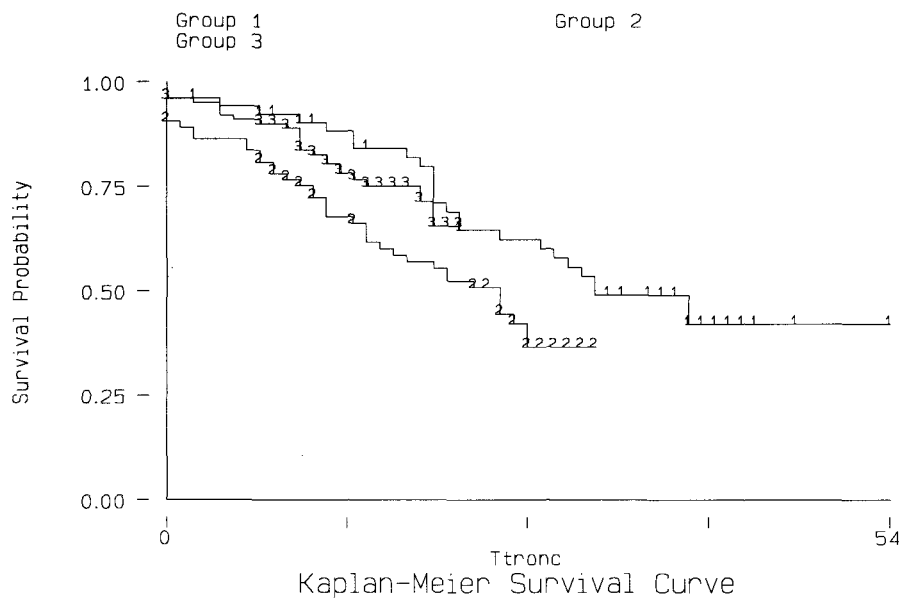
---

[9] Pour une explication en français, voir Courgeau et Lelièvre, 1989, chapitres III et IV.

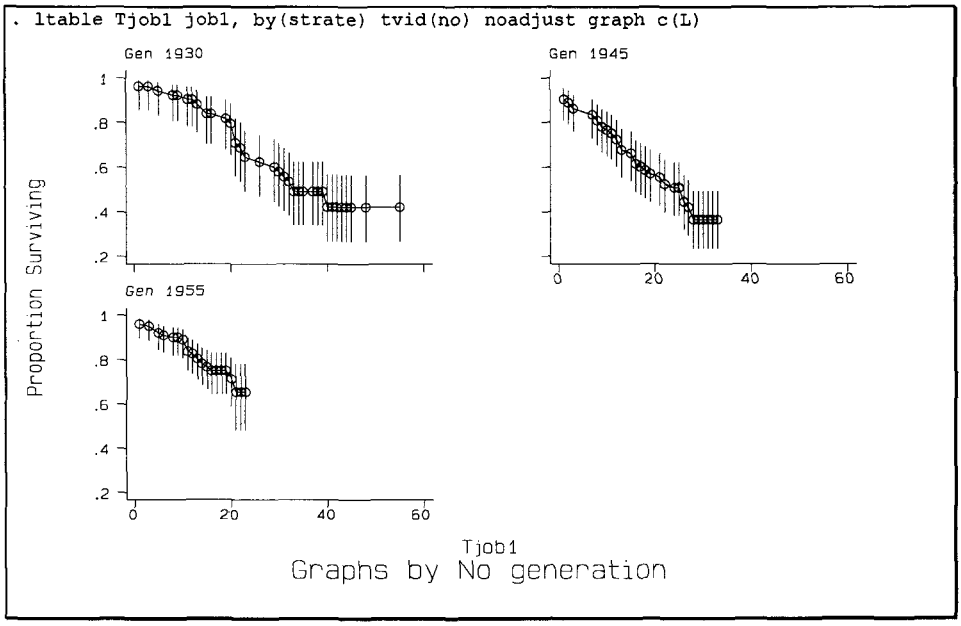
Les conditions d'analyse sont bien évidemment différentes et on fera intervenir d'autres variables par la suite. L'homogénéité des cohortes n'est jamais vérifiée en sciences sociales, en supposant même qu'on puisse suivre les personnes enquêtées jusqu'à leur décès. **Le but de l'analyse est certes de faire apparaître des groupes les plus homogènes possibles, mais il est surtout de mettre en relief les variables les plus discriminantes, celles qui expliquent la plus grande part de l'hétérogénéité entre les groupes.**

Pour représenter les courbes de Kaplan-Meier avec la commande « kapmeier », on doit n'utiliser qu'une observation par individu. Ceci oblige à créer une variable indicatrice de l'observation à prendre en compte, à savoir celle correspondant à la période pendant laquelle la troncature ou l'événement a eu lieu (repérée par la variable "dernier") :

```
. firstocc no v601 =jobl[_n+1]==. & jobl!=., gen(dernier)
. kapmeier Tjobl jobl if dernier==1, by(strate)
```



On peut aussi utiliser la commande « ltable » avec l'option « tvid(*ident*) » qui permet de relier les enregistrements entre eux pour un même individu identifié par la variable *ident*.



Avec la commande « ltable », il n'est pas nécessaire de créer la variable de contrôle "dernier". On remarque aussi que cette commande permet de représenter les courbes sur un graphique différent pour chaque groupe de générations, tandis que la commande « kapmeier » fait figurer les résultats pour chaque groupe sur un même graphique.

Dans cet exemple, la population soumise au risque en début d'observation est constituée des Dakaroises (femmes présentes à Dakar à l'âge de 12 ans), soit un échantillon de 228 femmes parmi les 497 enquêtées de l'enquête biographique IFAN-ORSTOM. Dans cet échantillon les troncatures à droite sont aussi bien causées par la date d'enquête (119 observations soit 52,2 % de l'échantillon) que par une émigration hors de Dakar (18 observations soit 7,9 % de l'échantillon).

Les courbes de Kaplan-Meier représentent à chaque âge, la proportion de femmes qui n'ont encore jamais eu d'emploi selon les trois groupes de générations qui ont servi pour la stratification de l'échantillon de l'enquête biographique IFAN-ORSTOM. En comparant les groupes de générations 1930-44 et 1945-54, on voit très nettement apparaître une accélération du calendrier d'insertion dans la vie professionnelle, qui se traduit dans les graphiques par un décalage vers la gauche de la fonction de séjour de la seconde génération par rapport à celle de la première. Les femmes du premier groupe de générations (1930-44) ont eu un premier emploi plus tardif.

L'équivalent chiffré de ces courbes est obtenu par la commande « ltable », sans l'option « graph » :

```
. ltable Tjob1 job1, by(strate) tvid(no) noadjust c(L)
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
strate 1								
0	1	53	2	0	0.9623	0.0262	0.8574	0.9904
2	3	51	0	1	0.9623	0.0262	0.8574	0.9904
4	5	50	1	0	0.9430	0.0320	0.8336	0.9813
7	8	49	1	1	0.9238	0.0366	0.8095	0.9707
8	9	47	0	1	0.9238	0.0366	0.8095	0.9707
10	11	46	1	1	0.9037	0.0410	0.7838	0.9588
11	12	44	0	1	0.9037	0.0410	0.7838	0.9588
12	13	43	1	0	0.8827	0.0451	0.7570	0.9456
14	15	42	2	0	0.8406	0.0518	0.7060	0.9171
15	16	40	0	1	0.8406	0.0518	0.7060	0.9171
18	19	39	1	0	0.8191	0.0548	0.6805	0.9017
19	20	38	1	0	0.7975	0.0574	0.6556	0.8858
20	21	37	4	0	0.7113	0.0654	0.5608	0.8182
21	22	33	1	0	0.6898	0.0669	0.5381	0.8004
22	23	32	2	1	0.6466	0.0693	0.4937	0.7640
25	26	29	1	0	0.6244	0.0704	0.4709	0.7448
28	29	28	1	0	0.6021	0.0714	0.4484	0.7254
29	30	27	1	0	0.5798	0.0721	0.4263	0.7057
30	31	26	1	0	0.5575	0.0727	0.4045	0.6857
31	32	25	1	0	0.5352	0.0731	0.3831	0.6654
32	33	24	2	0	0.4906	0.0735	0.3410	0.6241
33	34	22	0	2	0.4906	0.0735	0.3410	0.6241
34	35	20	0	1	0.4906	0.0735	0.3410	0.6241
36	37	19	0	2	0.4906	0.0735	0.3410	0.6241
37	38	17	0	1	0.4906	0.0735	0.3410	0.6241
38	39	16	0	2	0.4906	0.0735	0.3410	0.6241
39	40	14	2	1	0.4205	0.0780	0.2677	0.5658
40	41	11	0	2	0.4205	0.0780	0.2677	0.5658
41	42	9	0	1	0.4205	0.0780	0.2677	0.5658
42	43	8	0	1	0.4205	0.0780	0.2677	0.5658
43	44	7	0	2	0.4205	0.0780	0.2677	0.5658
44	45	5	0	2	0.4205	0.0780	0.2677	0.5658
47	48	3	0	2	0.4205	0.0780	0.2677	0.5658
54	55	1	0	1	0.4205	0.0780	0.2677	0.5658
strate 2								
0	1	74	7	1	0.9054	0.0340	0.8118	0.9537
1	2	66	1	0	0.8917	0.0362	0.7951	0.9443
2	3	65	2	0	0.8643	0.0399	0.7624	0.9246
6	7	63	2	0	0.8368	0.0431	0.7305	0.9039
7	8	61	2	1	0.8094	0.0459	0.6994	0.8824
8	9	58	2	1	0.7815	0.0483	0.6682	0.8600
9	10	55	1	1	0.7673	0.0495	0.6525	0.8484
10	11	53	1	1	0.7528	0.0506	0.6365	0.8365
11	12	51	2	1	0.7233	0.0528	0.6042	0.8119
12	13	48	3	0	0.6781	0.0556	0.5558	0.7734
14	15	45	1	1	0.6630	0.0563	0.5399	0.7603
15	16	43	3	0	0.6167	0.0584	0.4919	0.7195
16	17	40	1	0	0.6013	0.0589	0.4762	0.7056
17	18	39	1	0	0.5859	0.0594	0.4606	0.6917
18	19	38	1	0	0.5705	0.0598	0.4451	0.6776
20	21	37	1	0	0.5551	0.0601	0.4298	0.6634
21	22	36	2	0	0.5242	0.0606	0.3996	0.6347
23	24	34	1	3	0.5088	0.0608	0.3846	0.6201
24	25	30	0	6	0.5088	0.0608	0.3846	0.6201
25	26	24	3	2	0.4452	0.0633	0.3192	0.5636
26	27	19	1	3	0.4218	0.0642	0.2954	0.5427
27	28	15	2	3	0.3655	0.0668	0.2377	0.4941
28	29	10	0	2	0.3655	0.0668	0.2377	0.4941

29	30	8	0	2	0.3655	0.0668	0.2377	0.4941
30	31	6	0	2	0.3655	0.0668	0.2377	0.4941
31	32	4	0	2	0.3655	0.0668	0.2377	0.4941
32	33	2	0	2	0.3655	0.0668	0.2377	0.4941
strate 3								
0	1	101	4	1	0.9604	0.0194	0.8979	0.9849
2	3	96	1	0	0.9504	0.0216	0.8849	0.9790
4	5	95	3	0	0.9204	0.0270	0.8471	0.9594
5	6	92	1	0	0.9104	0.0285	0.8348	0.9523
7	8	91	1	1	0.9004	0.0299	0.8227	0.9451
8	9	89	0	2	0.9004	0.0299	0.8227	0.9451
9	10	87	1	1	0.8900	0.0313	0.8102	0.9375
10	11	85	5	2	0.8377	0.0372	0.7486	0.8973
11	12	78	1	2	0.8269	0.0382	0.7363	0.8887
12	13	75	2	1	0.8049	0.0403	0.7110	0.8710
13	14	72	2	18	0.7825	0.0421	0.6858	0.8526
14	15	52	1	4	0.7675	0.0439	0.6674	0.8410
15	16	47	1	5	0.7511	0.0459	0.6472	0.8284
16	17	41	0	4	0.7511	0.0459	0.6472	0.8284
17	18	37	0	10	0.7511	0.0459	0.6472	0.8284
18	19	27	0	6	0.7511	0.0459	0.6472	0.8284
19	20	21	1	8	0.7154	0.0560	0.5889	0.8090
20	21	12	1	4	0.6558	0.0767	0.4835	0.7827
21	22	7	0	5	0.6558	0.0767	0.4835	0.7827
22	23	2	0	2	0.6558	0.0767	0.4835	0.7827

Les courbes de Kaplan-Meier représentent la distribution de la durée avant la réalisation d'un événement. Puisqu'elles ont une signification probabiliste, on peut y associer un **intervalle de confiance** qui tiendra compte des effectifs soumis au risque à chaque durée. Les traits verticaux le long de chaque courbe du graphique produit par « ltable » représentent l'intervalle de confiance à 95 %, et cet intervalle est donné sous forme chiffrée dans les deux dernières colonnes du tableau précédent.

Habituellement, pour résumer l'allure de la distribution, on calculera un **indice de valeur centrale, la médiane** (ou deuxième quartile), c'est-à-dire la durée de séjour pour laquelle 50 % de la cohorte est encore "survivante". Parfois, on calculera le premier quartile durée de séjour pour lequelle 75 % de la cohorte est encore survivante. Le troisième quartile est estimé avec moins de fiabilité lorsque les données sont fortement tronquées aux durées élevées. Lorsque l'événement est rare, le deuxième et le troisième quartiles (et parfois même le premier quartile) ne sont pas forcément atteints et ne peuvent donc être calculés.

Le tableau produit par la commande « ltable » montre que le premier quartile (l'âge auquel 25 % des femmes ont déjà obtenu un premier emploi) n'est atteint qu'à la durée 20 et la médiane à la durée 32 dans le premier groupe de générations (1930-44). Étant donné qu'il faut rajouter 12 ans à ces durées pour calculer les âges exacts, le premier quartile est en fait atteint à 32 ans et la médiane à 44 ans. Dans le groupe d'année de naisssance 1945-54, le premier quartile est de 22 ans et la médiane est de 37 ans. Les durées médianes sont aussi obtenues avec la commande « survsum » :

```
. survsum Tjob1 job1 if dernier==1, by(strate)
    Group          Obs.          Died        Median          Mean
-------------------------------------------------------------------
        1            53            26            32      54.30769
        2            74            40            25        31.375
        3           101            25             .         55.52
-------------------------------------------------------------------
    Total           228            91             .             .
```

On remarque immédiatement que la médiane n'est pas calculée pour le troisième groupe de générations. Au moment de l'enquête, ce groupe pouvait avoir n'importe quel âge entre 25 et 35 ans, ce qui veut dire qu'il n'avait pas encore atteint la durée de 23 ans écoulée depuis l'âge de 12 ans (35-12=23). À la durée de 22 ans (correspondant à des femmes âgées de 34 ans), d'après les calculs de « ltable », 65,58 % des femmes n'avaient encore jamais travaillé à Dakar. La médiane n'était donc pas encore atteinte et elle figure comme manquante dans le tableau produit par la commande « survsum ».

En fait, le calendrier d'entrée dans la vie active s'est sensiblement modifié entre les générations 1945-54 et 1955-64. Les entrées précoces sont devenues plus rares de sorte qu'entre les deux groupes de générations, les quartiles sont décalés de quelques années. Dans le groupe de génération 1955-64, le premier quartile est atteint à 31 ans et même si les conditions changent au cours des prochaines années, la médiane sera certainement supérieure à 40 ans pour ce groupe de générations.

De toute façon, l'importance des troncatures pour le groupe de générations 1955-64 (75,2 % des observations) impose le conditionnel. En fait, la fonction de séjour au-delà de 25 ans pour ce groupe doit être interprétée comme une estimation de ce qui se passerait dans les années à venir si les conditions des années précédentes se maintenaient.

c) Les procédures « logrank » et « mantel »

Un certain nombre de tests non paramétriques sont disponibles pour évaluer la significativité de la différence entre les trois courbes.

```
. logrank Tjob1 job1 if dernier==1, by(strate)
    Group        Events      Predicted
------------------------------------
        1            26          33.26
        2            40          29.41
        3            25          28.33
------------------------------------
                chi2(2) =       5.79
                Pr>chi2 =     0.0552
```

La commande « logrank » mesure la différence globale entre les trois courbes, qui est ici significative au seuil de 10 % ($Pr>\chi 2=0{,}0552$). Pour mesurer la différence entre les courbes prises deux à deux, on utilisera le test de Mantel :

```
. * comparaison des generations 30-44 et 45-54
. mantel Tjob1 job1 if dernier==1 & strate!=3, by(strate)

    Group         Events      Predicted
-------------------------------------
        1             26          34.71
                      z =          2.14
                 Pr>|z| =        0.0322

. * comparaison des generations 45-54 et 55-64
. mantel Tjob1 job1 if dernier==1 & strate!=1, by(strate)

    Group         Events      Predicted
-------------------------------------
        2             40          32.92
                      z =          1.80
                 Pr>|z| =        0.0716

. * comparaison des generations 30-44 et 55-64
. mantel Tjob1 job1 if dernier==1 & strate!=2, by(strate)

    Group         Events      Predicted
-------------------------------------
        1             26          28.64
                      z =          0.72
                 Pr>|z| =        0.4693
```

On remarque en particulier que la différence entre le premier et le dernier groupe de générations n'est pas significative au seuil de 10 % ($Pr>|z| = 0{,}4693$) : le calendrier d'entrée en activité des générations 1955-64 a rejoint celui des générations 1930-44. En simplifiant, on pourrait presque dire que le comportement des filles a rejoint celui de leur mère.

4. La description des événements concurrents

Qu'arrive-t-il lorsque l'événement n'est pas défini de manière univoque ? Par exemple, les femmes peuvent acquérir un emploi salarié ou un emploi indépendant. Comme pour le modèle « mlogit », on considère alors que la variable dépendante est polytomique, c'est-à-dire qu'elle distingue plusieurs modalités. Mais cette fois-ci, l'analyse des différents types de risques tiendra compte du temps. Pour cela, il nous faut utiliser un autre estimateur que celui de Kaplan-Meier, conçu pour un risque unique.

a) La notion de risques concurrents

Jusqu'alors, l'événement considéré a été du même type pour tous les individus qui le connaissent. Or, certains événements, même s'ils marquent la sortie d'un même état (par exemple celui de célibataire, d'hébergé, etc.), peuvent mener à des états forts différents (marié monogame ou polygame, locataire ou propriétaire, etc.). La sortie d'un état importe autant que l'orientation vers d'autres états. Par exemple, ce n'est pas seulement le moment de l'entrée dans la vie active en tant que tel qui est intéressant, mais aussi la nature du premier emploi que l'enquêtée obtient.

Dans le jargon démographique, on parle alors de risques multiples ou concurrents (*competing risks*). Cela signifie essentiellement que l'événement étudié est scindé en plusieurs modalités exclusives l'une de l'autre : une fois qu'on a accédé à un premier emploi salarié, on ne peut accéder à un premier emploi indépendant, et le statut d'indépendant ne peut être connu que lors de la période d'emploi suivante. On a déjà vu que cette conception de la variable dépendante peut être traitée dans un modèle multinomial logit, mais le temps n'intervenait pas dans ce modèle.

Plusieurs auteurs ont souligné l'intérêt de scinder un événement en plusieurs catégories dans l'analyse longitudinale. En fait, des calculs menés pour tous les risques pris globalement peuvent conduire à des estimations biaisées, étant donné que différents types de passages d'un statut à un autre peuvent obéir à différents mécanismes. Ainsi, la distribution du temps d'occurrence d'un événement peut varier fortement selon le type d'événement. En outre, la difficulté à trouver des catégories adéquates pour l'analyse est moins technique que sémantique : il faut définir à l'avance les catégories qui ont un sens pour qualifier l'événement étudié.

b) L'estimateur de Aalen et sa représentation graphique : « graalen »

Dans le cas de risques concurrents, l'estimateur de Kaplan-Meier nécessite de faire l'hypothèse, rarement vérifiée, d'indépendance entre chacun des risques. Les tables à extinctions multiples font aussi cette hypothèse très forte, et on ne saurait trop conseiller au lecteur d'éviter ce genre de techniques.

L'estimateur de Aalen[10] est plus indiqué car il ne pose aucune restriction sur l'interdépendance entre les différents types de sortie d'observation. En fait, il s'agit d'une généralisation du cadre théorique de l'estimateur de Kaplan-Meier, sans l'hypothèse d'indépendance des risques. Par exemple, on peut mourir de différentes maladies ou accidents, mais considérer chaque cause de décès comme indépendante l'une de l'autre est difficile : pour mourir d'une cause donnée (un

[10] Pour une explication en français, voir Courgeau et Lelièvre, 1989, chapitre III et IV.

cancer) il faut avoir survécu à d'autres risques (un accident, une maladie infectieuse, etc.), et à l'inverse, certains risques sont associés entre eux (maladies cardio-vasculaires et respiratoires, suicide et anémie), souvent parce qu'ils ont des causes communes (tabagisme, toxicomanie). Ces processus de sélection et d'association ne sont généralement pas aléatoires, et il en est de même pour d'autres événements plus réjouissants de la vie sociale tels que l'accès au premier emploi, la migration, le début d'une vie de couple, etc.

Concrétement, il s'agit de calculer l'**intensité cumulée** (la somme cumulée des quotients instantanés) pour chaque type d'événement (une cause de l'événement), et d'en tracer la courbe. La courbe de Kaplan-Meier, si elle est estimée pour une cause donnée de l'événement, considére les autres causes comme des pertes (des troncatures), c'est-à-dire qu'elle considère les autres risques comme absents. L'estimateur de Aalen est, au contraire, multivarié et mesure l'intensité relative de chaque type d'événement. De ce fait, il n'est pas soumis à l'hypothèse d'indépendance entre chacun des types d'événement.

La commande « graalen » fournie avec ce manuel calcule les estimateurs des quotients instantanés pour chaque durée, et en fait la somme cumulée. Le niveau de chacune des courbes par type d'événement ainsi produites n'est pas interprétable en lui-même, mais on pourra **comparer les pentes des courbes entre elles : au moment *t*, la pente est une estimation de l'intensité de chaque type d'événement**. On pourra ainsi comparer les intensités de mortalité par cancer et par infection pour un âge ou un groupe d'âges donné.

L'interprétation est moins compliquée qu'il n'y paraît. Reprenons l'exemple de l'entrée dans la vie active. Il faudra d'abord calculer un indice de troncature correspondant à chacun des risques :

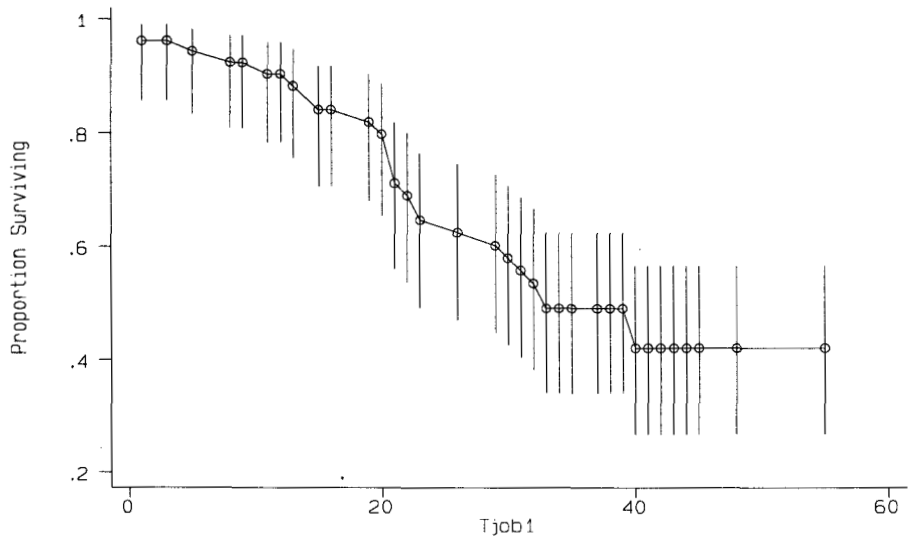
```
. censor no duree v601=v611==2 if immigree==0, gen(sal) before(emig==1 | v611==3)

. censor no duree v601=v611==3 if immigree==0, gen(ind) before(emig==1 | v611==2)

. censor no duree v601=emig if immigree==0, gen(emig1) before(v611==2 | v611==3)
```

Ainsi, la nouvelle variable "ind" (pour "indépendante"), est codée "1" si la femme devient indépendante ("=v611==3") avant d'émigrer ou de devenir salariée ("before(emig==1 | v611==2"). Si la femme devient salariée, si elle émigre ou si elle n'a pas eu encore d'emploi au moment de l'enquête, la variable "ind" est codée "0". On peut vérifier le résultat obtenu par ces commandes :

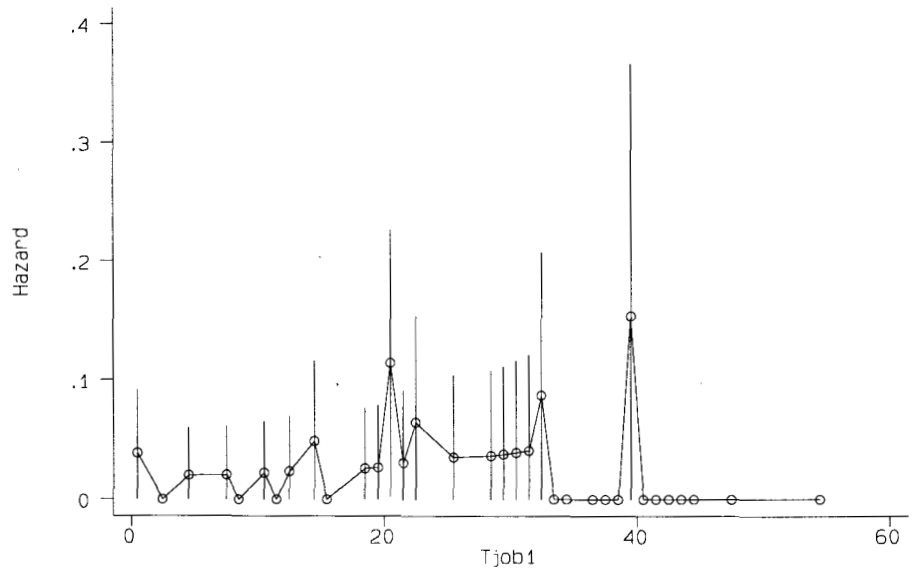
```
. by no : list v611 duree immigree emig job1 sal ind emig1
-> no=       3
v611     duree  immigree      emig      job1       sal       ind      emig1
   0        32         0         0         1         0         1          0
   3        37         0         0         .         .         .          .
-> no=       4
v611     duree  immigree      emig      job1       sal       ind      emig1
   0        19         0         0         0         0         0          0
-> no=      11
v611     duree  immigree      emig      job1       sal       ind      emig1
   0        19         1         0         .         .         .          .
   3        22         1         0         .         .         .          .
-> no=      13
v611     duree  immigree      emig      job1       sal       ind      emig1
   0         7         0         0         0         0         0          0
   0        20         0         0         1         0         1          0
   3        23         0         0         .         .         .          .
(etc)
-> no=      63
v611     duree  immigree      emig      job1       sal       ind      emig1
   0         4         0         0         1         1         0          0
   2        11         0         0         .         .         .          .
   0        22         0         0         .         .         .          .
-> no=      65
v611     duree  immigree      emig      job1       sal       ind      emig1
   0        12         1         0         .         .         .          .
   0        24         1         1         .         .         .          .
   3        27         1         0         .         .         .          .
-> no=      68
v611     duree  immigree      emig      job1       sal       ind      emig1
   0        24         1         0         .         .         .          .
-> no=      69
v611     duree  immigree      emig      job1       sal       ind      emig1
   0        16         1         0         .         .         .          .
   0        39         1         1         .         .         .          .
(etc)
-> no=     113
v611     duree  immigree      emig      job1       sal       ind      emig1
   0         9         0         0         0         0         0          1
   0        17         0         1         .         .         .          .
-> no=     114
v611     duree  immigree      emig      job1       sal       ind      emig1
   3        36         1         0         .         .         .          .
(etc)
```

La commande « graalen » trace ensuite la courbe du cumul des quotients pour chacun des risques. Pour mieux comprendre le rapport entre les courbes de Kaplan-Meier et de Aalen, on peut comparer les quatres graphiques suivants, qui représentent l'entrée dans la vie active des Dakaroises des générations 1930-44 :

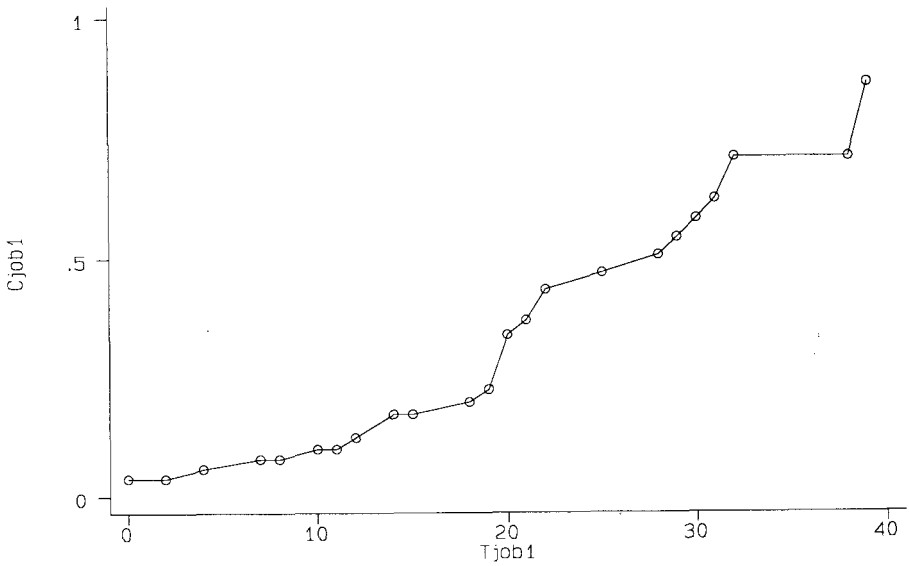
```
. ltable Tjob1 job1 if strate==1, tvid(no) graph notab
```



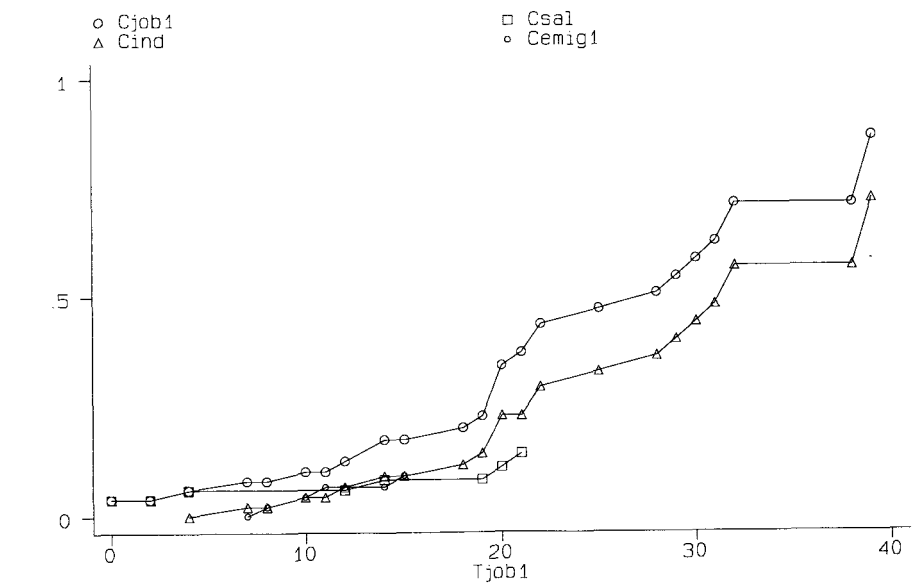

```
. ltable Tjob1 job1 if strate==1, tvid(no) hazard graph notab
```



```
. graalen Tjob1 job1 if strate==1, tvid(no)
```



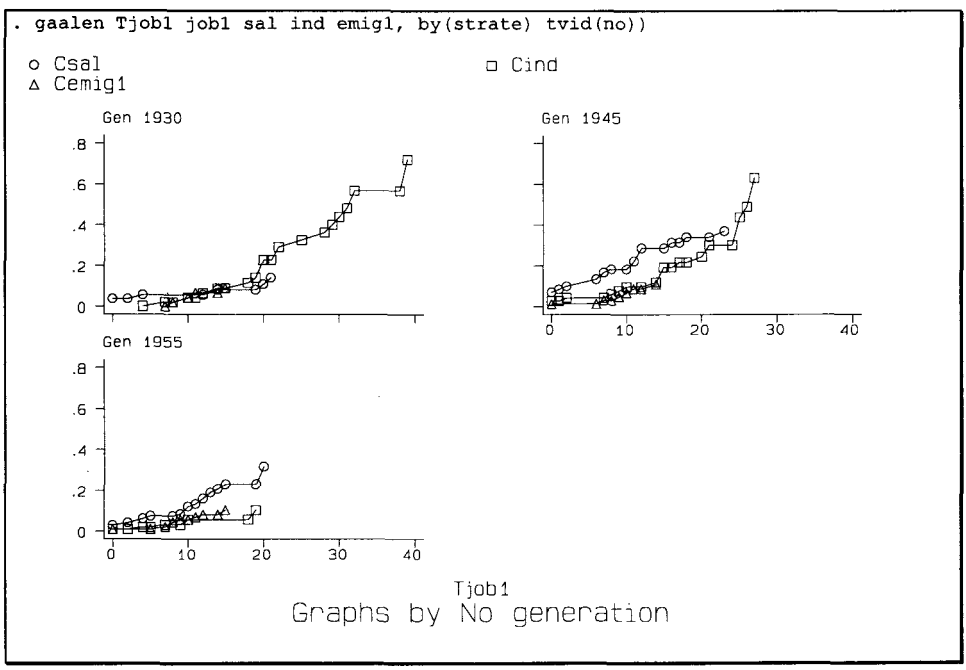

```
. graalen Tjob1 job1 sal ind emig1 if strate==1, tvid(no)
```



Dans le premier graphique figure la courbe de Kaplan-Meier pour le groupe de générations 1930-44, et dans le deuxième la série des quotients qui ont servi au tracé de cette courbe. La troisième figure exprime la même chose, mais sous la forme des quotients cumulés, plutôt que sous la forme d'une courbe de séjour (en l'état de personne n'ayant encore jamais travaillé).

Pour tracer la quatrième figure, la commande « graalen » a créé des variables "Cjob1", "Csal", "Cind" et "Cemig1", pour représenter les quotients cumulés du risque global d'avoir un emploi "job1", et les quotients des risques de devenir salariée "sal", indépendante "ind" ou d'émigrer "emig1". Les variables "C..." ne sont pas conservées à l'issue de l'exécution de la commande. On remarque sans surprise que la courbe pour "Cjob1" est identique à celle du troisième graphique.

Les mêmes calculs peuvent être faits selon une variable de contrôle, par exemple selon le groupe de générations (variable "strate") :



En interprétant ces courbes, il faut se rappeler que les niveaux ne sont pas comparables entre chacun des graphiques, mais seulement à l'intérieur d'un même graphique, entre les courbes tracées pour les événements concurrents. On ne peut donc comparer les courbes "Csal" entre deux générations, mais on peut comparer les courbes "Csal" avec les courbes "Cind" et "Cemig1" pour une même génération.

Pour décrire la distribution des événements par groupe de générations, on peut créer une variable indiquant le type de troncature :

```
. gen byte event=sal + 2*ind + 3*emig1

. lab def event 0 "enquete" 1 "salarie" 2 "indep" 3 "emig"

. lab val event event

. lab var event "Type de troncature"

. tab strate event, row
```

No generation	Type de troncature enquete	salarie	indep	emig	Total
Gen 1930	23	6	20	4	53
	43.40	11.32	37.74	7.55	100.00
Gen 1945	28	20	20	6	74
	37.84	27.03	27.03	8.11	100.00
Gen 1955	68	19	6	8	101
	67.33	18.81	5.94	7.92	100.00
Total	119	45	46	18	228
	52.19	19.74	20.18	7.89	100.00

Ce tableau ne permet pas de situer les événements dans le temps, c'est-à-dire de déterminer le calendrier des événements. Cependant, une lecture croisée de ce tableau et des courbes de Aalen par groupes de générations est instructive. Ainsi, on peut voir que la contribution du salariat dépasse celle de l'emploi indépendant à partir des générations 1945-54. Mais ce phénomène n'a pu être observé qu'avant la durée 35 (47 ans) : on ne sait si le salariat va continuer sa progression dans les jeunes générations, ou au contraire céder la place à l'emploi indépendant comme dans les premières générations où il a constitué la seule voie d'accès à la vie active après la durée 32 (44 ans). On voit aussi que l'émigration n'est pas négligeable dans les dernières générations 1955-64 : elle a autant d'importance que l'accès à l'emploi indépendant.

5. Conclusions sur les tables de séjour

La technique des tables de séjour et les diverses techniques associées (tests non paramétriques, courbes de Aalen) apportent une description très riche de l'événement. La dimension temporelle y est pleinement intégrée : on se rapproche du principe de priorité temporelle à la base de la relation de causalité. C'est donc un pas qualitatif important pour l'analyse.

Une description exhaustive de l'événement est nécessaire avant la modélisation. D'abord, la simple constitution des fichiers et le calcul des indices obligent à une certaine rigueur dans les définitions à donner à l'événement, au temps d'observation, aux événements perturbateurs, etc. Ce travail est une étape essentielle de l'analyse où l'on ne s'attarde jamais assez longtemps.

La recherche de différents types d'événements permet déjà de décrire des processus complexes. À chaque risque peut être associé un calendrier spécifique. Avec une description simple, on peut aussi mieux voir l'influence des événements perturbateurs tels que l'émigration, et la limite imposée par l'utilisation de données tronquées, qui sont les données collectées en général dans les enquêtes rétrospectives.

Une fois cette étape descriptive bien maîtrisée, on peut introduire les facteurs explicatifs dans l'analyse de l'événement. Cette prochaine étape est l'aboutissement logique de la recherche des causes en sciences sociales.

CHAPITRE 6

LA MODÉLISATION DE L'ÉVÉNEMENT

Ce qui semble être, dans ce manuel, un long détour par la régression et par les tables de survie, est en fait essentiel pour bien comprendre *l'utilité* de la modélisation en sciences sociales, ainsi que ses limites.

La régression nous a appris comment on peut mesurer l'influence de variables explicatives sur un phénomène, toutes choses égales par ailleurs. Les hypothèses du modèle de régression sont assez fortes, mais nous avons vu quels sont les moyens de les contourner, ou d'en minimiser l'importance. La régression reste, malgré ses défauts, un moyen très efficace de contrôler l'analyse causale et de dépasser la simple analyse descriptive par tableau croisé.

Les tables de séjour permettent de décrire des situations qui évoluent au cours du temps. Ce n'est plus l'état actuel de l'individu auquel on s'intéresse, mais la transition d'un état à un autre au cours de son existence. Les calculs tiennent compte en particulier des troncatures, ce qui est bien utile dans le cas des données d'enquêtes. Cependant la technique reste descriptive : elle nous renseigne peu sur les facteurs explicatifs en dehors du temps lui-même.

1. Le modèle semi-paramétrique de Cox

L'idée (géniale) qu'a eue David R. Cox en 1972, fut de combiner deux types d'analyse : régression et tables de survie. **On peut voir le modèle de Cox comme le contrôle par la régression de l'effet des variables explicatives dans l'analyse de survie, ou bien comme l'introduction de la dimension temporelle dans la régression.** Les avantages d'une technique permettent de combler les lacunes de l'autre.

Dans le cas du modèle logistique, que nous avons illustré au chapitre 4 par l'analyse de la proportion de salariées lors de leur premier emploi, le risque est calculé à un moment donné de la vie de l'individu : on mesure sa probabilité d'occuper un premier emploi salarié, quel que soit l'âge auquel il a effectivement obtenu cet emploi. La durée, le temps écoulé jusqu'au premier emploi, est donc une dimension importante qui fait défaut dans un modèle de type logistique. En particulier, on ne tient pas compte de l'interruption du temps d'observation par la date d'enquête, ou par l'émigration, c'est-à-dire que toute une partie de l'échantillon dont l'observation est tronquée n'est pas prise en compte dans l'analyse si l'on ne prend pas explicitement en compte le temps.

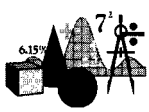
Par ailleurs, si l'on se contente de faire la description de l'entrée dans la vie active par la technique des tables de survie, on sera difficilement en mesure de contrôler l'influence de variables cruciales : scinder l'échantillon en diverses catégories selon la génération, l'origine sociale, etc., mène à de petits sous-échantillons, aux effectifs insuffisants pour l'analyse, en particulier pour mesurer l'influence combinée de plusieurs facteurs explicatifs.

Pour résoudre à la fois le problème de la durée et celui des facteurs explicatifs, l'idée de David R. Cox fut de faire une régression non pas sur une caractéristique acquise par l'individu à l'issue de sa vie (ou au moment de l'enquête), mais sur une caractéristique acquise chaque année de son existence. En quelque sorte, chaque année vécue par chaque membre de l'échantillon constitue une observation jusqu'au moment de l'enquête. La modalité de référence, telle que l'exige le modèle de régression, n'est pas unique pour l'ensemble de l'échantillon, mais elle est propre à chaque durée d'observation : cette série de probabilités permet d'établir une courbe de séjour de référence (par exemple en l'état de "non encore actif" s'il s'agit de l'analyse du premier emploi) appelée encore fonction de séjour de base.

Ce modèle de régression calcule alors l'effet des variables explicatives sur le risque annuel de connaître l'événement. À chaque variable est associé un coefficient de régression qui mesure l'influence moyenne de cette variable sur le risque annuel. En d'autres termes, **l'effet des variables est proportionnel à la probabilité annuelle de connaître l'événement** (d'où l'appellation de ces modèles dits "à risques proportionnels", traduction de l'anglais *proportional hazard models*).

Cette hypothèse de proportionnalité est assez forte, et il est nécessaire de la vérifier pour chaque variable du modèle. Si cette hypothèse n'est pas vérifiée, le modèle devient incohérent, et il faut alors envisager de stratifier l'échantillon selon la variable incriminée. Une méthode graphique permet de tester cette hypothèse avant et après l'introduction de chaque variable dans le modèle.

2. La fonction de séjour de base et les risques relatifs : « cox », « coxbase », « coxhaz », « coxvar » et « coxrel »



Les modèles à risques proportionnels prennent la forme suivante :

$$h_j(t;z_j) = h_0(t) * \exp(\sum_i b_i z_{ij})$$

où $h_0(t)$ est le quotient instantané pour la catégorie de référence (z_{ij}=0), et b_i une série de coefficients associés aux variables indicatrices zij. Le modèle a une composante non paramétrique, la fonction de séjour de base formée de la série des quotients $h_0(t)$, et une composante paramétrique, le vecteur des variables indépendantes. Du fait de ces deux composantes, le modèle à risques proportionnels est aussi appelé **modèle semi paramétrique.**

a) Effet isolé de la durée d'observation

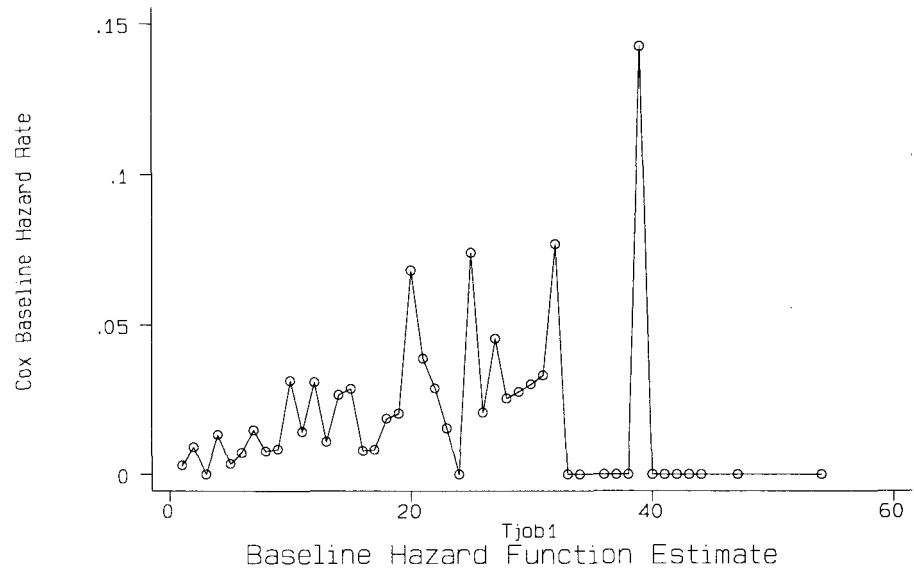
On peut exécuter la commande « cox » sans liste de variables indépendantes :

```
. cox Tjob1, dead(job1) tvid(no)

Iteration 0 :  Log Likelihood =-436.83557

Cox regression                                        Number of obs =     347
                                                      chi2(0)       =    0.00
                                                      Prob > chi2   =       .
Log Likelihood = -436.83557                           Pseudo R2     = 0.0000

------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       t     P>|t|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
------------------------------------------------------------------------------
```

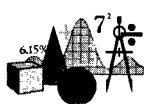
Dans la commande « cox », on remarque l'utilisation de l'option « tvid() » pour désigner l'identifiant individuel qui relie plusieurs observations (périodes) entre elles de la même façon que l'option « tvid() » dans la commande « ltable ».

On peut obtenir une représentation des quotients instantanés avec la commande « coxhaz » :

```
. coxhaz Tjob1 job1, connect(l)
```



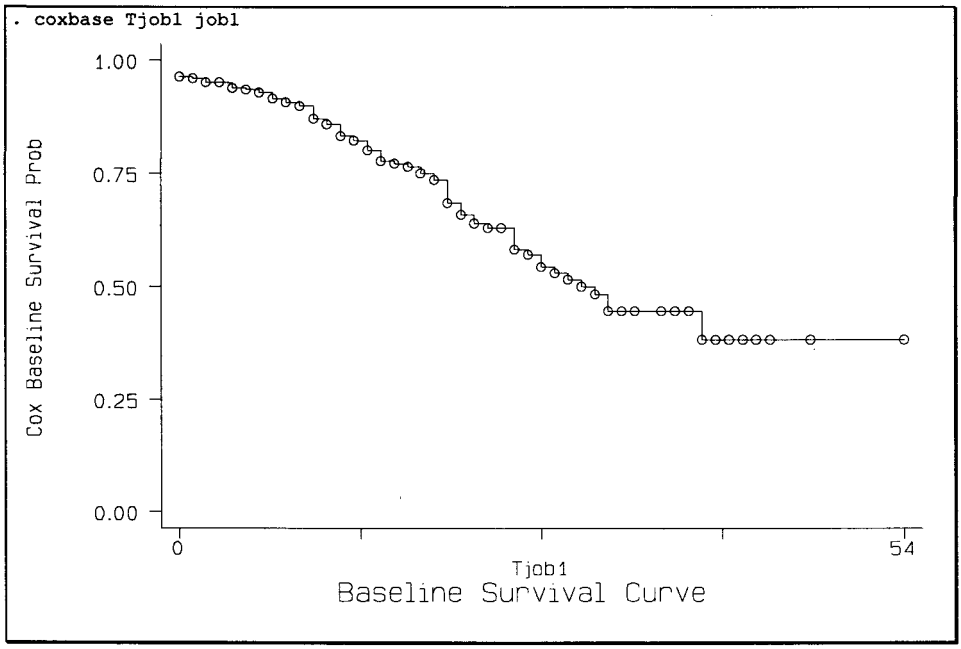

```
Variables created :

 63. baserelh      float   %9.0g           Cox Baseline Relative Hazard
 64. basesurv      float   %9.0g           Cox Baseline Survival Prob
 65. ownsurv       float   %9.0g           Survival Probability
 66. basehaz       float   %9.0g           Cox Baseline Hazard Rate
```



On a ainsi obtenu la série des quotients pour la catégorie de référence :

$$h_j(t;z_j) = h_0(t)$$

La commande « coxbase » : permet de tracer la courbe de séjour de base, tous les paramètres du modèle étant fixés à 0.

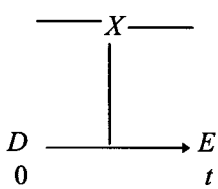


Aucune variable explicative n'a encore été introduite dans le modèle. Le minimum de variables nécessaires pour l'analyse descriptive de l'événement est constitué de la variable qualifiant la durée jusqu'à l'événement, et de l'indice de troncature, à l'aide de l'option « dead() », ainsi nommée selon la tradition héritée de l'analyse de survie. Le résultat de la commande « coxbase » est simplement ce qu'on obtient avec la courbe de Kaplan-Meier :

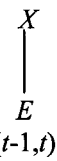
```
. kapmeier Tjob1 job1 if dernier==1
. ltable Tjob1 job1, graph notab tvid(no)
. ltable Tjob1 job1, haz graph notab tvid(no)
```

(graphiques omis)

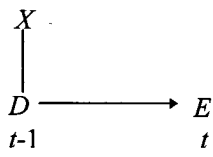
b) Le schéma causal du modèle de Cox



Mais le but du modèle de régression n'est évidemment pas de faire de l'analyse descriptive. Le modèle semi-paramétrique élémentaire correspond au schéma de relation causale où la variable explicative X est une condition permanente qui joue sur le temps entre le début d'observation D et l'événement E.



Il est possible de mettre à jour une telle relation causale dans la mesure où l'observation se fait sur des intervalles de temps réduits et non sur la totalité du temps d'observation. En effet, la régression est calculée sur le quotient instantané, selon le schéma de relation causale ci-contre, pour une unité de temps donnée. Le risque est calculé pour le milieu de l'intervalle (t-1,t) mais il est supposé constant tout au long de cet intervalle (hypothèse d'uniformité) dans la mesure où il est suffisamment court.



On se rapproche du schéma de la relation causale élémentaire telle qu'on l'a décrite au premier chapitre de ce manuel, qui pose comme principe la priorité temporelle de la cause sur l'effet : bien que la condition permanente X ne soit pas un événement, on peut la considérer comme tel sur l'intervalle (t-1,t). En effet, si X est définie au début de chaque intervalle de temps et si le risque calculé est supposé constant sur l'intervalle, on se rapproche de la relation causale, où D (l'entrée en observation en début d'intervalle) fait office d'événement explicatif, puisqu'il faut être présent au temps t-1 pour subir le risque dans l'intervalle (t-1,t). Le schéma se lit ainsi : entrer en observation en t-1 avec la caractéristique X est une cause possible de l'événement E avant la fin de l'intervalle en t.

On est très près de la relation causale de base, bien qu'elle ne soit pas exactement vérifiée : l'effet X est calculé sur l'ensemble des intervalles de temps, c'est-à-dire qu'il représente un risque moyen tout au long du temps d'observation. Chaque variable X (condition permanente) n'est donc pas associée à une unité de temps particulière, ce qui la distingue d'une cause située précisément dans le temps (un événement). Cela a pour conséquence que l'on ne peut analyser avec un tel modèle les variations de calendrier (variation du risque selon l'intervalle de temps) pour un événement selon telle ou telle variable explicative, puisque c'est un effet moyen sur tout le temps d'observation qui est calculé.

c) Interprétation des résultats de la commande « cox »

Pour exécuter la commande « cox » avec variables explicatives, il suffit d'ajouter les noms de ces variables à la suite du nom de la variable de durée avant l'événement :

```
. cox Tjob1 gener2 gener3, dead(job1) tvid(no)

Iteration 0 :  Log Likelihood =-436.83557
Iteration 1 :  Log Likelihood =-433.92053
Iteration 2 :  Log Likelihood =-433.90546
Iteration 3 :  Log Likelihood =-433.90546

Cox regression                                        Number of obs =     347
                                                      chi2(2)       =    5.86
                                                      Prob > chi2   =  0.0534
Log Likelihood = -433.90546                           Pseudo R2     =  0.0067

------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       z     P>|z|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .5975556   .2681007     2.229   0.026        .072088    1.123023
  gener3 |    .160257   .3118209     0.514   0.607       -.4509007    .7714148
------------------------------------------------------------------------------
```

Dans la présentation du modèle sous forme additive, **un coefficient positif ou négatif signifie que l'événement est connu plus ou moins rapidement par rapport à la catégorie de référence**. La durée est bien prise en compte, puisqu'on mesure l'effet des variables explicatives sur le temps que met un individu à connaître l'événement.

La présentation des résultats fournis par la commande « cox » est équivalente à celle des résultats de la commande « logit » dans un modèle logistique, c'est-à-dire que le modèle est présenté sous forme additive. Si l'on préfère présenter les résultats sous forme de quotients relatifs (dans un modèle multiplicatif), il suffit de calculer l'exponentielle des coefficients, ce que fait très bien la commande « cox » avec l'option « hr » (pour *hazard ratio*), qui équivaut à la commande « logistic » pour obtenir les rapports de chance :

```
. cox, hr

Cox regression                                        Number of obs =     347
                                                      chi2(2)       =    5.86
                                                      Prob > chi2   =  0.0534
Log Likelihood = -433.90546                           Pseudo R2     =  0.0067

------------------------------------------------------------------------------
   Tjob1 |
    job1 | Haz. Ratio   Std. Err.       z     P>|z|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |    1.81767   .4873186     2.229   0.026        1.07475    3.074134
  gener3 |   1.173813   .3660193     0.514   0.607        .6370541   2.162824
------------------------------------------------------------------------------
```

On interprète l'effet du groupe de générations comme multipliant par 1,8 les chances annuelles d'obtenir un premier emploi par les générations 1945-54 par rapport aux générations 1930-44, et par seulement 1,2 pour les générations 1955-64, toujours par rapport aux générations 1930-44. Seul le coefficient pour les générations 1945-54 est significatif au seuil de 5 %.

Contrairement au modèle logistique, calculer le risque pour la catégorie de référence n'a pas de sens : en effet, ce risque est différent d'une année d'observation à l'autre, comme on l'a vu avec la courbe des quotients instantanés. On notera cependant au passage que le risque dans un modèle semi-paramétrique n'est pas exprimé sous forme logit comme dans le modèle logistique, mais s'interprète comme le rapport entre un nombre d'événements et une population soumise au risque au milieu de l'intervalle de temps.

d) La vérification de l'hypothèse de proportionnalité : la commande « loglogs2 »

Comme on l'a dit précédemment, le modèle semi-paramétrique impose que l'effet des variables explicatives soit proportionnel à la probabilité annuelle de connaître l'événement. Cette hypothèse, assez forte, n'est pas forcément vérifiée, et il est donc important de la tester.

Pour vérifier l'hypothèse de proportionnalité, la commande « loglogs2 » trace un graphique du $\log\left(-\log\left(S(t)\right)\right)$ selon $\log(t)$, où $S(t)$ représente la fonction de séjour calculée par l'estimateur de Kaplan-Meier (obtenu avec la commande « kapmeier »).

La commande « loglogs2 » est identique à la commande « loglogs » telle que fournie dans la version 4.0 de STATA, avec trois options supplémentaires « slope », « tmin(#) » et « tmax(#) ».

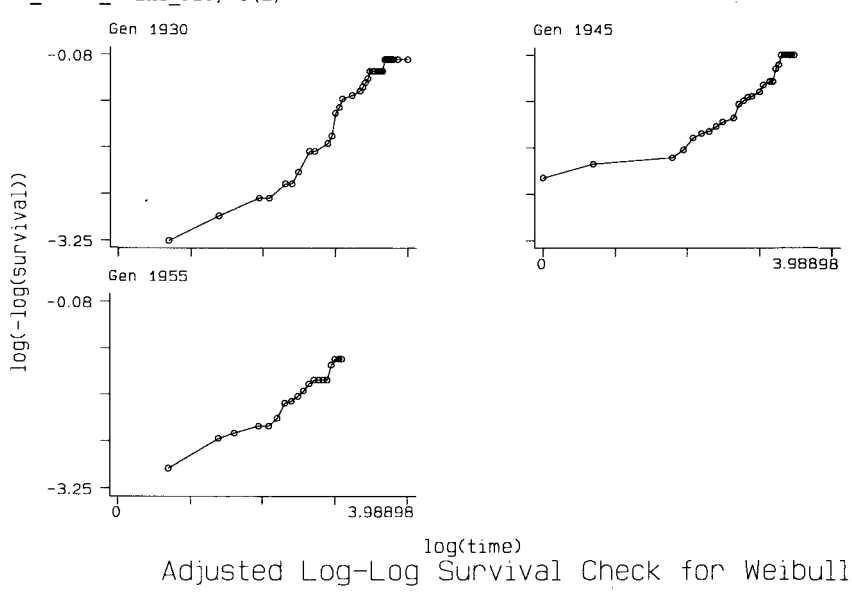
Si l'option « by(*byvar*) » est spécifiée, un graphique séparé pour chaque catégorie de l'échantillon définie par *varlist* sera tracé. L'option « adjustfor(*varlist*) » ajuste la fonction de séjour $S(t)$ selon un modèle semi-paramétrique, avec *byvar* et *varlist* comme variables indépendantes. On peut donc tester l'hypothèse de proportionnalité sur les variables indépendantes d'un modèle semi-paramétrique.

Considérons les résultats du modèle estimé avec l'instruction et le groupe de générations comme variables explicatives :

```
. cox Tjob1 gener2 gener3 ins_2 ins_3 ins_4 ins_5 ins_6 ins_7 ins_8 ins_910,
dead(job1) tvid(no)

Iteration 0:  Log Likelihood =-436.83557
(etc.)
Iteration 4:  Log Likelihood =-428.06128

Cox regression                                    Number of obs =     347
                                                  chi2(10)      =   17.55
                                                  Prob > chi2   =  0.0631
Log Likelihood = -428.06128                       Pseudo R2     =  0.0201

------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .5731449    .283006     2.025   0.043      .0184633    1.127827
  gener3 |  -.0284077   .3557704    -0.080   0.936     -.7257049    .6688894
   ins_2 |   .2366023   .4729104     0.500   0.617     -.6902851     1.16349
   ins_3 |   .2854057   .5301826     0.538   0.590     -.7537331    1.324544
   ins_4 |  -.2933274   .3332683    -0.880   0.379     -.9465214    .3598665
   ins_5 |  -.1399417   .6104345    -0.229   0.819     -1.336371    1.056488
   ins_6 |    .967581   .3895223     2.484   0.013      .2041313    1.731031
   ins_7 |   1.011043   .4736803     2.134   0.033      .0826465    1.939439
   ins_8 |   .2100297   .6304098     0.333   0.739     -1.025551     1.44561
 ins_910 |    .441251   1.037453     0.425   0.671     -1.592119    2.474621
------------------------------------------------------------------------------
```

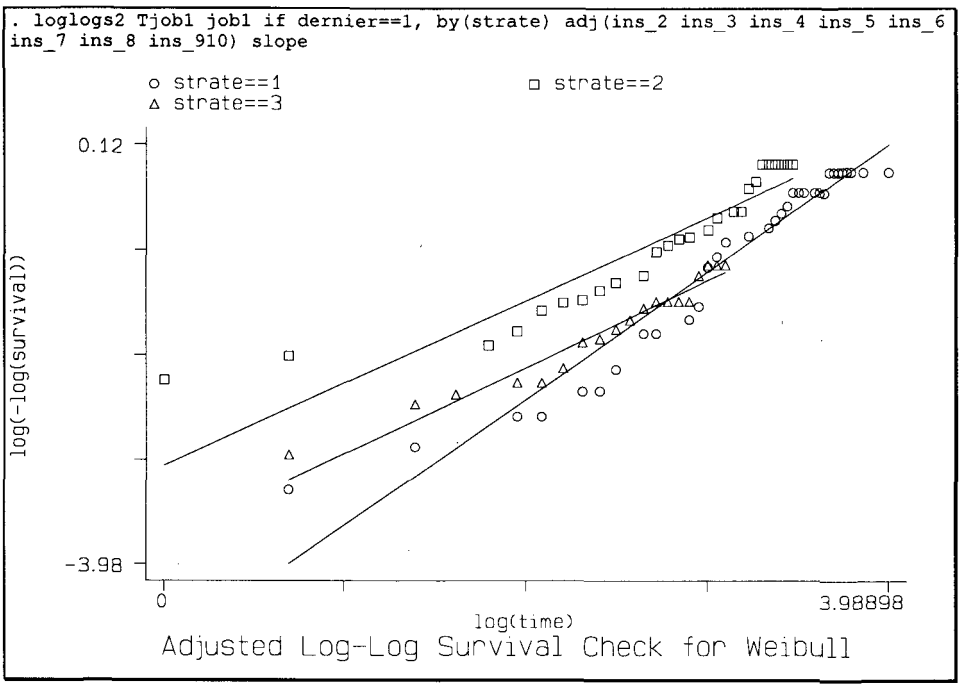
On pourra tester l'hypothèse de proportionnalité sur la variable "strate" de la façon suivante :

```
. loglogs2 Tjob1 job1 if dernier==1, by(strate) adj(ins_2 ins_3 ins_4 ins_5 ins_6
ins_7 ins_8 ins_910) c(l)
```



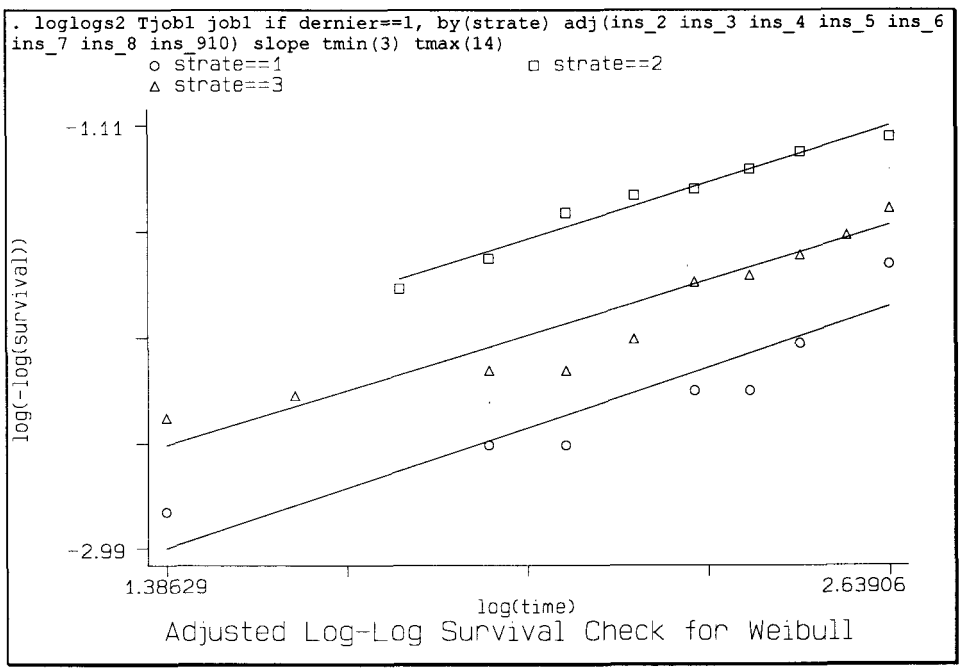
En principe, si l'hypothèse de proportionnalité est vérifiée, les courbes doivent suivre la même pente. Mais ces graphiques sont difficiles à interpréter : les pentes sont-elles vraiment différentes ? L'option « slope » que nous avons ajoutée au programme « loglogs » permet de regrouper ces trois graphiques en un seul et de tracer les pentes (*slopes*) pour chaque catégorie de la variable indiquée dans l'option « by() », de façon à vérifier l'hypothèse de proportionnalité :

```
. loglogs2 Tjob1 job1 if dernier==1, by(strate) adj(ins_2 ins_3 ins_4 ins_5 ins_6
ins_7 ins_8 ins_910) slope
```

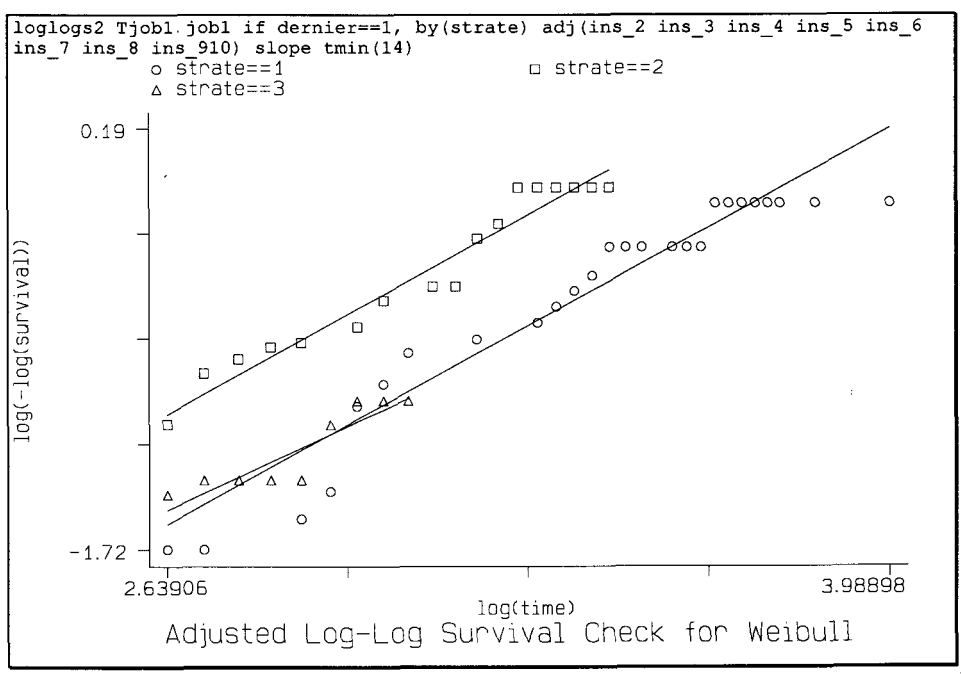


D'après ce graphique, l'hypothèse de proportionnalité ne semble pas vérifiée. Les pentes ne sont pas parallèles lorsqu'on considère toute la période d'observation.

Cependant, si l'on y regarde de plus près, la situation n'est pas aussi désespérée. Les options « tmin(#) » et « tmax(#) » permettent de tracer la pente pour un intervalle donné de la période d'observation. Par exemple, de la durée 3 à la durée 14, c'est-à-dire de la durée 1,1 (log 3=1,0986) à la durée 2,6 (log 14=2,639) sur le graphique "log-log", on a :



Les pentes paraissent maintenant remarquablement parallèles. Il en est de même après la durée 14 :



Ces tests graphiques sont très utiles pour vérifier l'hypothèse de proportionnalité sur des variables indépendantes fixes au cours du temps, c'est-à-dire définies dès l'entrée en observation et jusqu'au moment de l'événement ou de la troncature.

Malheureusement, ces tests ne sont pas valables pour les variables indépendantes qui varient en fonction du temps, dont on verra plus loin l'intérêt pour l'analyse des biographies. Nous reviendrons en conclusion sur les difficultés de mesure de l'ajustement du modèle de Cox par rapport aux données.

3. L'analyse des événements concurrents

Comme on l'a suggéré lors de l'analyse du premier emploi avec le modèle logistique et avec les courbes de quotients cumulés, dites de Aalen, le premier emploi peut revêtir différentes formes. Les femmes peuvent par exemple acquérir un emploi salarié ou indépendant. Cela suggère de faire l'analyse des différents types de sortie d'un état donné ("non encore actif", pour le cas de l'entrée dans la vie active).

Le traitement des événements concurrents est très simple : il suffit de créer un indice de troncature pour chaque événement, ce que nous avons déjà fait pour illustrer les courbes de Aalen, et de faire la régression pour cette variable. Le modèle explicatif du risque d'avoir un premier emploi salarié donne les résultats suivants :

```
. cox Tsal gener2 gener3, dead(sal) tvid(no)

Iteration 0 :  Log Likelihood =-232.32805
(etc.)
Iteration 3 :  Log Likelihood =-229.60174

Cox regression                                        Number of obs =     347
                                                      chi2(2)       =    5.45
                                                      Prob > chi2   =  0.0655
Log Likelihood = -229.60174                           Pseudo R2     =  0.0117

------------------------------------------------------------------------------
    Tsal |
     sal |      Coef.   Std. Err.       z     P>|z|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   1.003692   .4657514     2.155   0.031        .0908356    1.916548
  gener3 |   .7423944   .4743442     1.565   0.118       -.1873032    1.672092
------------------------------------------------------------------------------
```

On peut tester la différence entre les deux coefficients associés aux groupes de générations et voir qu'elle n'est pas significative :

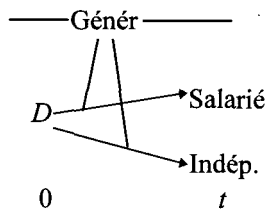
```
. test gener2=gener3

 ( 1)  gener2 - gener3 = 0.0

           chi2(  1) =    0.64
         Prob > chi2 =    0.4242
```

Le modèle explicatif du risque de devenir indépendante lors de l'obtention d'un premier emploi donne les résultats suivants :

```
. cox Tind gener2 gener3, dead(ind) tvid(no)

Iteration 0 :  Log Likelihood =-205.51069
(etc.)
Iteration 3 :  Log Likelihood =-203.63456

Cox regression                                        Number of obs =     347
                                                      chi2(2)       =    3.75
                                                      Prob > chi2   =  0.1532
Log Likelihood = -203.63456                           Pseudo R2     =  0.0091

------------------------------------------------------------------------------
    Tind |
     ind |      Coef.   Std. Err.       z     P>|z|       [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .3343448   .3406716     0.981   0.326      -.3333593    1.002049
  gener3 |  -.5441694   .5143454    -1.058   0.290      -1.552268    .4639291
------------------------------------------------------------------------------
```

Cette fois le test d'inégalité des coefficients est significatif au seuil de 10 % :

```
. test gener2=gener3

 ( 1)  gener2 - gener3 = 0.0

           chi2(  1) =    3.26
         Prob > chi2 =    0.0709
```



En somme, en distinguant les emplois salariés et indépendants, on voit très nettement que l'augmentation de la probabilité annuelle d'obtenir un emploi pour les générations 1945-54 est essentiellement due à l'emploi salarié. Par rapport à ce groupe de génération, la probabilité d'obtenir un emploi a baissé pour les générations 1955-64, mais plus particulièrement en ce qui concerne l'emploi indépendant.

4. Les variables indépendantes fonction du temps

Les variables explicatives les plus communément utilisées sont celles qui caractérisent l'individu à sa naissance : il s'agit par exemple, du sexe, de

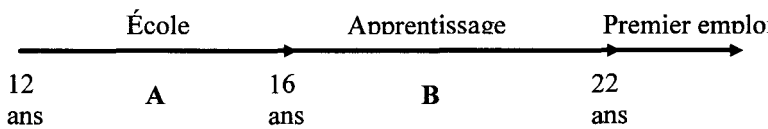
l'appartenance à un groupe ethnique, à une caste, etc. On suppose que l'effet de chacune de ces variables est constant tout au long de la vie de l'individu : c'est ce qu'on a appelé des conditions permanentes. Leur effet est supposé proportionnel à la probabilité annuelle de connaître l'événement étudié comme on l'a expliqué plus haut.

Les événements qu'a connu l'individu depuis sa naissance ou le début de l'observation, et qui ont pu influer sur ses chances de connaître l'événement étudié, peuvent aussi être introduits sous la forme de variables explicatives : ces variables sont dites alors "fonction du temps". Leur introduction a pour effet de rendre le modèle plus dynamique puisque cela permet de suivre au plus près le cheminement des individus.

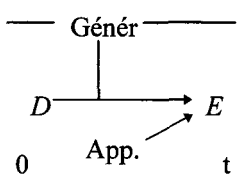
Nous verrons comment introduire les variables fonction du temps lorsqu'elles sont définies selon un processus interne (ou endogène) et lorsqu'elles sont définies selon un processus externe (ou exogène). Rappelons cependant que, dans les deux cas, il n'est pas possible de vérifier graphiquement l'hypothèse de proportionnalité des quotients, comme on l'a fait précédemment pour les variables permanentes.

a) L'effet des différentes périodes précédant l'événement

Prenons le cas d'un individu que l'on observe depuis l'âge de 12 ans sur le graphique suivant :

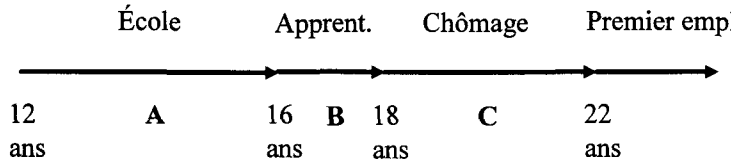


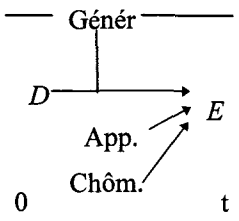
À l'issue de sa période d'étude, cet individu aurait pu trouver un travail mais il a débuté un apprentissage à 16 ans. Ses chances de trouver un travail ont certainement été modifiées par cet événement, et par conséquent la probabilité de trouver un emploi est sans doute différente en période A et en période B. On cherchera à mesurer la dépendance stochastique entre l'événement étudié (l'accès au premier emploi) et l'événement perturbateur (le changement de statut pré-professionnel : dans l'exemple, le passage du statut d'élève à celui d'apprenti).



Par rapport au modèle simple où n'intervient qu'une condition permanente (par exemple l'appartenance à une génération), on a introduit une variable qui évolue dans le temps, c'est-à-dire entre le début de l'observation et l'événement.

L'individu aurait pu connaître un parcours différent, où l'apprentissage n'aurait duré que deux ans, et aurait été suivi par une période de chômage. Ses chances d'obtenir un emploi seraient alors modifiées en conséquence :



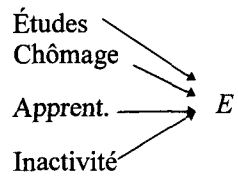


Cela veut dire qu'il ne faut pas seulement prendre en compte le moment de l'événement perturbateur en tant que tel (l'âge d'entrée en apprentissage) mais bien la période depuis cet événement jusqu'à l'événement suivant (le début du chômage ou le début de la vie active dans notre exemple). Le schéma causal ci-contre résume les relations entre variables.

En définitive, l'ensemble des changements pré-professionnels est susceptible de modifier l'accès au premier emploi. On peut calculer à chaque âge la probabilité d'entrer dans la vie active, selon qu'on est étudiant, apprenti ou chômeur, pour l'ensemble des individus. L'entrée dans chacun de ces statuts constitue une **variable indépendante qu'on dit "fonction du temps"**, parce qu'elle intervient en cours d'observation, ce qui la distingue des conditions permanentes qui caractérisent l'individu au début de l'observation, et dont l'effet est supposé constant tout au long de l'observation. Les variables fonction du temps sont des événements qui peuvent modifier le cours de la vie d'un individu : on est donc là au plus près de la relation causale élémentaire, qui pose comme principe la priorité temporelle de la cause sur l'effet.

b) L'interprétation des résultats

On voit maintenant l'originalité du traitement de l'analyse de survie avec le logiciel STATA : en acceptant plusieurs lignes d'observations par individu (représentant chacune une période) et en les reliant par l'identifiant de l'individu grâce à l'option « tvid() », la commande « cox » peut traiter les variables indépendantes fonction du temps comme n'importe quelle autre variable. La prise en compte de ces variables indépendantes fonction du temps se fait simplement en introduisant la liste de ces variables dans le modèle.



Par exemple, l'effet du statut d'activité (ou d'inactivité) peut être introduit sous forme polytomique, chaque statut étant exclusif des autres. Nous pouvons créer les modalités suivantes (les périodes d'études constituent la catégorie de référence) :

```
. gen byte chomage=v604==2

. gen byte inactiv=v604>=4

. gen byte apprent=v611==4
```

Il suffit alors d'ajouter ces variables aux précédentes dans le modèle :

```
. cox Tjob1 gener2 gener3 chomage apprent inactiv, dead(job1) tvid(no)

Iteration 0 :  Log Likelihood =-436.83557
(etc.)
Iteration 5 :  Log Likelihood = -416.1552

Cox regression                                    Number of obs =     347
                                                  chi2(5)       =   41.36
                                                  Prob > chi2   =  0.0000
Log Likelihood =  -416.1552                       Pseudo R2     =  0.0473

------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       z     P>|z|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .3867146   .2727985     1.418   0.156      -.1479606    .9213898
  gener3 |  -.3991418   .3312584    -1.205   0.228      -1.048396    .2501128
 chomage |   .2815345   .4845052     0.581   0.561      -.6680783    1.231147
 apprent |   .6999419   .4503057     1.554   0.120      -.1826411    1.582525
 inactiv |  -1.333213   .2856869    -4.667   0.000      -1.893149   -.7732772
------------------------------------------------------------------------------
```

On voit que la différence entre les générations 1930-44 et 1945-54 est beaucoup moins grande, et n'est d'ailleurs pas significative au seuil de 15 %.Le modèle fait aussi état d'un retard de calendrier des générations 1955-64 par rapport aux générations 1930-44, mais la différence n'est pas non plus significative.

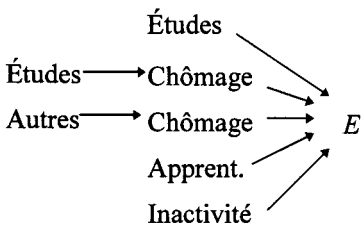
L'interprétation est très différente du premier modèle où n'intervenait que la variable "groupe de générations". Les changements sont dus à l'introduction des variables qui qualifient l'activité ou l'inactivité avant le premier emploi. Les coefficients produits par la commande « cox » s'interprètent de la même façon que ceux produits par la commande « logit ».

Ainsi, le coefficient de -1,333, significatif à un seuil inférieur à 1 pour 1 000, pour la catégorie "inactiv" signifie que les femmes inactives (au foyer) ont leur chance d'obtention d'un premier emploi réduite de 73,6 % (ou si l'on veut, divisée par 3,8) par rapport à celles qui font des études (catégorie de référence). Le statut d'apprentie ou d'aide familiale semble en revanche plus favorable (multiplication des chances par 2) bien que le coefficient ne soit significatif qu'au seuil de 12 %. Par

rapport aux études, le chômage a un léger effet positif, mais non significatif, sur le risque d'obtention d'un premier emploi.

c) L'effet markovien ou la "mémoire des événements"

Il est possible que la succession des transitions intermédiaires vers l'événement étudié ait un effet propre. Par exemple, le chômage après l'apprentissage peut avoir un effet différent du chômage après les études. Dans ce cas, chaque période conserverait la "mémoire" de la période précédente, et pourquoi pas, de toutes les périodes connues antérieurement. On appelle ce type d'effet, un **effet markovien** d'ordre 2 ou plus, par opposition à l'effet markovien d'ordre 1 qui stipule que la probabilité de connaître un événement dépend seulement de l'état actuel (d'ordre 1) dans lequel se trouve l'individu et non pas des états antérieurs (d'ordre 2 ou plus) dans lesquels il a pu se trouver.



Par exemple, en ce qui concerne l'entrée dans la vie active, on peut tenir compte des effets markoviens d'ordre 2, pour qualifier les périodes de chômage selon qu'elles ont eu lieu après les études ou non (schéma causal ci-contre).

On créera les variables adéquates pour modifier le modèle :

```
. quietly by no : gen byte chomet=chomage==1 & v604[_n-1]==3 & no==no[_n-1]

. quietly by no : gen byte chomaut=chomage==1 & v604[_n-1]!=3 & no==no[_n-1]
```

Il suffit ensuite de remplacer dans le modèle précédent, la variable "chomage" par les variables "chomet" et "chomaut" :

```
. cox Tjob1 gener2 gener3 chomet chomaut appr inact, dead(job1) tvid(no)
Iteration 0 :  Log Likelihood =-436.83557
(etc.)
Iteration 4 :  Log Likelihood =-415.37774
Cox regression                                    Number of obs =     347
                                                  chi2(6)       =   42.92
                                                  Prob > chi2   =  0.0000
Log Likelihood = -415.37774                       Pseudo R2     =  0.0491
------------------------------------------------------------------------------
       Tjob1 |
        job1 |      Coef.   Std. Err.      t    P>|t|     [95 % Conf. Interval]
-------------+----------------------------------------------------------------
     gener2  |   .4160835   .2726379     1.526  0.128     -.1201803    .9523474
     gener3  |  -.4428013   .3330378    -1.330  0.185     -1.097868    .2122658
     chomet  |   .6142377   .5225903     1.175  0.241     -.4136686    1.642144
    chomaut  |  -.4211404   1.042258    -0.404  0.686     -2.471206    1.628925
     apprent |   .7982103   .4606059     1.733  0.084     -.1077761    1.704197
     inactiv |   -1.32081    .286476    -4.611  0.000     -1.884292   -.7573273
------------------------------------------------------------------------------
```

L'effet des autres variables se maintient, mais on voit que le chômage a un effet inverse selon qu'il suit les études ou une autre situation (apprentissage). Dans le cas du chômage après les études, les chances sont plus grandes d'obtenir un emploi par rapport à la période des études, alors que c'est l'inverse dans le cas du chômage après l'apprentissage. La femme au chômage a d'autant moins de chances d'obtenir un premier emploi qu'elle était apprentie auparavant.

d) Les chaînes de transition

Le modèle de Cox donne la possibilité de calculer les risques pour deux ou plusieurs événements concurrents. On pourra donc tester le schéma causal selon qu'il s'agit d'un emploi salarié ou indépendant. On constate que les variables significatives ne sont pas les mêmes dans les deux cas. Pour le salariat, le modèle donne les résultats suivants :

```
. cox Tsal gener2 gener3 chomet chomaut appr inact, dead(sal) tvid(no)

Iteration 0 :  Log Likelihood =-232.32805
(etc.)
Iteration 32 :  Log Likelihood =-214.34595

Cox regression                                    Number of obs =     347
                                                  chi2(6)       =   35.96
                                                  Prob > chi2   =  0.0000
Log Likelihood = -214.34595                       Pseudo R2     =  0.0774

------------------------------------------------------------------------------
    Tsal |
     sal |      Coef.   Std. Err.       z     P>|z|     [95 % Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .6580211   .4777176     1.377   0.168    -.2782883     1.59433
  gener3 |   .0008383   .5000517     0.002   0.999     -.979245    .9809216
  chomet |   .7127041   .5890486     1.210   0.226    -.4418099    1.867218
 chomaut |  -35.87011   6.46e+07    -0.000   1.000    -1.27e+08    1.27e+08
 apprent |   .1445094   .6536236     0.221   0.825    -1.136569    1.425588
 inactiv |  -1.715093   .3820457    -4.489   0.000    -2.463889   -.9662973
------------------------------------------------------------------------------
```

La différence de risque d'obtention d'un premier emploi salarié entre la première et la troisième génération est quasiment nulle. L'apprentissage n'a pratiquement aucun effet et l'inactivité rend encore plus improbable l'accès au salariat.

On remarque aussi que le coefficient pour la modalité "chomaut" est particulièrement élevé dans les valeurs négatives. Cela vient du fait qu'aucune femme ayant connu le chômage après une autre période que les études, n'avait accédé à l'emploi salarié. En effet, comme la commande « mlogit », « cox » permet de faire des estimations tenant compte des risques nuls pour certaines catégories de l'échantillon.

Dans le cas de risque nul, le coefficient tend vers moins l'infini. Cela a deux conséquences sur l'estimation du modèle. D'abord, le calcul peut ne pas converger, d'où le nombre important d'itérations successives (32) avant la production des résultats. Il faudra alors fixer un nombre maximum d'itérations avec l'option « iter(#) », ou bien fixer un seuil de tolérance pour lequel on considérera la convergence atteinte avec l'option « ltol(#) ». Lorsqu'on s'aperçoit que le modèle ne converge pas ou converge lentement, on peut interrompre la commande et la relancer avec les options adéquates :

```
. cox Tsal gener2 gener3 chomet chomaut appr inact, dead(sal) tvid(no) iter(10)
(listing omis)

. cox Tsal gener2 gener3 chomet chomaut appr inact, dead(sal) tvid(no)
ltol(.000001)
(listing omis)
```

La deuxième conséquence d'un risque nul est que le coefficient calculé ne pourra jamais être *significatif*, bien que l'on soit en présence d'un effet particulièrement *signifiant*. Il ne faut surtout pas négliger ou éliminer les modalités pour lesquelles le risque est nul : le modèle y gagnera en pertinence, bien qu'il sera impossible d'en évaluer la significativité (chapitre sur le modèle logistique).

Par rapport à l'accès au salariat, les effets de l'apprentissage et de l'inactivité sont quasiment inverses pour l'accès à l'emploi indépendant : l'apprentissage augmente fortement et significativement la probabilité (par près de 10), et l'inactivité n'a plus aucun effet significatif :

```
. cox Tind gener2 gener3 chomet chomaut appr inact, dead(ind) tvid(no)

Iteration 0 :  Log Likelihood =-205.51069
(etc.)
Iteration 4 :  Log Likelihood =-196.98216

Cox regression                                    Number of obs =     347
                                                  chi2(6)       =   17.06
                                                  Prob > chi2   =  0.0091
Log Likelihood = -196.98216                       Pseudo R2     =  0.0415

------------------------------------------------------------------------------
        Tind |
         ind |      Coef.   Std. Err.      z    P>|z|     [95 % Conf. Interval]
-------------+----------------------------------------------------------------
     gener2  |   .2844368     .34754     0.818   0.413    -.3967291    .9656026
     gener3  |    -1.0045   .5574695    -1.802   0.072    -2.097121    .0881197
     chomet  |   .9645942   1.147673     0.840   0.401    -1.284804    3.213992
    chomaut  |   1.159314   1.135531     1.021   0.307    -1.066287    3.384914
    apprent  |   2.294355   .7309039     3.139   0.002     .8618101    3.726901
    inactiv  |  -.2958885   .5282479    -0.560   0.575    -1.331235    .7394583
------------------------------------------------------------------------------
```

L'accès à l'emploi indépendant à la suite d'une période de chômage postérieure à une période d'inactivité ou d'apprentissage, est maintenant possible, mais n'est pas significativement différent de l'accès à l'emploi indépendant après des études (modalité de référence).

5. La description d'une interaction entre événements : la commande « nonpar »

Comme illustration des interactions possibles entre deux événements, on se propose ici d'étudier les effets réciproques du mariage et du premier emploi. On utilisera la durée en mois depuis l'anniversaire des 12 ans jusqu'à l'événement, que celui-ci soit un changement d'activité ou un changement de statut matrimonial.

Après avoir construit à partir des données le fichier adéquat, nous verrons successivement comment décrire une interaction et comment évaluer cette interaction dans un modèle semi-paramétrique.

a) La préparation des fichiers pour l'analyse de l'interaction d'événements de différente nature : la commande « tmerge »

L'influence de l'itinéraire de formation (études, apprentissage, etc.) sur la probabilité d'obtenir un emploi peut être analysée en tant que rendement du capital humain, c'est-à-dire comme une **variable interne** (ou endogène) au processus d'entrée ou de mobilité sur le marché du travail. Mais il est d'autres variables qui peuvent influer sur ce processus : on pense par exemple aux **variables externes** (ou exogènes) au processus d'accès à l'emploi, comme le mariage, l'itinéraire d'une autre personne du ménage (le conjoint, les enfants...), la fermeture d'une usine, un changement de la législation du travail, etc. Ces événements explicatifs ont pour particularité de n'avoir pas forcément lieu avant l'événement étudié au contraire des événements internes au processus.

Nous allons prendre pour exemple le cas où la situation matrimoniale est introduite comme variable explicative pour expliquer l'entrée en activité des femmes. Il nous faut d'abord créer un fichier où figureront à la fois les états matrimoniaux successifs et les périodes d'activité jusqu'à l'enquête. Ce nouveau fichier, produit de la fusion de deux fichiers biographiques (l'un pour l'état matrimonial MAR.DTA, l'autre pour l'emploi JOB.DTA) est simplement créé avec la commande « tmerge » figurant sur la disquette accompagnant ce manuel.

Pour bien déterminer la succession des événements qui ont eu lieu la même année, il est souhaitable de travailler sur des dates aussi précises que possibles (en mois, par exemple) pour éviter les événements simultanés (un mariage qui se produit la même année que l'entrée en activité). Dans le cas où deux événements se

sont produits à la même date (le même mois), il faudra faire un choix pour déterminer lequel s'est passé avant l'autre (voir la section suivante).

Le calcul d'une durée au mois près suppose évidemment qu'on dispose d'une variable indiquant le mois de l'événement. En annexe 3 figurent des conseils pratiques pour résoudre le problème de la datation imprécise des événements. Dans le cas où ce problème n'aurait pas pu être résolu au moment de la collecte, on serait contraint de ne considérer la date qu'à l'année près.

Il nous faut d'abord recalculer une durée (en mois) avant l'événement dans chaque fichier, avant de les fusionner (les variables indiquant le mois sont supposées avoir été contrôlées et les erreurs nettoyées) :

```
. use job

. drop event job1 sal ind emig1 T*

. gen int duract=(v638*12+v637) - (v306*12+v305)

. lab var duract "Mois avant nouvelle activite"

. sort no duract

. by no : list duract v637 v638 v305 v306 v611

-> no=       3
       duract  v637  v638  v305  v306  v611
  1.      532    10    84     6    40     0
  2.      593    11    89     6    40     3

-> no=       4
       duract  v637  v638  v305  v306  v611
  3.      378    11    89     5    58     0

-> no=      11
       duract  v637  v638  v305  v306  v611
  4.      372     6    86     6    55     0
  5.      413    11    89     6    55     3

-> no=      13
       duract  v637  v638  v305  v306  v611
  6.      226     4    73     6    54     0
  7.      384     6    86     6    54     0
  8.      425    11    89     6    54     3
(etc)
-> no=     113
       duract  v637  v638  v305  v306  v611
 74.      257    11    81     6    60     0
 75.      353    11    89     6    60     0

-> no=     114
       duract  v637  v638  v305  v306  v611
 76.      581    11    89     6    41     3

-> no=     116
       duract  v637  v638  v305  v306  v611
 77.      374     8    65     6    34     0
 78.      665    11    89     6    34     3
(etc)
. save job, replace
```

```
. use mar

. gen int durmar=(fin*12+moisfin)-(v306*12+v305)

. lab var durmar "Mois avant changement statut"

. sort no durmar

. by no : list durmar moisfin fin v305 v306 statut

-> no=       3
        durmar    moisfin       fin  v305  v306    statut
   1.      203          5        57     6    40   celibat
   2.      593         11        89     6    40  mariage1

-> no=       4
        durmar    moisfin       fin  v305  v306    statut
   3.      378         11        89     5    58   celibat

-> no=      11
        durmar    moisfin       fin  v305  v306    statut
   4.      192          6        71     6    55   celibat
   5.      413         11        89     6    55  mariage1

-> no=      13
        durmar    moisfin       fin  v305  v306    statut
   6.      212          2        72     6    54   celibat
   7.      425         11        89     6    54  mariage1
(etc)
-> no=     113
        durmar    moisfin       fin  v305  v306    statut
  87.      226          4        79     6    60   celibat
  88.      288          6        84     6    60  mariage1
  89.      298          4        85     6    60  veu mar1
  90.      353         11        89     6    60  mar2veu1

-> no=     114
        durmar    moisfin       fin  v305  v306    statut
  91.      166          4        55     6    41   celibat
  92.      451          1        79     6    41  mariage1
  93.      456          6        79     6    41  veu mar1
  94.      581         11        89     6    41  mar2veu1

-> no=     116
        durmar    moisfin       fin  v305  v306    statut
  95.      215          5        52     6    34   celibat
  96.      264          6        56     6    34  mariage1
  97.      274          4        57     6    34  div mar1
  98.      644          2        88     6    34  mar2div1
  99.      665         11        89     6    34  veu2div1
(etc)
. save mar, replace
```

Les deux fichiers originaux JOB.DTA et MAR.DTA doivent être classés selon la variable "no" (identifiant l'individu). De plus, les dates d'enquête (troncatures) doivent être les mêmes dans les deux fichiers, c'est-à-dire que les variables de durée ("durmar" et "duract", dans notre cas) doivent avoir la même valeur pour chaque dernier enregistrement de chaque individu.

On peut alors exécuter la commande :

```
. tmerge no job(duract) mar(durmar) jobmar(durev)
set maxvar 114 width 141

The variable '_File' indicates in which file the time change is originated,
either job , mar , or both.
 Record from|
     file...|      Freq.     Percent        Cum.
------------+-----------------------------------
        job |         436       24.97       24.97
        mar |         778       44.56       69.53
       both |         532       30.47      100.00
------------+-----------------------------------
      Total |        1746      100.00
file jobmar.dta saved
```

puis libeller la nouvelle variable de durée (appelée ici "durev"), et enfin sauvegarder le nouveau fichier :

```
. lab var durev "Mois avant evenement"

. save jobmar, replace
```

La commande « tmerge » conserve toutes les variables contenues dans les deux fichiers originaux (y compris les variables de durée) et ajoute la nouvelle variable de durée ainsi qu'une variable "_File" indiquant de quel fichier provient chaque période d'observation.

On peut vérifier le contenu du nouveau fichier :

```
. by no : list durev duract durmar v611 statut _File

-> no=       3
         durev    duract    durmar  v611    statut     _File
   1.      203       532       203     0   celibat       mar
   2.      532       532       593     0  mariage1       job
   3.      593       593       593     3  mariage1      both

-> no=       4
         durev    duract    durmar  v611    statut     _File
   4.      378       378       378     0   celibat      both

-> no=      11
         durev    duract    durmar  v611    statut     _File
   5.      192       372       192     0   celibat       mar
   6.      372       372       413     0  mariage1       job
   7.      413       413       413     3  mariage1      both

-> no=      13
         durev    duract    durmar  v611    statut     _File
   8.      212       226       212     0   celibat       mar
   9.      226       226       425     0  mariage1       job
  10.      384       384       425     0  mariage1       job
  11.      425       425       425     3  mariage1      both
(etc)
```

```
-> no=      113
         durev    duract    durmar  v611     statut      _File
 124.      226       257       226     0    celibat        mar
 125.      257       257       288     0   mariage1        job
 126.      288       353       288     0   mariage1        mar
 127.      298       353       298     0   veu mar1        mar
 128.      353       353       353     0   mar2veu1       both

-> no=      114
         durev    duract    durmar  v611     statut      _File
 129.      166       581       166     3    celibat        mar
 130.      451       581       451     3   mariage1        mar
 131.      456       581       456     3   veu mar1        mar
 132.      581       581       581     3   mar2veu1       both

-> no=      116
         durev    duract    durmar  v611     statut      _File
 133.      215       374       215     0    celibat        mar
 134.      264       374       264     0   mariage1        mar
 135.      274       374       274     0   div mar1        mar
 136.      374       374       644     0   mar2div1        job
 137.      644       665       644     3   mar2div1        mar
 138.      665       665       665     3   veu2div1       both
(etc)
```

b) Les principes de la description d'une interaction

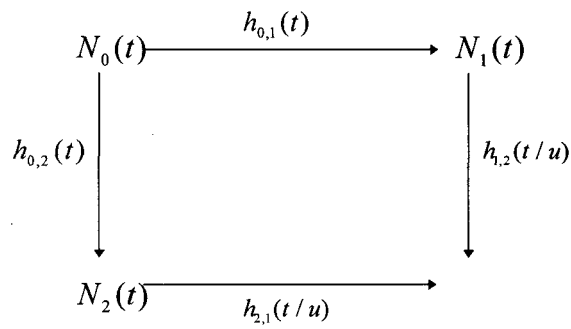
Paradoxalement, les techniques d'analyse semi-paramétrique et paramétrique de l'interaction des événements ont évolué plus vite que les analyses non paramétriques. Il semble cependant logique, dans un but pédagogique, que la description du phénomène passe avant la modélisation statistique.

L'analyse non paramétrique des interférences entre événements a été mise au point à l'INED par Éva Lelièvre, en collaboration avec Daniel Courgeau (Courgeau et Lelièvre, 1989). Cette technique s'inscrit dans le cadre plus général de l'analyse des processus markoviens. Son originalité vient de sa faible exigence du point de vue des hypothèses : le modèle non paramétrique n'impose aucune forme à la distribution de l'événement étudié.

Ce modèle permet de déceler une influence locale d'un événement sur un autre, sans que nécessairement le second exerce une influence sur le premier. Ainsi, on pourra déceler l'influence du premier mariage sur l'accès au premier emploi, et réciproquement, l'effet de l'entrée en activité sur la première union.

Une différence fondamentale du modèle non paramétrique, par rapport aux autres modèles d'analyse de survie, est qu'aucune formule de régression linéaire n'est utilisée pour mesurer l'interaction. On se soustrait ainsi à l'hypothèse de proportionnalité que l'on fait dans les modèles de régression logistique ou à risques proportionnels.

À l'instant t, on observe $N_0(t)$ femmes présentes à Dakar à l'âge de 12 ans (ces "Dakaroises" sont au nombre de 228 dans notre échantillon). Posons $h_{0,1}(t)$, le quotient instantané d'entrée pour le premier emploi, c'est-à-dire une mesure du risque d'entrée en activité. Le quotient instantané correspondant au premier mariage est $h_{0,2}(t)$. Pour les femmes ayant déjà travaillé, le quotient $h_{1,2}(t/u)$ représente la probabilité qu'elles ont de se marier en ayant déjà été actives. De même, le quotient $h_{2,1}(t/u)$ représente la probabilité d'entrer en activité après avoir quitté le célibat. Le schéma suivant formalise l'analyse :



Pour voir si la probabilité d'entrée dans la vie active diffère avant et après l'entrée en union, nous testons l'égalité :

$$h_{0,1}(t) = h_{2,1}(t/u)$$

La différence entre les quotients est entièrement attribuée au premier mariage. Il en est de même dans l'autre sens : pour voir si entrée dans la vie professionnelle influe sur la probabilité de se marier, nous testons :

$$h_{0,2}(t) = h_{1,2}(t/u)$$

L'intervalle de temps t choisi est la durée écoulée depuis l'anniversaire des 12 ans, comptée en nombre de mois.

c) Le traitement des simultanéités

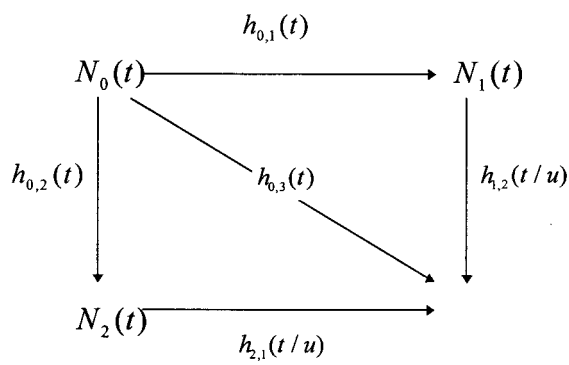
Le modèle non paramétrique est estimé à l'aide de la commande « nonpar », dont le programme figure sur la disquette accompagnant ce manuel. Il est conseillé d'utiliser d'abord « nonpar » avec les options « graph » et « summary » pour vérifier rapidement la cohérence des données, avant de passer aux tables de séjour qui peuvent être assez volumineuses.

La seule difficulté pour l'estimation du modèle non paramétrique réside dans le traitement des événements qui ont eu lieu au cours du même intervalle de temps. On parle alors de **simultanéités**. Leur traitement doit faire l'objet d'une attention particulière car il peut influencer les conclusions de l'analyse. Il s'agit en fait d'une difficulté moins liée à la conception du modèle qu'au recueil des données et à la conceptualisation du temps avant l'événement.

Sur la base des expériences passées, l'équipe IFAN-ORSTOM a conçu son enquête dans la perspective d'une analyse des interférences des événements professionnels, matrimoniaux et migratoires. L'ordre de succession de ces événements a pu être établi grâce à un système original de codification pour les événements se produisant la même année (annexe 3).

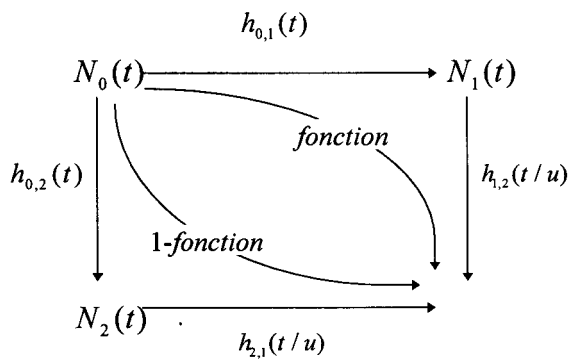
Lorsqu'on travaille au mois près, on n'observe que 12 cas de simultanéités (5,2 % des 228 femmes non immigrées) du premier emploi et du premier mariage. À l'année près, le nombre de simultanéités passe à 14 (6,1 %). C'est peu, mais néanmoins, il est important de passer en revue toutes les options possibles pour le traitement des simultanéités qu'offre la commande « nonpar », car elles peuvent sensiblement changer les conclusions de l'analyse :

- « **nosim** » : le traitement le plus simple consiste à ne pas tenir compte des individus qui ont expérimenté les deux événements simultanément. Cela signifie que l'échantillon restant a connu les deux événements avec un décalage d'au moins une unité de temps ;
- « **apart1** » : on peut avec cette option prendre en compte les individus qui ont connu les deux événements simultanément, mais en calculant un quotient séparé pour le passage direct du stade initial (les événements n'ont pas encore eu lieu) au stade final (les deux événements ont eu lieu durant le même intervalle de temps). On calculera en tout cinq séries de quotients :

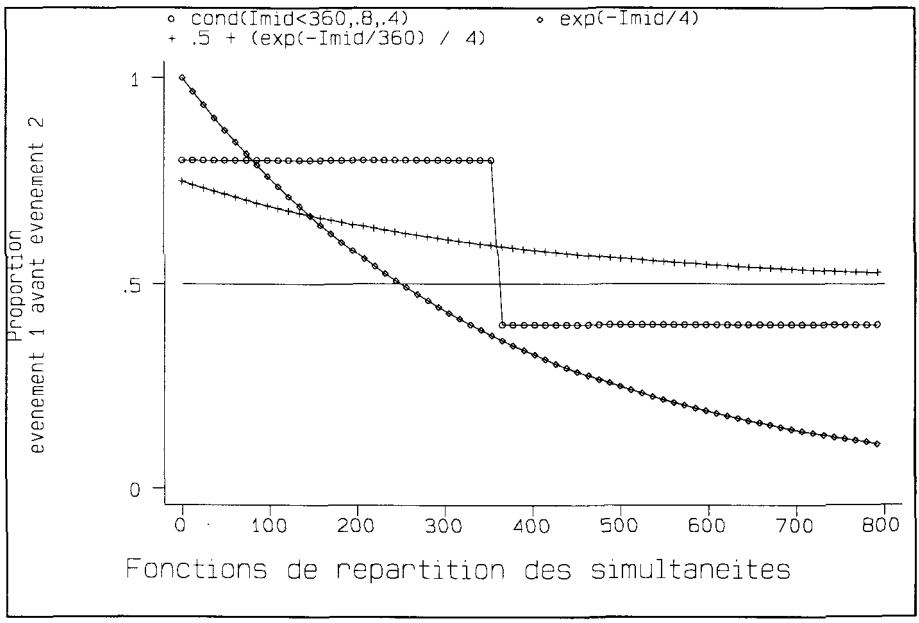


- « **dist**(*fonction*) » : les événements simultanés peuvent être distribués entre les deux chemins possibles : une partie des individus concernés sera

comptée parmi les individus qui ont connu l'événement 1 avant l'événement 2, et le complément sera compté parmi les individus qui ont connu l'événement 2 avant l'événement 1. Il s'agit alors de choisir la bonne répartition. L'option « dist() » offre la possibilité de fixer une répartition pour l'ensemble des intervalles de temps, ou bien de choisir une répartition variant en fonction du temps d'exposition (voir plus loin les applications concrètes).



Dans le premier cas, le paramètre fonction est un chiffre : il doit prendre une valeur entre 0 et 1, qui représente la proportion de simultanéités pour lesquelles l'événement 1 est considéré antérieur à l'événement 2 (dans l'ordre d'apparition de la liste de variables de la commande « nonpar ») : par exemple « dist(0,5) ». On peut aussi utiliser une fonction qui dépendra de la durée d'observation. Dans ce cas, on précise « dist(function) » (où "function" est écrit en toutes lettres), et le programme demande de préciser la forme de cette fonction en utilisant le terme « Imid » pour indiquer le milieu de l'intervalle. Par exemple, on peut préciser « cond(Imid<360,.8,.4) » pour indiquer que 80 % des simultanéités seront considérées comme l'événement 1 avant l'événement 2 pour les durées inférieures à 360 mois et 40 % pour les durées supérieures ou égale à 360 mois. Des fonctions plus complexes peuvent être introduites telles que exp(-Imid/4) ; ou 0,5 + (exp(-Imid/360)/4), qui sont représentées sur le graphique suivant, où l'échelle est donnée en nombre de mois :



- **« apart2 », « with() »** : ces options de traitement des simultanéités sont plus rarement utilisées. Le lecteur qui voudrait s'en enquérir consultera le menu « help nonpar » pour plus d'information.

d) La création des indices de troncature

L'interaction du premier emploi et du premier mariage sera utilisée comme exemple pour la création des indices de troncature. Le traitement des troncatures est différent de ce qu'on a vu jusqu'à présent : la troncature a lieu lorsque les deux événements ont eu lieu ou bien lorsque vient le moment de l'enquête (ou de l'émigration) sans qu'un (ou aucun) des événements n'ait eu lieu.

Concrètement, on doit d'abord créer des indices de troncature classiques. Pour le premier emploi et le premier mariage des Dakaroises, on aura :

```
. use jobmar
. censor no durev =v611==2 | v611==3 if immigree==0, gen(job1) before(emig==1)
. censor no durev =statut==1 if immigree==0, gen(mar1) before(emig==1)
```

Le fichier et les indices de troncature pour chaque événement se présentent comme suit :

```
. by no : list durev immigr emig job1 mar1
(extrait du listing)
-> no=       3
        durev  immigree       emig       job1       mar1
  1.      203          0          0          0          1
  2.      532          0          0          1          .
  3.      593          0          0          .          .

-> no=       4
        durev  immigree       emig       job1       mar1
  4.      378          0          0          0          0

-> no=      11
        durev  immigree       emig       job1       mar1
  5.      192          1          0          .          .
  6.      372          1          0          .          .
  7.      413          1          0          .          .

-> no=      13
        durev  immigree       emig       job1       mar1
  8.      212          0          0          0          1
  9.      226          0          0          0          .
 10.      384          0          0          1          .
 11.      425          0          0          .          .
(etc)
-> no=     113
        durev  immigree       emig       job1       mar1
124.      226          0          0          0          1
125.      257          0          0          0          .
126.      288          0          1          .          .
127.      298          0          1          .          .
128.      353          0          1          .          .

-> no=     114
        durev  immigree       emig       job1       mar1
129.      166          1          0          .          .
130.      451          1          0          .          .
131.      456          1          0          .          .
132.      581          1          0          .          .

-> no=     116
        durev  immigree       emig       job1       mar1
133.      215          0          0          0          1
134.      264          0          0          0          .
135.      274          0          0          0          .
136.      374          0          0          1          .
137.      644          0          0          .          .
138.      665          0          0          .          .
(etc)
```

On voit par exemple que pour l'individu n°3 le mariage a eu lieu avant le premier emploi. L'individu n°113 n'a connu que le mariage. L'individu n°4 n'a connu aucun des événements au moment de l'enquête.

La commande « nonpar » et les autres commandes d'analyse des données biographiques font un traitement différent des troncatures. Dans l'analyse des interactions entre deux événements biographiques, l'individu est toujours soumis au risque tant que les deux événements ou la date d'enquête (voire l'émigration) n'ont pas eu lieu. En conséquence, pour la commande « nonpar », les indices de troncature seront codés 1 après l'occurrence du premier événement, jusqu'à l'occurrence de l'autre événement (ou de la troncature par enquête ou émigration). On recalcule simplement ces indices en faisant la somme cumulée des indices de

troncature classiques, en prenant bien soin de ne pas tenir compte des immigrées et des périodes postérieures à une émigration :

```
. quietly by no : gen byte emigc=sum(emig) if immigr==0
. quietly by no : gen byte job1c=sum(job1) if immigr==0 & emigc==0
. lab var job1c "Indice troncature job1 cumule"
. quietly by no : gen byte mar1c=sum(mar1) if immigr==0 & emigc==0
. lab var mar1c "Indice troncature mar1 cumule"
. by no : list durev immigr emigc job1 job1c mar1 mar1c, noo nod
-> no=       3
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     203         0          0          0          0          1          1
     532         0          0          1          1          .          1
     593         0          0          .          1          .          1
-> no=       4
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     378         0          0          0          0          0          0
-> no=      11
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     192         1          .          .          .          .          .
     372         1          .          .          .          .          .
     413         1          .          .          .          .          .
-> no=      13
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     212         0          0          0          0          1          1
     226         0          0          0          0          .          1
     384         0          0          1          1          .          1
     425         0          0          .          1          .          1
(etc)
-> no=     113
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     226         0          0          0          0          1          1
     257         0          0          0          0          .          1
     288         0          1          .          .          .          .
     298         0          2          .          .          .          .
     353         0          3          .          .          .          .
-> no=     114
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     166         1          .          .          .          .          .
     451         1          .          .          .          .          .
     456         1          .          .          .          .          .
     581         1          .          .          .          .          .
-> no=     116
   durev  immigree      emigc       job1      job1c       mar1      mar1c
     215         0          0          0          0          1          1
     264         0          0          0          0          .          1
     274         0          0          0          0          .          1
     374         0          0          1          1          .          1
     644         0          0          .          1          .          1
     665         0          0          .          1          .          1
(etc)
```

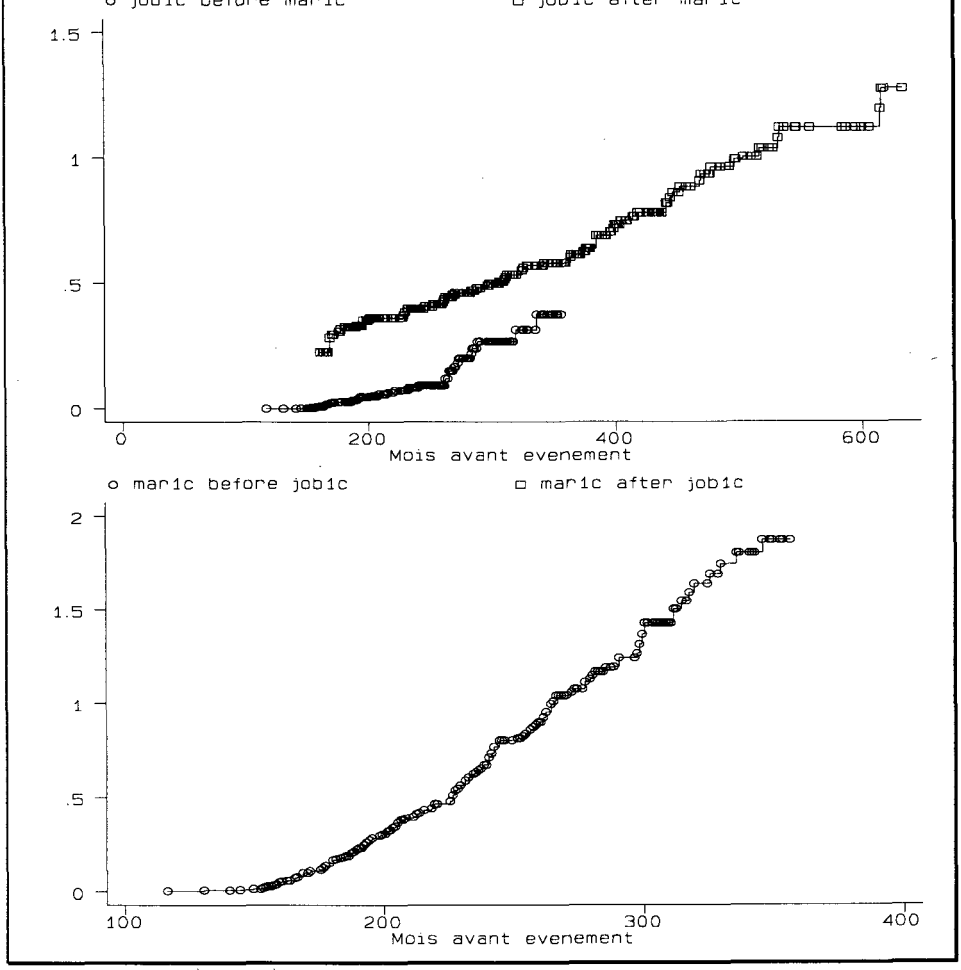
e) Le diagnostic graphique : l'option « graph » de la commande « nonpar »

Le diagnostic graphique repose sur les mêmes principes que les courbes de Aalen auxquelles le chapitre précédent consacrait une partie. Le niveau des courbes

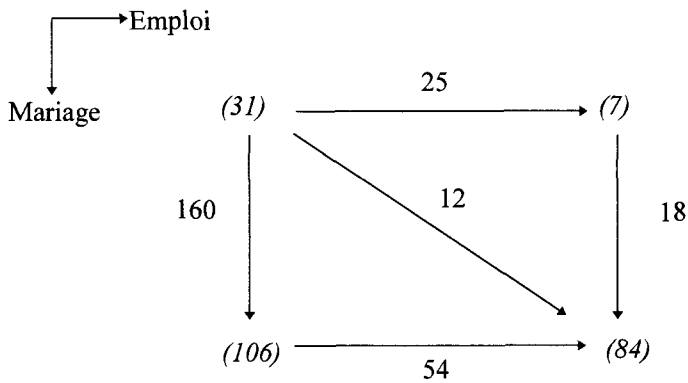
n'est pas pertinent, mais on peut comparer leur pente. Un risque plus important de connaître l'événement est repéré par une pente plus abrupte.

L'option « graph » produit deux graphiques pour représenter les quotients cumulés pour chaque événement, avant et après occurrence de l'autre événement, soit en principe deux courbes pour chaque graphique. L'option « summarize » fournit en outre un résumé des événements et des troncatures dans l'échantillon :

```
. nonpar durev job1c mar1c, tvid(no) dist(.5) graph summ
(warning:  job1c or mar1c have missing values; 966 obs not used)
Sample :     228
       Number of events job1c before mar1c :      25
                         job1c after mar1c :      54
       Number of events mar1c before job1c :     160
                         mar1c after job1c :      18
       Number of simultaneous events mar1c and job1c :      12
       Truncations :      31 before job1c and mar1c
                           7 after job1c
                         106 after mar1c
                          84 after job1c and mar1c
```



Ces graphiques seront mieux compris à l'aide du tableau qui les précède. Ce tableau indique la taille de l'échantillon, le nombre d'événements et de troncatures de chaque type. Pour souligner le rapport entre ces chiffres et les événements qu'ils représentent, on a reporté ces chiffres sur le diagramme des transitions présenté dans la section sur les principes de la description d'une interaction. Les chiffres en italiques et entre parenthèses représentent les troncatures :

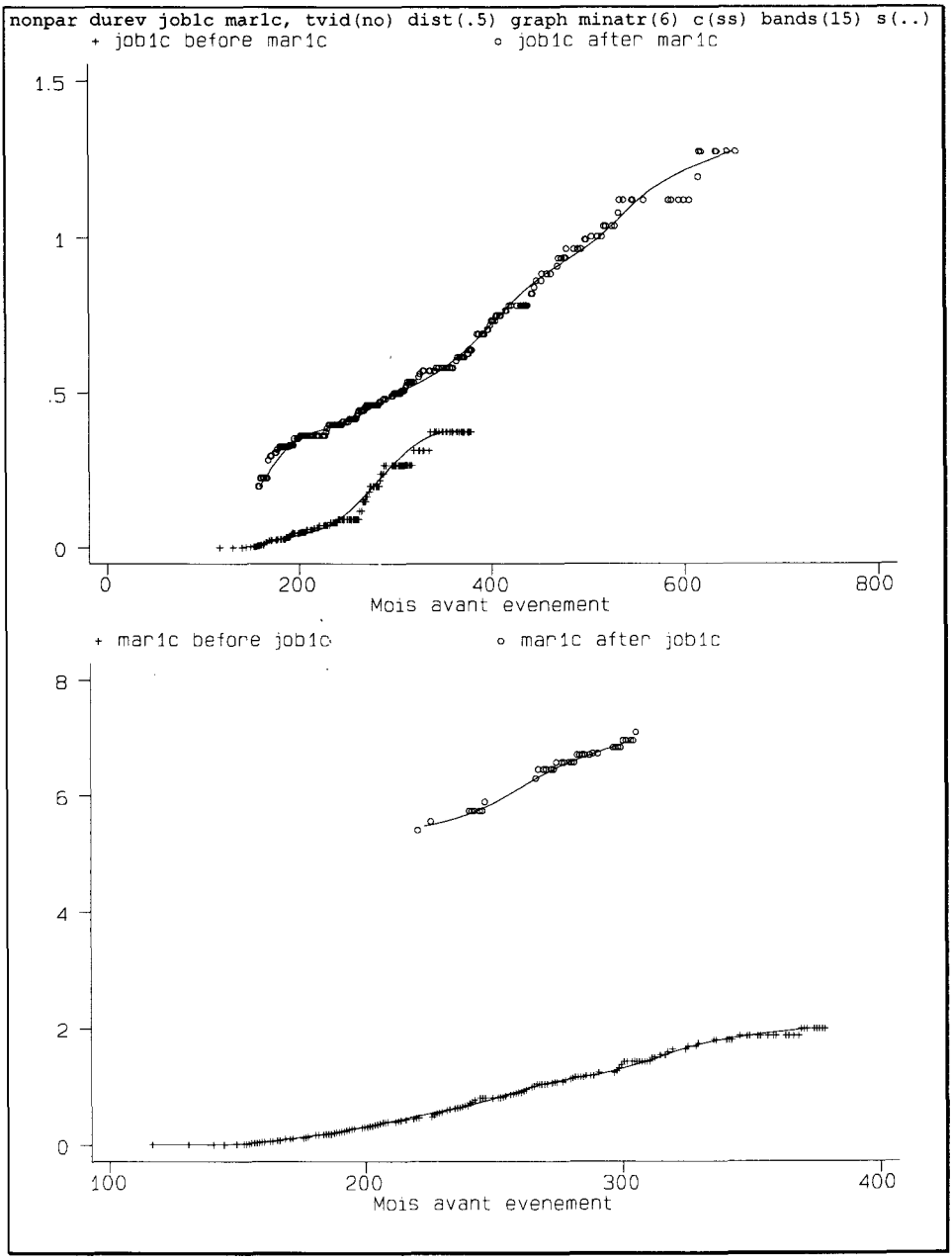


On voit très nettement à la lecture de ce diagramme que le mariage avant l'emploi et l'emploi après le mariage sont des événements bien plus fréquents dans l'échantillon que l'emploi avant le mariage et le mariage après l'emploi.

La taille de l'échantillon peut poser des problèmes pour le calcul de certains risques. Par défaut, la commande « nonpar » prévoit un effectif minimum de 10 individus soumis au risque. Ainsi, la courbe des quotients cumulés pour le risque d'obtenir un emploi après le mariage ne débute qu'à la durée 160 (ce qui correspond à l'âge de 25 ans et 4 mois dans notre fichier). De même, dans le deuxième graphique, la courbe des quotients cumulés pour le risque de se marier après un emploi (*marlc after joblc*) n'est pas représentée, parce que les effectifs de femmes soumises au risque étaient inférieurs à 10 : on ne peut donc comparer la probabilité de se marier avant et après un emploi sur ce graphique.

Pour comparer chacun des risques (avant et après occurrence de l'événement concurrent), il s'agit de considérer, comme on le fait pour les courbes de Aalen, la pente des courbes et non leur niveau. La commande « nonpar » produit des courbes en escalier, ce qui n'est pas toujours très lisible, mais on peut effectuer un lissage simple à l'aide des options graphiques « c(ss) » (*cubic spline*) et « bands(15) » (la section [3] connect du manuel de STATA pour de plus amples explications sur cette technique de lissage). De plus, le seuil de 10 individus soumis au risque peut être abaissé (bien que cela ne soit pas très conseillé du point de vue du calcul des tests statistiques) avec l'option « minatr(#) » (comme

"*minimum at risk*") pour faire apparaître la courbe des quotients cumulés pour le risque de se marier après un emploi. On aura donc :



Dans le premier graphique on voit très nettement que la probabilité d'obtenir un emploi est plus élevée après qu'avant le mariage jusqu'à la durée 200 environ (avant 29 ans), mais à partir de la durée 250 environ (33 ans) c'est l'inverse qui se produit. À partir de 40 ans, les courbes n'ont plus beaucoup d'intérêt puisqu'il ne reste qu'une minorité de femmes non mariées.

Dans le deuxième graphique, c'est justement la probabilité de se marier qui est calculée. Cette fois, puisqu'on a abaissé le seuil de la population soumise au risque à 6 (au lieu de 10 par défaut), le deuxième graphique fait état de la courbe des quotients cumulés entre la durée 228 (31 ans) et la durée 312 (38 ans) pour la probabilité de se marier après avoir obtenu un emploi. Ainsi, entre 31 et 38 ans (c'est-à-dire à des âges au premier mariage très élevés pour les femmes sahéliennes), la probabilité de se marier semble légèrement plus élevée pour les femmes ayant déjà travaillé que pour les femmes restées toujours inactives. Mais, comme on l'a dit plus haut, les effectifs soumis au risque sont très faibles.

EXERCICE 5

Les simultanéités sont au nombre de 12 dans le fichier des Dakaroises, et dans les exemples qui précèdent, l'option « dist(.5) » fixe leur répartition uniformément entre emploi avant mariage et mariage avant emploi. Une autre fonction peut être testée. Voyez les résultats de la fonction « exp(-Imid/360) » :

```
. nonpar durev joblc marlc, tvid(no) dist(function) graph minatr(6) c(ss)
bands(20) s(..)
(warning:  joblc or marlc have missing values; 966 obs not used)

Simultaneities will be distributed as a function of mid-interval time.
Please enter a function(Imid)
e.g. cond(Imid<4,.8,.4) , exp(-Imid/4) or .5+(exp(-Imid/4)/4) :. exp(-Imid/360)
```

Quelles différences voyez-vous entre les deux traitements des simultanéités ? Les résultats peuvent aussi être produits en considérant à part les simultanéités. Voyez les résultats du remplacement de l'option « dist() » par les options « nosim » ou « apart1 ».

La commande « nonpar » permet d'augmenter l'intervalle de temps avec l'option « interval(#) ». Voyez les résultats avec un intervalle de 12 mois :

```
. nonpar durev joblc marlc, tvid(no) dist(.5) int(12) graph minatr(6) c(ss)
bands(20) s(..) summ
```

Le nombre de simultanéités a-t-il beaucoup augmenté ? Les conclusions de l'analyse sont-elles différentes ?

f) Les tables de séjour

La commande « nonpar », sans l'option « graph » présente les effectifs et les quotients, ainsi que le résultat des tests sous la forme de tableaux :

```
. nonpar durev joblc marlc, tvid(no) dist(.5) minatr(6) int(12)
(warning:  joblc or marlc have missing values; 966 obs not used)
```

	joblc BEFORE marlc				joblc AFTER marlc				Hoem	
t,t+i	AtRisk	joblc	Rate	Cumul	AtRisk	joblc	Rate	Cumul	test	CumTest
[108-	227.50	0.00	0.0000	0.0000	0.50	0.00	-	-	-	0.000
[120-	226.50	0.00	0.0000	0.0000	1.50	0.00	-	-	-	0.000
[132-	225.50	0.00	0.0000	0.0000	2.00	0.00	-	-	-	0.000
[144-	222.00	0.50	0.0023	0.0023	4.25	0.50	-	-	-	-0.693
[156-	211.25	3.25	0.0154	0.0176	11.12	0.25	0.0225	0.1401	-0.155	-0.848
[168-	196.75	1.75	0.0089	0.0265	21.38	1.75	0.0819	0.2220	-1.172	-2.021
[180-	178.50	2.50	0.0140	0.0405	34.25	0.50	0.0146	0.2366	-0.026	-2.047
[192-	158.50	1.50	0.0095	0.0500	50.25	1.50	0.0299	0.2664	-0.797	-2.844
[204-	140.50	2.00	0.0142	0.0642	64.50	0.00	0.0000	0.2664	1.414	-1.430
[216-	126.00	1.00	0.0079	0.0722	77.00	0.00	0.0000	0.2664	1.000	-0.430
[228-	109.50	1.00	0.0091	0.0813	88.00	3.00	0.0341	0.3005	-1.150	-1.580
[240-	93.00	1.00	0.0108	0.0921	98.50	1.00	0.0102	0.3107	0.041	-1.540
[252-	78.75	2.25	0.0286	0.1206	107.12	3.25	0.0303	0.3410	-0.070	-1.609
[264-	64.25	4.25	0.0661	0.1868	111.62	2.25	0.0202	0.3612	1.322	-0.287
[276-	52.00	2.00	0.0385	0.2252	114.50	2.00	0.0175	0.3786	0.703	0.416
[288-	41.75	1.25	0.0299	0.2552	117.62	2.25	0.0191	0.3978	0.365	0.780
[300-	30.50	0.00	0.0000	0.2552	115.00	3.00	0.0261	0.4239	-1.732	-0.952
[312-	22.50	1.00	0.0444	0.2996	110.50	1.00	0.0090	0.4329	0.780	-0.172
[324-	18.50	0.00	0.0000	0.2996	106.50	4.00	0.0376	0.4705	-2.000	-2.172
[336-	15.50	1.00	0.0645	0.3641	101.50	1.00	0.0099	0.4803	0.838	-1.334
[348-	12.00	0.00	0.0000	0.3641	96.50	0.00	0.0000	0.4803	.	-1.334
[360-	9.00	0.00	0.0000	0.3641	89.50	3.00	0.0335	0.5138	-1.732	-3.066
[372-	7.00	0.00	0.0000	0.3641	82.50	2.00	0.0242	0.5381	-1.414	-4.480
[384-	5.50	1.00	-	-	74.50	5.00	0.0671	0.6052	-	-3.858
[396-	4.00	0.00	-	-	67.00	3.00	0.0448	0.6500	-	-5.590
[408-	3.00	0.00	-	-	61.50	2.00	0.0325	0.6825	-	-7.004
[420-	2.50	0.00	-	-	57.50	0.00	0.0000	0.6825	-	-7.004
[432-	2.00	0.00	-	-	52.50	2.00	0.0381	0.7206	-	-8.418
[444-	2.00	0.00	-	-	46.50	3.00	0.0645	0.7851	-	-10.150
[456-	2.00	0.00	-	-	42.50	0.00	0.0000	0.7851	-	-10.150
[468-	2.00	0.00	-	-	38.00	3.00	0.0789	0.8641	-	-11.882
[480-	2.00	0.00	-	-	34.00	0.00	0.0000	0.8641	-	-11.882
[492-	1.75	0.25	-	-	31.62	1.25	0.0395	0.9036	-	-11.523
[504-	0.50	0.00	-	-	29.50	0.00	0.0000	0.9036	-	-11.523
[516-	0.00	0.00	-	-	27.00	1.00	0.0370	0.9406	-	-11.523
[528-	0.00	0.00	-	-	23.50	2.00	0.0851	1.0257	-	-11.523
[540-	0.00	0.00	-	-	21.00	0.00	0.0000	1.0257	-	-11.523
[552-	0.00	0.00	-	-	19.50	0.00	0.0000	1.0257	-	-11.523
[576-	0.00	0.00	-	-	18.00	0.00	0.0000	1.0257	-	-11.523
[588-	0.00	0.00	-	-	16.00	0.00	0.0000	1.0257	-	-11.523
[600-	0.00	0.00	-	-	14.50	0.00	0.0000	1.0257	-	-11.523
[612-	0.00	0.00	-	-	12.50	2.00	0.1600	1.1857	-	-11.523
[624-	0.00	0.00	-	-	10.00	0.00	0.0000	1.1857	-	-11.523
[636-	0.00	0.00	-	-	8.50	0.00	0.0000	1.1857	-	-11.523
[648-	0.00	0.00	-	-	7.50	0.00	0.0000	1.1857	-	-11.523
[660-	0.00	0.00	-	-	6.00	0.00	-	-	-	-11.523
[672-	0.00	0.00	-	-	4.00	0.00	-	-	-	-11.523
[708-	0.00	0.00	-	-	2.00	0.00	-	-	-	-11.523
[792-	0.00	0.00	-	-	0.50	0.00	-	-	-	-11.523

	job1c BEFORE mar1c				job1c AFTER mar1c				Hoem	
t,t+i	AtRisk	mar1c	Rate	Cumul	AtRisk	mar1c	Rate	Cumul	test	CumTest
[108-	227.50	1.00	0.0044	0.0044	0.00	0.00	-	-	-	0.000
[120-	226.50	1.00	0.0044	0.0088	0.00	0.00	-	-	-	0.000
[132-	225.50	0.00	0.0000	0.0088	0.00	0.00	-	-	-	0.000
[144-	222.00	4.50	0.0203	0.0291	0.25	0.50	-	-	-	-0.700
[156-	211.25	10.25	0.0485	0.0776	1.12	0.25	-	-	-	-1.091
[168-	196.75	11.75	0.0597	0.1373	2.38	1.75	-	-	-	-2.306
[180-	178.50	16.50	0.0924	0.2298	3.25	0.50	-	-	-	-2.586
[192-	158.50	17.50	0.1104	0.3402	4.25	1.50	-	-	-	-3.424
[204-	140.50	13.00	0.0925	0.4327	5.00	0.00	-	-	-	0.181
[216-	126.00	13.00	0.1032	0.5359	5.50	2.00	-	-	-	-0.826
[228-	109.50	15.00	0.1370	0.6729	5.50	0.00	-	-	-	3.047
[240-	93.00	13.00	0.1398	0.8126	5.50	2.00	-	-	-	2.187
[252-	78.75	10.25	0.1302	0.9428	5.12	1.25	-	-	-	1.674
[264-	64.25	8.25	0.1284	1.0712	6.62	1.25	0.1887	4.6257	-0.345	1.329
[276-	52.00	6.00	0.1154	1.1866	8.50	1.00	0.1176	4.7434	-0.018	1.311
[288-	41.75	7.25	0.1737	1.3602	9.12	1.25	0.1370	4.8803	0.265	1.576
[300-	30.50	4.00	0.1311	1.4914	6.50	4.00	0.6154	5.4957	-1.539	0.036
[312-	22.50	3.00	0.1333	1.6247	4.50	0.00	-	-	-	1.769
[324-	18.50	3.00	0.1622	1.7869	5.00	0.00	-	-	-	3.501
[336-	15.50	1.00	0.0645	1.8514	4.50	2.00	-	-	-	2.316
[348-	12.00	0.00	0.0000	1.8514	3.50	0.00	-	-	-	2.316
[360-	9.00	1.00	0.1111	1.9625	3.00	0.00	-	-	-	3.316
[372-	7.00	0.00	0.0000	1.9625	2.50	1.00	-	-	-	2.316
[384-	5.50	0.00	-	-	2.50	0.00	-	-	-	2.316
[396-	4.00	0.00	-	-	2.50	0.00	-	-	-	2.316
[408-	3.00	0.00	-	-	1.50	0.00	-	-	-	2.316
[420-	2.50	1.00	-	-	1.00	0.00	-	-	-	3.316
[432-	2.00	0.00	-	-	0.50	0.00	-	-	-	3.316
[444-	2.00	0.00	-	-	0.00	0.00	-	-	-	3.316
[456-	2.00	0.00	-	-	0.00	0.00	-	-	-	3.316
[468-	2.00	0.00	-	-	0.00	0.00	-	-	-	3.316
[480-	2.00	0.00	-	-	0.00	0.00	-	-	-	3.316
[492-	1.75	0.25	-	-	0.12	0.25	-	-	-	2.853
[504-	0.50	1.00	-	-	0.00	0.00	-	-	-	2.853
[516-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[528-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[540-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[552-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[576-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[588-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[600-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[612-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[624-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[636-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[648-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[660-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[672-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[708-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853
[792-	0.00	0.00	-	-	0.00	0.00	-	-	-	2.853

Étant donné que les simultanéités selon l'année sont en nombre relativement faible quand l'année est utilisée comme unité de temps, on a produit ces tableaux par année avec l'option « interval(12) » pour établir les statistiques sur une base annuelle plutôt que mensuelle. Les tableaux sont en effet de taille plus raisonnable pour des intervalles de temps annuels.

Des tests existent pour comparer les courbes deux à deux sur toute la durée de soumission au risque, comme on le fait par exemple pour deux courbes de Kaplan-Meier. Cependant, étant donné les fortes variations de calendrier que chaque type d'événement peut connaître, il est préférable d'utiliser des tests localisés sur des durées précises. C'est ce que permet de faire le test de Hoem sur l'écart entre deux quotients, même avec de petits effectifs soumis au risque. Le test permet aussi, lorsqu'il est cumulé, de mesurer le degré de significativité de l'écart entre deux courbes de quotients cumulés.

Dans notre exemple, la différence mesurée par le test cumulé semble être significative (au seuil de 10 %, valeur du test supérieure à 1,63) entre les durées 168 et 192, de même qu'à la durée 324. Cependant, ces tests cumulés tiennent compte des quotients calculés avant la durée 156 sur des effectifs soumis au risque inférieurs à 10 pour l'emploi après le mariage. Dès lors, il faut retrancher la valeur du test cumulé jusqu'à la durée 156 (soit -0,693) : on aboutit à un test cumulé montrant une tendance significative plus élevée d'obtenir un emploi après le mariage seulement à la durée 192 (soit à 28 ans), avec un test égal à -2,151 (soit -2,844+0,693).

On peut aussi choisir de faire les calculs des quotients cumulés sur un groupe d'âges plus réduit. Par exemple, d'après le premier graphique, il semble que la probabilité d'obtenir un emploi soit plus élevée (la courbe étant plus abrupte) à partir de la durée 204 environ. En effet, à partir de ce moment, on aboutit à la durée 216 (30 ans) à un test cumulé de 1,414+1,000=2,414 significatif à 5 %.

Dans le deuxième tableau, malgré l'abaissement à 6 du nombre minimum d'individus soumis au risque, les tests, cumulés ou non, ne sont pas significatifs.

EXERCICE 6

Calculez le test cumulé à partir de la durée 300 pour la probabilité d'obtenir un emploi. La pente est-elle significativement différente entre les deux courbes ?

Le programme « nonpar » donne aussi la possibilité de représenter graphiquement les quotients non cumulés. Interprétez le résultat de la commande suivante :

```
. nonpar durev jobic maric, tvid(no) dist(.5) int(12) minatr(6) graph hazard
c(ss) bands(20)
```

6. L'évaluation semi-paramétrique d'une interaction entre événements : la commande « cox » avec variables indépendantes fonction du temps

L'effet du mariage sur l'entrée en activité va être évalué de la même façon que l'effet du changement de statut d'activité. Le passage du statut de célibataire à celui de mariée est considéré au même titre que le passage des études au chômage ou à l'apprentissage, c'est-à-dire comme une variable fonction du temps. Rappelons à nouveau que le test graphique de l'hypothèse de proportionnalité n'est pas autorisé pour de telles variables explicatives fonction du temps.

Du point de vue pratique, il suffit de créer une variable indiquant si l'enquêtée a déjà été mariée :

```
. gen byte djmarie=statut!=0 if statut!=.
```

Le lecteur se demandera sans doute comment sont traités les événements simultanés, par exemple lorsque le mariage a lieu en même temps qu'un changement d'activité. Dans le modèle semi-paramétrique estimé ici, la période matrimoniale prise en compte est celle qui précède juste l'événement : la femme sera donc considérée comme célibataire au moment de l'accès à l'emploi, c'est-à-dire que ces simultanéités sont donc toutes considérées comme "emploi avant mariage". On suppose dans ce cas que le mariage n'a pas eu le temps de faire sentir ses effets sur les chances d'obtenir un emploi. Si l'on estime le contraire, il faut alors créer artificiellement un décalage d'au moins une unité de temps entre les deux événements pour placer le mariage avant l'emploi.

La nouvelle variable indépendante est introduite dans le modèle :

```
. cox Tjob1 djmarie, dead(job1) tvid(no)

Iteration 0:  Log Likelihood =-435.99761
Iteration 1:  Log Likelihood =-435.66569
Iteration 2:  Log Likelihood =-435.66547

Cox regression                                        Number of obs =     581
                                                      chi2(1)       =    0.66
                                                      Prob > chi2   =  0.4151
Log Likelihood = -435.66547                           Pseudo R2     =  0.0008

------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 djmarie |  -.2651063   .3228533     -0.821   0.412      -.8978871    .3676745
------------------------------------------------------------------------------
```

L'effet du mariage seul sur l'accès à l'emploi ne semble pas significatif. En contrôlant pour le groupe de générations, on obtient le résultat suivant :

```
. cox Tjob1 gener2 gener3 djmarie, dead(job1) tvid(no)

Iteration 0:  Log Likelihood =-435.99761
(etc.)
Iteration 3:  Log Likelihood =-432.38669

Cox regression                                        Number of obs =     581
                                                      chi2(3)       =    7.22
                                                      Prob > chi2   =  0.0652
Log Likelihood = -432.38669                           Pseudo R2     =  0.0083
------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .6009767   .2699555     2.226   0.026        .0718736    1.13008
  gener3 |   .0690105   .3261595     0.212   0.832       -.5702504    .7082715
 djmarie |   -.363997   .3450322    -1.055   0.291       -1.040248    .3122538
------------------------------------------------------------------------------
```

L'effet du mariage n'est toujours pas significatif. La variable "statut" permet d'être plus précis quant à la définition du statut matrimonial, car elle représente non seulement le statut matrimonial courant, mais elle résume les statuts matrimoniaux antérieurs. Le code "1" est réservé au statut de mariée, le code "2" au statut de divorcée, le code "3" au statut de veuve. On forme ainsi un code où les unités représentent le premier mariage, les dizaines le deuxième mariage, les centaines le troisième mariage, et ainsi de suite. Par exemple, le code "23" (label "div2veu1") signifie "divorcée d'un deuxième mariage après veuvage du premier mariage" :

```
. tab statut
      statut|
 matrimonial|      Freq.     Percent        Cum.
------------+-----------------------------------
    celibat |        659       37.74       37.74
   mariage1 |        653       37.40       75.14
   div mar1 |        132        7.56       82.70
   veu mar1 |         64        3.67       86.37
   mar2div1 |        115        6.59       92.96
   mar2veu1 |         28        1.60       94.56
   div2div1 |         25        1.43       95.99
   div2veu1 |          9        0.52       96.51
   veu2div1 |          8        0.46       96.96
   veu2veu1 |          2        0.11       97.08
   mar3d2d1 |         14        0.80       97.88
   mar3d2v1 |          5        0.29       98.17
   mar3v2d1 |          4        0.23       98.40
   mar3v2v1 |          2        0.11       98.51
   div3d2d1 |          7        0.40       98.91
   div3d2v1 |          4        0.23       99.14
   div3v2d1 |          3        0.17       99.31
   m4d3d2d1 |          5        0.29       99.60
   d4d3d2d1 |          4        0.23       99.83
   m5dv4321 |          3        0.17      100.00
------------+-----------------------------------
      Total |       1746      100.00
```

On peut ainsi prendre en compte les périodes de mariage effectif (que ce soit un premier mariage ou un remariage) en excluant les périodes de divorce ou de veuvage :

```
. gen marie=statut

. recode marie 1 12 13 122 123 132 133 1222 12222=1 *=0
(434 changes made)

. cox Tjob1 gener2 gener3 marie, dead(job1) tvid(no)

Iteration 0:  Log Likelihood =-435.99761
(etc.)
Iteration 3:  Log Likelihood =-430.96527

Cox regression                                        Number of obs =     581
                                                      chi2(3)       =   10.06
                                                      Prob > chi2   =  0.0180
Log Likelihood = -430.96527                           Pseudo R2     =  0.0115

------------------------------------------------------------------------------
       Tjob1 |
        job1 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gener2  |   .6158649   .2693446     2.287   0.022      .0879593    1.143771
     gener3  |   .0517234   .3171527     0.163   0.870     -.5698845    .6733313
      marie  |  -.5332555   .2639491    -2.020   0.043     -1.050586   -.0159247
------------------------------------------------------------------------------
```

Cette fois-ci, l'effet du statut de mariée sur l'entrée en activité devient très significatif. Outre le statut matrimonial (variable indépendante exogène), on peut faire intervenir les variables endogènes telles que le statut d'activité avant le premier emploi :

```
. cox Tjob1 gener2 gener3 marie chomet chomaut apprent inact, dead(job1) tvid(no)

Iteration 0:  Log Likelihood =-435.99761
(etc)
Iteration 4:  Log Likelihood =-412.67498

Cox regression                                        Number of obs =     581
                                                      chi2(7)       =   46.65
                                                      Prob > chi2   =  0.0000
Log Likelihood = -412.67498                           Pseudo R2     =  0.0535

------------------------------------------------------------------------------
       Tjob1 |
        job1 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gener2  |   .4499737    .274985     1.636   0.102     -.0889869    .9889344
     gener3  |  -.4698926   .3343646    -1.405   0.160     -1.125235     .18545
      marie  |  -.2513087   .2801794    -0.897   0.370     -.8004503    .2978329
     chomet  |   .4977143   .5214266     0.955   0.340      -.524263    1.519692
    chomaut  |  -.5190265   1.045282    -0.497   0.620     -2.567741    1.529688
    apprent  |   .7070444   .4648793     1.521   0.128     -.2041024    1.618191
    inactiv  |  -1.381418    .289426    -4.773   0.000     -1.948683   -.8141535
------------------------------------------------------------------------------
```

On remarque que l'effet du statut de mariée redevient non significatif : en fait, il semble que ce soit moins le statut de mariée en lui-même que le fait d'être femme au foyer qui freine l'accès au premier emploi. Lorsqu'on ne fait pas intervenir le statut d'activité avant le premier emploi, le statut de mariée ne joue que parce qu'il est fortement corrélé avec le fait d'être femme au foyer.

EXERCICE 7

La variable "statut" permet de distinguer les femmes selon le statut matrimonial courant ainsi que selon leur histoire matrimoniale : par exemple, une femme dans son troisième mariage peut avoir le code 132 correspondant à une période de mariage (1) qui suit une période de veuvage après un second mariage (3), lui-même consécutif à un premier mariage s'étant terminé par un divorce (2). Construisez deux variables supplémentaires pour les statuts de femme divorcée et veuve, et comparez les modèles obtenus :

```
. cox Tjob1 gener2 gener3 marie divorcee veuve, dead(job1) tvid(no)

. cox Tjob1 gener2 gener3 chomet chomaut apprent inact, dead(job1) tvid(no)

. cox Tjob1 gener2 gener3 marie divorcee veuve chomet chomaut apprent inact,
dead(job1) tvid(no)
```

Le statut matrimonial et le statut d'activité agissent-ils indépendamment l'un de l'autre sur la probabilité d'obtenir un emploi ? Les statuts de veuve, de mariée et de célibataire sont-ils significativement différents l'un de l'autre ? Qu'en déduisez-vous sur l'effet de la variable "djmarie" utilisée précédemment ?

De la même façon qu'on a testé l'influence du changement matrimonial sur l'accès à l'emploi, on peut tester l'influence du changement d'activité sur l'entrée en première union. Après avoir créé les variables supplémentaires pour les statuts de salariée et d'indépendante, voyez le résultat du modèle :

```
. gen byte salarie=v611==2

. gen byte independ=v611==3

. cox Tmar1 gener2 gener3 chomet chomaut apprent inact salarie independ,
dead(mar1) tvid(no)
```

Dans ce modèle, l'activité est-elle une variable endogène ou exogène ? Quelle est la catégorie de référence pour l'activité ? Comment sont traitées les simultanéités entre premier emploi et premier mariage ? Comment l'émigration et le statut d'immigrée sont-ils pris en compte ? Les conclusions sur les influences réciproques du mariage et de l'emploi sont-elles différentes de celles obtenues à partir de l'analyse non paramétrique ?

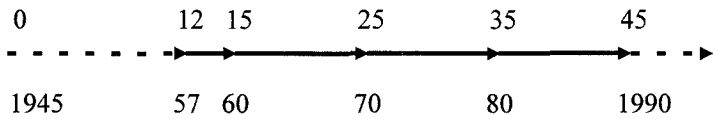
7. Âges de la vie et effet de périodes

Habituellement, l'effet de la période de conjoncture est introduit indirectement par l'effet de génération : on sait que chaque génération a connu des contextes différents, et que par conséquent, au même âge, leurs conditions de vie ne sont pas les mêmes. En termes statistiques, on parlera de processus dit "non stationnaire", c'est-à-dire non constant au cours du temps.

Mais puisqu'on cherche à repérer les effets de l'environnement économique, de l'état du marché de l'emploi etc., bref, d'un processus macro-économique, sur l'insertion et la mobilité professionnelles, n'est-il pas possible de créer une variable qui nous permette de qualifier directement les périodes traversées par chacune des générations ?

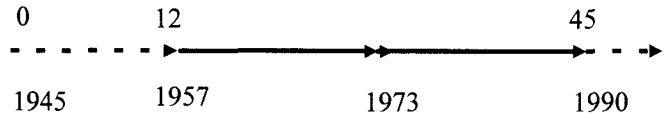
a) Conceptualisation du problème

Au lieu d'utiliser la génération, une caractéristique acquise au moment de la naissance et dont l'effet est supposé constant tout au long de la vie, il est plus précis de créer une variable indépendante fonction du temps qui indique, pour chaque individu, les périodes qu'il a traversées. On saura qu'un enquêté né en 1945, par exemple, a connu depuis l'âge de 12 ans, les périodes 1957-1959, 1960-1969, etc., jusqu'au moment de l'enquête en 1989. On créera donc une variable indicatrice pour chaque période traversée : pour un enquêté né en 1945, l'indicateur de la période 1957-1959 sera activé à 12, 13 et 14 ans, l'indicateur de la période suivante de 15 à 24 ans, etc.



Pour qualifier chacune de ces périodes, il n'est pas nécessaire d'utiliser des décennies, comme dans l'exemple ci-dessus. L'intérêt d'une telle variable est justement de permettre de mesurer l'effet de périodes particulières, choisies par exemple en fonction de la situation économique ou politique.

On peut à la limite tester l'effet d'une année donnée, s'il s'est produit un événement remarquable durant cette année (par exemple le choc pétrolier de 1973) :



Si la taille de l'échantillon et les risques d'imprécision dans la datation des événements ne permettent généralement pas une telle précision dans les enquêtes rétrospectives, on peut penser qu'à l'aide de données plus précises (par exemple, sur les chômeurs dans les agences pour l'emploi), on pourrait aboutir à une plus grande finesse. Tout dépend de l'adéquation entre les données et le phénomène étudié.

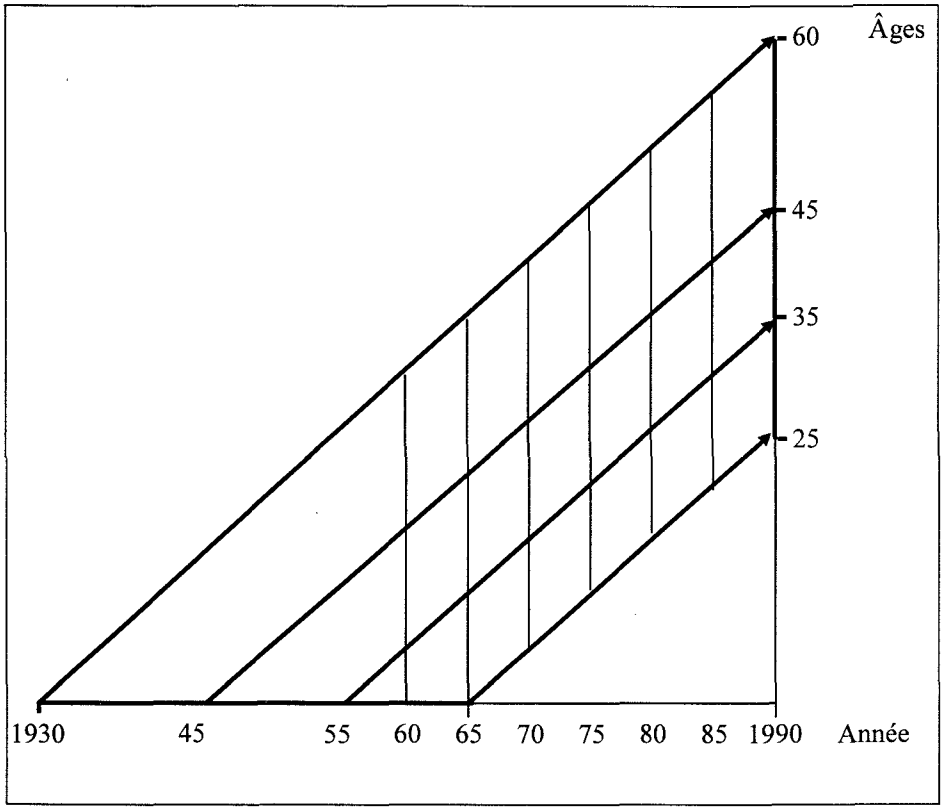


Figure 3. Diagramme de Lexis figurant les trois générations de l'enquête IFAN-ORSTOM et les périodes quinquennales de 1960 à 1990

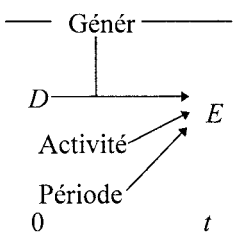
L'effet de la génération ne doit cependant pas être négligé. L'effet de période ne s'y substitue pas nécessairement. En fait, les deux peuvent agir en toute indépendance. Sur le diagramme de Lexis de la figure 3, sont reportées les surfaces couvertes par le résultat du croisement entre générations et périodes quinquennales pour l'enquête IFAN-ORSTOM.

Dans un même modèle, on peut tester chacun de ces effets à l'aide d'une variable groupe de générations et d'une variable indicatrice de la période. Il est possible que les deux variables agissent en interaction. Par exemple, l'effet des années 1975-79 n'est peut-être pas le même pour les générations 1945-54 qui avaient en moyenne 27 ans et demi pendant cette période, et pour les générations 1955-64 qui avaient en moyenne 17 ans et demi. De même, l'effet de l'âge peut différer selon qu'il concerne des générations qui traversent les années 1970-74 ou 1975-79.

L'interrelation entre les dimensions de l'âge, de la génération et de la période est un problème classique d'analyse descriptive en démographie, que l'on traite généralement avec les tables d'extinction (ou de mortalité). Il est en revanche peu souvent traité en analyse probabiliste. La commande « slice » et le traitement par interaction de la génération et de la période permettent d'introduire d'une façon originale les principes de l'analyse descriptive dans un modèle probabiliste.

b) La commande « slice » et la définition de l'indicateur de période

Pour prendre en compte l'effet des périodes ou des groupes d'âges traversés par les enquêtés, il suffit de créer une nouvelle variable qui indique lorsque l'individu change de période ou de groupe d'âges, et de l'ajouter aux autres variables indépendantes dans le modèle.



Dans le schéma causal ci-contre, "Période" est une variable indépendante fonction du temps qui indique les périodes successives traversées par l'individu jusqu'à l'événement ou la troncature. Selon ce modèle, elle agit indépendamment de la génération ("Génér."), et elle est du même type que la variable indiquant les périodes d'activité successives ("Activité").

Pour construire un fichier qui marque les transitions d'une période à l'autre, on doit disposer d'une variable de temps avec l'unité qui convient, et s'en servir pour "couper le temps en tranches" (d'où le terme anglais *slice*). Chaque passage d'une période à l'autre sera considéré comme un événement. Dans

STATA, cela veut dire qu'il faudra créer autant de lignes d'observation que de passages d'une période à l'autre. On n'aura plus seulement les périodes telles qu'elles ont été recueillies sur le terrain (périodes d'études, de chômage, etc.) mais aussi toutes les périodes ou tous les groupes d'âges traversés par l'individu. Pour créer une variable indicatrice de la période, on peut s'aider du programme « slice » qui se trouve sur la disquette accompagnant ce manuel.

Dans le fichier "job.dta", la variable "v638" qui définit l'année de fin de période d'activité aurait suffi pour découper le temps jusqu'au premier emploi. Dans le fichier "jobmar.dta", cette variable n'est pas appropriée, puisque les périodes matrimoniales (fichier "mar.dta") y ont été introduites. La variable "durev" (durée avant l'événement) est préférable. Étant donné qu'aucune femme de l'échantillon n'est née avant 1920, on peut calculer une nouvelle variable indiquant la durée en mois depuis janvier 1920 comme il suit :

```
. use jobmar

. gen dur1920=(v306*12+v305) - 20*12 + durev
```

Cette variable est calculée pour toutes les périodes, pour les migrantes ou non, avant ou après émigration. Pour bien circonscrire les calculs aux seules non migrantes avant une éventuelle émigration, il vaut mieux recalculer la variable avec ces restrictions avant d'exécuter la commande « slice » :

```
. gen dur1920=(v306*12+v305) - 20*12 + durev if immigree==0 & emigc==0
(966 missing values generated)

. slice dur1920, tvid(no) gen(periode) int(480,540,600,660,720,780) sav(jobmar2)
   -480     + 223
[480-540    + 222
[540-600    + 218
[600-660    + 213
[660-720    + 209
[720-780    + 203
[780-      --------
           1288 records added to 1746
```

La commande « slice » crée alors une nouvelle variable "periode", égale à "0" pour les années précédant 1960 (pour dur1920<480), "1" pour les années 60-64 (480<=dur1920<540), "2" pour les années 65-69 (540<=dur1920<600), etc., jusqu'à "6" pour les années 85-89. D'ailleurs, pour plus de commodité, un nouveau libellé s'impose pour la variable "periode" :

```
. lab def periode 0 "<1960" 1 "1960-64" 2 "1965-69" 3 "1970-74" 4 "1975-79" 5
"1980-84" 6 "1985-89", modify
```

Le découpage du temps a été fait au mois près parce que le fichier "jobmar.dta" (construit pour analyser les interactions entre vies active et matrimoniale) l'exige. Cette précision au mois n'est pas indispensable dans la

plupart des analyses, en particulier lorsqu'un seul événement est étudié ou lorsque la datation n'est pas précise. On veut généralement obtenir une indication de l'effet moyen de chaque période, et non pas l'effet d'un événement précis. Le passage d'une période à l'autre est défini exactement (à la seconde près pourrait-on dire), mais la signification que l'on donne à un tel changement est, elle, beaucoup plus souple : il est bien évident que les effets de période sont progressifs.

La commande « slice » a pour effet d'augmenter le nombre de périodes, et donc d'observations dans le fichier. Des précautions doivent être prises avant l'exécution de cette commande, et en particulier le fichier original (ici "jobmar.dta") doit être enregistré au préalable. La variable de durée en fin de période ("durev") doit être modifiée (puisque de nouvelles lignes d'observation sont créées), de même que les indicateurs de troncature :

```
. replace durev=dur1920 + 20*12 - (v306*12+v305)
. drop Tjob1 job1
. censor no durev =v611==2 | v611==3 if immigr==0, gen(job1) before(emig==1)
. drop sal ind Tsal Tind
. censor no durev =v611==2 if immigr==0, gen(sal) before(emig==1 | v611==3)
. censor no durev =v611==3 if immigr==0, gen(ind) before(emig==1 | v611==2)
. drop emig1 Temig1
. censor no durev =emig==1 if immigr==0, gen(emig1) bef(v611==2 | v611==3)
. drop mar1 Tmar1
. censor no durev =statut==1 if immigr==0, gen(mar1) before(emig==1)
```

Une fois ces corrections faites, il suffit de créer des variables dichotomiques pour les périodes traversées par l'enquêtée, et de fixer la modalité de référence, par exemple la période antérieure à 1960 :

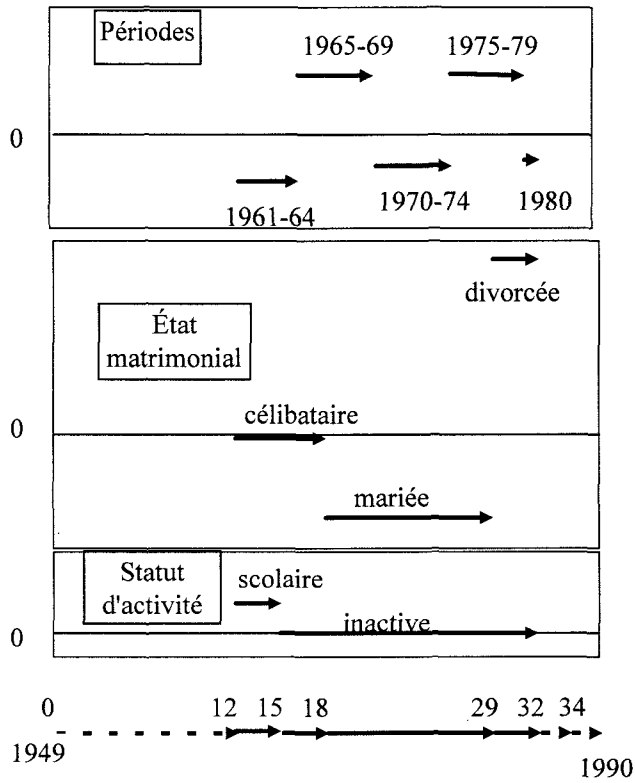
```
. tab periode, gen(per)
 Time period|
     periode|      Freq.     Percent        Cum.
------------+-----------------------------------
      <1960 |         291       14.07       14.07
    1960-64 |         259       12.52       26.60
    1965-69 |         293       14.17       40.76
    1970-74 |         312       15.09       55.85
    1975-79 |         306       14.80       70.65
    1980-84 |         303       14.65       85.30
    1985-89 |         304       14.70      100.00
------------+-----------------------------------
      Total |        2068      100.00

. drop per1

. desc per*

102. periode      byte   %8.0g       periode  Time period periode
111. per2         byte   %8.0g                periode==1960-64
112. per3         byte   %8.0g                periode==1965-69
113. per4         byte   %8.0g                periode==1970-74
114. per5         byte   %8.0g                periode==1975-79
115. per6         byte   %8.0g                periode==1980-84
116. per7         byte   %8.0g                periode==1985-89
```

On peut maintenant estimer la régression avec les nouvelles variables :

```
. cox Tjob1 gener2 gener3 per2-per6 marie divorcee veuve etudes apprent chomet
chomaut, dead(job1) tvid(no) ltolerance(0.000001)

Iteration 0:  Log Likelihood =-435.99761
(etc)
Iteration 3:  Log Likelihood =-417.42127

Cox regression                                        Number of obs =   1639
(Log Likelihood tolerance 1.00e-06)                   chi2(14)      =  37.15
                                                      Prob > chi2   = 0.0007
Log Likelihood = -417.42125                           Pseudo R2     = 0.0426

------------------------------------------------------------------------------
   Tjob1 |
    job1 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  gener2 |   .6028921   .2874607     2.097   0.036        .0394796    1.166305
  gener3 |  -.1734836   .3628861    -0.478   0.633       -.8847273    .5377601
    per2 |    -.18011   .4606375    -0.391   0.696       -1.082943    .7227229
    per3 |   .2526013   .4175859     0.605   0.545       -.5658519    1.071055
    per4 |  -.0920638   .4017164    -0.229   0.819       -.8794134    .6952859
    per5 |   .2654462   .3274664     0.811   0.418       -.3763762    .9072686
    per6 |  -.0727933   .3461038    -0.210   0.833       -.7511442    .6055576
   marie |  -.5792672   .3326314    -1.741   0.082       -1.231213    .0726785
divorcee |   1.466303    .759389     1.931   0.053       -.0220725    2.954678
   veuve |  -1.399301     .79819    -1.753   0.080       -2.963725    .1651223
  etudes |   .1594267   .3351388     0.476   0.634       -.4974333    .8162867
 apprent |   1.589703   .4464137     3.561   0.000        .7147478    2.464657
  chomet |   1.419607    .502117     2.827   0.005        .4354759    2.403739
 chomaut |  -.1209173    1.11889    -0.108   0.914       -2.313901    2.072067
------------------------------------------------------------------------------
```

Le modèle ainsi constitué est assez complexe. Les variables indépendantes sont de plusieurs catégories : endogènes ou exogènes, constantes ou fonction du temps. La prise en compte des variables indépendantes fonction du temps (étape de formation, période) permet de contrôler dans une large mesure le temps d'observation, c'est-à-dire le processus qui mène à l'événement. Les interactions entre le temps d'observation et les variables explicatives sont encore possibles, mais leur influence se trouve limitée par le fait que les variables indépendantes fonction du temps capturent en partie ces interactions. En découpant la période d'observation en tranches les plus homogènes possibles, on atténue l'hypothèse de proportionnalité des quotients.

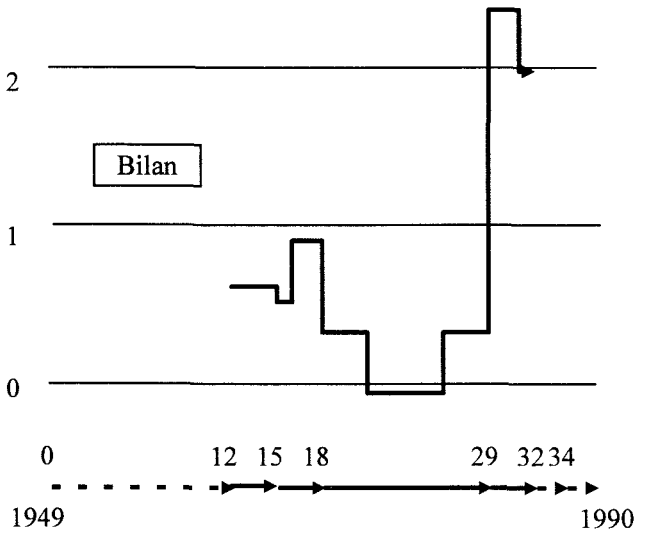
L'individu est encore en quelque sorte suivi au fur et à mesure des étapes de sa vie. Par exemple, Fanta Diallo est née en 1949 et rentre en cours d'observation à 12 ans, c'est-à-dire en 1961. À 15 ans, elle abandonne ses études. De 15 à 18 ans elle est inactive et célibataire. À 18 ans, elle se marie, mais reste inactive. Elle divorce à 29 ans, et, à 32 ans, elle trouve son premier emploi. À 34 ans, elle se remarie. Cette biographie est recueillie fin 1989, alors qu'elle avait 36 ans.



Le tableau des résultats de la régression précédente indique par rapport à une femme appartenant à la catégorie de référence du modèle (de la génération 1930-44, inactive, célibataire, observée avant 1960), les probabilités d'obtenir un emploi pour Fanta en fonction des événements (périodes, changements matrimoniaux, activités). L'effet de la génération (Fanta est née entre 1945 et 1954, soit un coefficient de 0,6) est supposé constant dans le modèle, et s'ajoute aux trois autres variables. Au total, l'effet combiné des quatre variables est le suivant :

Âge	Génération	Période	Statut matr.	Activité	Total
de 12 à 14 ans :	+ 0,603	- 0,180	+ 0	+ 0,159	+ 0,582
à 15 ans :	+ 0,603	- 0,180	+ 0	+ 0	+ 0,423
de 16 à 17 ans :	+ 0,603	+ 0,253	+ 0	+ 0	+ 0,856
de 18 à 20 ans :	+ 0,603	+ 0,253	- 0,579	+ 0	+ 0,277
de 21 à 25 ans :	+ 0,603	- 0,092	- 0,579	+ 0	- 0,068
de 26 à 28 ans :	+ 0,603	+ 0,265	- 0,579	+ 0	+ 0,289
de 29 à 30 ans :	+ 0,603	+ 0,265	+ 1,466	+ 0	+ 2,334
à 31 ans :	+ 0,603	- 0,073	+ 1,466	+ 0	+ 1,996

On peut représenter graphiquement le bilan de ces calculs :



Cet exercice n'a évidemment qu'une valeur d'illustration, et nul n'est besoin de le reproduire pour tous les individus de l'échantillon. Il a cependant le mérite de montrer que les variables indépendantes fonction du temps permettent de suivre l'individu tout au long de sa vie en fonction des événements qu'il connaît ou des périodes qu'il traverse. C'est en ce sens que l'on peut parler d'un modèle dynamique : le temps y est pleinement pris en compte, et cela constitue un avantage indéniable sur les autres modèles probabilistes et sur les analyses plus descriptives.

EXERCICE 8

L'ajout des variables indicatrices de période n'a eu aucun effet sur l'estimation du modèle. Cependant, la génération et la période peuvent avoir des effets conjugués (des interactions) que ne permettent pas de déceler la prise en compte de ces variables séparément. Comment schématiseriez-vous un modèle où l'effet de génération n'est plus une condition permanente, mais agit en interaction avec l'effet de période ?

Pour tester ce type d'interaction, il suffit de créer une variable combinant les informations sur la génération et la période, ainsi que la série de variables dichotomiques correspondantes :

```
. gen genper=strate*10 + periode
. tab genper, gen(gper)
. drop gper1
```

On peut maintenant tester le modèle :

```
. cox Tjob gper* chomet chomaut apprent inact , dead(job1) tvid(no) ltol(.000001)
```

Reportez les coefficients des variables « gper » sur le diagramme de Lexis tracé dans ce chapitre. Quel commentaire vous inspire ces résultats ? Peut-on conclure à une interaction entre effets de période et de génération ? Le modèle est-il meilleur que le précédent ?

EXERCICE 9

Dans le cas où le temps d'observation est défini indépendamment de l'âge (lorsque le début d'observation n'est pas fixé au même âge pour tout l'échantillon), l'interaction entre générations et périodes permet la prise en compte des "âges de la vie", c'est-à-dire de l'âge propre à chaque individu : cette possibilité est très intéressante, en particulier pour l'étude d'un sous-échantillon hétérogène selon l'âge. C'est le cas par exemple des immigrants que l'on peut étudier depuis le moment de leur arrivée dans le champ de l'enquête. On peut aussi introduire une troisième variable, le groupe d'âges quinquennal, soit au lieu de la variable de période de conjoncture, soit au lieu de la variable groupe de générations.

Une telle variable sera utile pour contrôler l'âge d'un échantillon non homogène du point de vue de l'âge. Ce serait le cas par exemple si l'on avait étudier l'accès à l'emploi des femmes migrantes à partir du moment de leur migration à Dakar (voir dans le chapitre les parties consacrées à la conceptualisation de l'événement et de la population soumise au risque).

Un exercice avancé consiste à considérer dans le fichier JOBMAR.DTA, non plus les femmes présentes à Dakar à l'âge de 12 ans, mais les femmes immigrées après 12 ans. On pourra tester la différence entre le modèle où intervient l'effet simple de l'âge d'arrivée, et le modèle avec effets des groupes d'âge traversés jusqu'à l'événement, avec ou sans interaction avec la période.

CONCLUSION

MÉRITES ET LIMITES DE L'ANALYSE QUANTITATIVE DES BIOGRAPHIES

Les premiers chapitres de ce manuel insistent beaucoup sur le temps. Le temps est au cœur de l'analyse causale. L'analyse biographique, en contrôlant cette dimension essentielle, rapproche au plus près la statistique du raisonnement causal. Certes, les théories générales en sciences sociales ne peuvent être facilement réduites à de simples modèles, mais, au niveau des théories dites auxiliaires (spécifiques ou partielles), l'analyse biographique est appelée à jouer un rôle de plus en plus important en sciences sociales. Elle donne une impulsion supplémentaire au va-et-vient incessant entre la vérification des théories et la recherche de nouvelles explications.

Le modèle semi-paramétrique est certainement le plus achevé dans l'analyse des biographies. Il contrôle les variables explicatives indépendamment l'une de l'autre, tout en respectant le principe de priorité temporelle cher au raisonnement causal. De plus, il prend en compte au maximum l'information sur les données incomplètes grâce au traitement des troncatures.

Mais surtout, le propre du modèle semi-paramétrique est de suivre au plus près la succession des événements. L'usage des variables indépendantes fonction du temps a permis un bond qualitatif important. On peut parler de modèles dynamiques dans le sens où c'est bien une chaîne causale que l'on teste. On tente de mesurer la probabilité de connaître un événement au fur et à mesure que se déroule la vie de l'individu jusqu'au moment de l'enquête.

Cela dit, on ne pourra jamais prévoir exactement les comportements humains, même avec le meilleur modèle qui soit. En sciences sociales, les événements doivent être considérés comme le résultat de processus complexes de décisions, parfois irrationnelles, toujours incertaines. On ne peut se soustraire à cet indéterminisme fondamental, qui pose une limite au traitement en termes de causalité. Cet indéterminisme a pour conséquence que les problèmes du temps "flou" et de l'hétérogénéité non observée ne seront jamais réduits à néant. On peut

en réduire la portée par plus de précision dans la collecte et dans la conceptualisation, mais on ne se satisfera jamais de n'avoir pas "tout" compris.

En guise de conclusion, nous revenons sur deux directions de recherche en analyse des biographies : il s'agit d'une part de poursuivre en amont le travail conceptuel et le recueil des données biographiques, et d'autre part d'améliorer en aval le diagnostic après modélisation.

1. La violation de l'hypothèse de proportionnalité

Les variables fixées en début d'observation (à la naissance ou à un âge fixe) sont supposées, dans un modèle à risques proportionnels, avoir une influence tout au long de la vie de l'individu. Or, une interaction entre le temps d'observation et ces variables est possible. L'appartenance sociale peut jouer, par exemple, sur les premières années d'observation de l'individu et décroître ensuite, ce qui remet en cause l'hypothèse de proportionnalité des quotients tout au long du temps d'observation.

On peut vérifier cette hypothèse avec, par exemple, un test graphique, mais ensuite, il est difficile de corriger le modèle. En effet, on imagine assez facilement la complexité d'un modèle faisant ainsi intervenir, pour chaque variable, une interaction selon le temps d'observation, qui ne devrait pas nécessairement être linéaire en fonction du temps. La solution pratique consiste bien souvent à stratifier l'échantillon selon la variable pour laquelle l'hypothèse de proportionnalité n'est pas vérifiée, au risque d'aboutir à de trop petits échantillons.

Les événements qui surviennent en cours d'observation permettent de lever en partie cette indétermination par rapport au temps, mais il n'est pas possible de vérifier l'hypothèse de proportionnalité à l'aide des techniques actuelles après introduction des variables indépendantes fonction du temps. Il reste encore un travail important à faire pour vérifier l'adéquation entre les données biographiques et le modèle semi-paramétrique dans sa version la plus évoluée (avec variables fonction du temps). Pour l'instant, ce travail de vérification se fait essentiellement au niveau de la mesure du gain total de vraisemblance, ou bien du degré de significativité des coefficients de la régression, mais les diagnostics sont pauvres en ce qui concerne l'analyse des résidus et la vérification de l'hypothèse de proportionnalité des quotients.

En outre, la prise en compte des variables fonction du temps n'évacue pas le problème de l'interaction entre le temps d'observation la nature de la période, que l'on appelle encore une dépendance sur la durée (*duration dependance*). Par exemple, la durée du chômage à l'issue d'une période de formation a certainement

une influence sur l'orientation vers tel ou tel emploi. Le chômeur peut dans un premier temps préférer rester au chômage dans l'espoir de trouver l'emploi qui lui convient, mais il acceptera certainement un emploi moins intéressant si son chômage dure trop longtemps. La question est encore de savoir quelle est l'ampleur de cette interaction, son sens et sa forme.

Si l'on soupçonne ce type d'interaction, il est nécessaire d'étudier chaque période séparément. Plutôt que d'étudier l'accès au premier emploi à l'issue de la formation, on peut étudier chacune des étapes avant le premier emploi : études, apprentissage, chômage. Outre l'ampleur de la tâche, il se pose alors d'autres problèmes méthodologiques relatifs à la taille de l'échantillon et surtout, à la datation de chacune de ces périodes.

2. Le traitement du temps flou : erreur de lecture chronologique et intentionnalité des acteurs sociaux

Pour analyser dans de bonnes conditions les données biographiques, il est nécessaire de bien définir les événements, le temps d'observation et les variables indépendantes, qu'elles soient ou non fonction du temps.

Dans certains cas, même la définition du temps de l'événement pose un problème. L'indétermination temporelle ne provient pas seulement d'un mauvais recueil des données (erreur de datation, omission d'événements) mais aussi lorsqu'un événement a lieu par anticipation d'un autre événement. C'est le cas par exemple de la migration et de l'obtention d'un nouvel emploi, de la reprise de la fécondité et d'un déménagement, ou du mariage et de la naissance d'un enfant : le premier événement peut, par anticipation, avoir lieu soit avant, soit simultanément au second, alors que la cause du premier événement est en fait le second événement.

Le problème de l'anticipation des événements provient en fait d'une indétermination conceptuelle de la causalité. Le principe de l'antériorité de la cause sur l'effet n'est pas remis en question, mais le processus qui mène à la réalisation de l'événement doit être défini de manière plus complexe, en faisant intervenir la décision des acteurs. Par exemple, dans le cas du mariage, la cérémonie à la mairie n'est que l'aboutissement de longs préparatifs que l'on peut par exemple faire débuter à la date des fiançailles. Il s'agit en quelque sorte de trouver, à l'aide d'un meilleur système de collecte, les événements les plus pertinents du point de vue de la biographie des individus. Mais comme on l'a dit plus haut, on se trouvera toujours confronté, tôt ou tard, à l'indéterminisme fondamental des actions humaines. Les limites pourront être repoussées, jamais supprimées.

ANNEXE 1

RÉFÉRENCES DES PROGICIELS UTILISÉS DANS CE MANUEL

Pour commander les progiciels STATA, Stat/Transfer leurs manuels d'utilisation, et le STATA Technical Bulletin (STB), s'adresser à :

Stata Corporation
702 University Drive East
College Station, TX 77840 USA
Tel. : (1) 409-696-4600 ou (1) 800-STATA-PC
Fax. : (1) 409-696-4601
Courrier électronique : stata@stata.com

Pour obtenir le progiciel ISSA, s'adresser à :

ISSA - *Integrated Statistical and Survey Analysis*
Macro International
Department DHS
1178 Beltsville Drive
Calverston, MD 20 705 U.S.A.
Tél. : 301-572-0200
Fax : 301-575-0999
E-Mail : james@macroint.com

ANNEXE 2

LA SAISIE DES FICHIERS BIOGRAPHIQUES ET LEUR TRANSFERT DANS STATA

Si les données ont été saisies avec un logiciel tel que SAS, SYSTAT, Dbase, Lotus ou un autre tableur, Stat/Transfer permet leur conversion directe en format STATA. Il n'est pas donc pas indispensable d'utiliser ISSA, mais on doit savoir que c'est un progiciel particulièrement adapté à la saisie des questionnaires biographiques : chaque type de périodes (emplois, migrations, états matrimoniaux, enfants...) peut constituer un module où les périodes sont saisies successivement selon un masque de saisie conforme au questionnaire. Les modules peuvent être extraits du fichier global, et combinés entre eux pour former de nouveaux fichiers.

Une fois la saisie terminée, on dispose d'un fichier ASCII, qu'on peut lire ensuite à l'aide d'un dictionnaire de variables selon le format imposé par STATA. La réécriture d'un dictionnaire est non seulement fastidieuse mais présente des risques d'erreur (emplacement des variables, codification, labels, etc.). ISSA ne permet pas la conversion directe des fichiers en format STATA mais une procédure d'exportation en format SPSS est disponible et conserve tout le dictionnaire déjà utilisé dans ISSA. Ensuite, il suffira de convertir ce fichier SPSS en STATA avec le progiciel Stat/Transfer commercialisé comme STATA.

Conversion du fichier ISSA en un fichier lisible par SPSS

À l'aide de ISSA, on crée une procédure en "batch" dans laquelle on spécifie le dictionnaire à lire, le fichier à convertir et la nature de la procédure (EXPORT). Une fois la fenêtre de la procédure disponible, on écrit les commandes suivantes :

```
split case by (nom de la section)
include (nom des variables).
```

On sauvergarde la procédure et on l'éxécute. Le fichier de sortie, à la fin de l'éxécution de la procédure est un fichier rectangulaire ASCII à extention « .SPS », lisible par le logiciel SPSS.

En règle générale, la section qui intervient dans la commande « split case by () », est une section multiple. On peut ensuite ajouter par la commande « include » des sections ou des variables simples.

Conversion du fichier « SPS » en un fichier SPSS exportable

À cette étape on passe au logiciel SPSS. Une fois SPSS lancé, on charge en mémoire le fichier « .SPS » et on éxécute la commande EXPORT de SPSS :

```
GET File 'path...\nom.SPS'
EXPORT OUTFILE 'nom.XPT'
```

Après l'éxécution de ces commandes, le fichier résultat qu'on obtient est un fichier exportable à extension « .XPT ».

Conversion du fichier exportable « .XPT » en un fichier STATA

Cette étape est la dernière à effectuer avant d'obtenir le fichier STATA. Une fois sorti de SPSS, lancer le logiciel Stat/Transfer. Choisir le fichier sous format *SPSS export file* comme fichier d'entrée (*input* ou origine), et le format *STATA System file* comme fichier de sortie (*output* ou destination). Sélectionner les variables si nécessaire et confirmez. Le fichier « .DTA » est prêt pour l'analyse avec STATA.

Quelques précautions à prendre avant de convertir les fichiers

Lorsque qu'après la conversion vous parcourez votre fichier de données rectangularisé à l'aide d'un éditeur de texte, vous pouvez rencontrer :

- des tildes (~) au début de certaines lignes
- des étoiles (*) à n'importe quel endroit du fichier

Les tildes au début d'une ligne représentent un enregistrement effacé mais non compacté. Avant de faire l'exportation de ISSA vers SPSS, veillez à ce que le fichier de données soit compacté, en plaçant dans ISSA le curseur côté dictionnaire et en tapant F7 pour *fileservices*. Choisir *compact*, puis préciser le nom du fichier de données et faire *proceed*. Dans le fichier obtenu, vous n'aurez plus de tildes (~).

S'agissant des étoiles qui apparaissent dans le fichier de données au cours du processus d'exportation, elles sont dues au fait que certains cas (qualifiés de "*not applicable*") ne sont pas prévus dans le dictionnaire. Il faut donc charger le dictionnaire et pour toutes les variables pouvant prendre la valeur " *not applicable*", mettre « *not applicable=blank* » au lieu de « *not applicable=none* ». Une exportation effectuée en ayant pris ces précautions donnera lieu à un fichier de données sans étoiles.

ANNEXE 3

DATATION DES ÉVÉNEMENTS DANS LE QUESTIONNAIRE BIOGRAPHIQUE

Il est nécessaire, pour l'analyse quantitative des biographies, que les événements soient enregistrés dans l'ordre le plus exact possible. La précision au mois près n'est pas nécessaire en soi, mais elle est bien utile pour établir l'ordre des événements qui ont eu lieu la même année.

Des principes ont été établis pour que les enquêteurs de l'équipe IFAN-ORSTOM puissent obtenir une réponse la plus précise possible auprès des enquêtés, en retenant l'ordre des événements à défaut de leur date au mois près. Une codification spéciale pour les mois imprécis a été adoptée : 33, 66, 88, pour deux événements qui ont eu lieu la même année :

1. L'enquêté déclare de lui-même des dates au mois près : ces dates sont alors inscrites sur le questionnaire.

2. L'enquêté ne se souvient pas des mois, mais il sait que les deux événements se sont passés en même temps (ou dans un intervalle inférieur à un mois). Si aucune date n'a pu être estimée, le code pour le mois est 66 pour les deux événements qui ont eu lieu simultanément.

3. L'enquêté ne se souvient pas des mois, mais il sait que l'un des événements s'est passé avant l'autre (à plus d'un mois d'intervalle). Si l'enquêteur réussit à dater approximativement un des événements, il doit faire en sorte de situer l'autre événement par rapport au premier, comme dans le 4ème cas ci-dessous. Si aucune date n'a pu être estimée, le code pour le mois est fixé à 33 pour le premier événement et 88 pour le second, pour que l'on sache au moment de l'analyse quel événement a eu lieu avant l'autre.

4. L'enquêté donne l'une des dates seulement (soit qu'il s'en souvient spontanément, soit que l'enquêteur l'ait aidé en lui posant des questions). Deux cas se présentent :

- l'autre événement à dater est postérieur au premier événement daté : on propose une date entre le premier événement et la fin de l'année, ou bien on fixe la date au milieu de cet intervalle. Rappelons que la datation précise n'est pas nécessaire, et qu'il suffit de connaître l'ordre des événements (la même remarque vaut pour le cas suivant) ;
- l'autre événement à dater est antérieur au premier événement daté : on propose une date entre le début de l'année et le premier événement.

5. L'enquêté ne se souvient ni des mois, ni de l'ordre des événements : si malgré toutes les propositions de dates, l'enquêteur n'obtient aucune réponse sûre, on code 99 pour les deux événements. Ce cas n'arrive que très rarement et correspond plutôt à un oubli de l'enquêteur ou à une erreur de codification. Les enquêtés situent aisément deux événements qu'ils ont vécus l'un par rapport à l'autre dans leur ordre chronologique.

En fait, le recoupement des informations suffit à établir la chronologie, même si l'enquêteur n'avait pas inscrit les bons codes sur le questionnaire lors de son passage sur le terrain. Parfois, la confrontation de plusieurs questionnaires permet de vérifier la cohérence des réponses, dans le cas où deux individus d'un même ménage (mari et femme, deux frères, etc.) sont soumis au questionnaire biographique, et ont connu des événements ensemble. Pour l'équipe de supervision de l'enquête IFAN-ORSTOM, une des tâches prioritaires a en effet consisté à vérifier, dès le questionnaire rempli, la cohérence de la datation des événements. Les enquêteurs, en particulier au début de l'enquête, étaient renvoyés sur le terrain lorsque les erreurs de datation n'étaient pas corrigibles sans vérification auprès de l'enquêté.

En somme, le bon recueil de la chronologie des événements dans l'enquête biographique nécessite une attention particulière de l'équipe encadrante, qui doit nécessairement être très présente sur le terrain. Les contrôles manuels ajoutés au contrôle automatique, lors de la saisie informatique, font que les cas de simultanéité dans la datation sont réduits au minimum, ce qui présente un avantage considérable pour l'analyse.

RÉFÉRENCES BIBLIOGRAPHIQUES

ANDERSEN Per Kragh, BORGAN Ornulf, GILL Richard D. et KEIDING Neil, 1993. – "Statistical Models based on Counting Processes", *Springer Series in Statistics*, Springer Verlag, 768 p.

BLALOCK H.M., 1982. – *Conceptualization and measurement in the social sciences.* – Sage pubications, Beverly Hills.

BOCQUIER Philippe, 1996. – *L'insertion et la mobilité professionnelles à Dakar.* – Paris, Université Paris V - René DESCARTES - Sorbonne, 375 p. (Thèse de doctorat en démographie, sous la direction de Yves CHARBIT).

COURGEAU Daniel et LELIEVRE Éva, 1989. – *Analyse démographique des biographies.* – INED, 268 p.

COURGEAU Daniel et LELIÈVRE Éva, 1992. – *Event history analysis in demography*, Clarendon Press, OUP, Oxford, 226 p.

COURGEAU Daniel et LELIÈVRE Éva, 1995. – "L'apport de l'analyse biographique en démo-économie", *in : L'analyse longitudinale en économie sociale.* – ADEPS/Université Nancy II/CNRS. (15èmes journées de l'Association d'Économie Sociale, 14-15 septembre 1995).

COX David R., 1972. – Regression models and life tables (with discussion), *Journal of royal statistical society*, b34, p. 187-220.

DROESBEKE Jean-Jacques, FICHET Bernard et TASSI Philippe (éds), 1989. – *Analyse statistique des durées de vie. Modélisation des données censurées.* – Economica, 1989, 282 p.

DUCHÊNE Josiane, WUNSCH Guillaume, VILQUIN Éric, 1987. – *Chaire Quételet '87 L'explication en sciences sociales, la recherche des causes en démographie.* – Louvain-la-Neuve, Institut de Démographie, Université Catholique de Louvain/Éditions Ciaco, 478 p.

HECKMAN James et SINGER Burton, 1985. – "Longitudinal analysis of labor market data", *in : Econometric Society Monographs.* – Cambridge, Cambridge University Press, 410 p.

KALBFEISCH John et PRENTICE Ross L., 1980. – *The statistical analysis of failure time data.* – New York/Chichester/Brisbane/Toronto, John Wiley and sons, 321 p.

LELIÈVRE Éva et BRINGE Arnaud. – *Manuel pratique pour l'analyse statistique des biographies. Présentation des modèles de durée et utilisation des logiciels SAS, TDA et STATA*, préface de Daniel COURGEAU. – Paris, INED, à paraître.

MAYER Karl Ulrich et TUMA Nancy Brandon, 1990. – "Event History Analysis in Life Course Research", *in : Life Course Studies.* – Madison, The University of Wisconsin Press, 297 p.

RIANDEY Benoît, 1985. – "L'enquête biographique familiale, professionnelle et migratoire (INED 81). Le bilan de la collecte", *in : Chaire Quételet 83 : Migratoins internes, collecte des données et méthodes d'analyse*, p. 117-134. – Louvain-la-Neuve, Université catholique de Louvain, 460 p.

STATA CORPORATION, 1995. – *Stata reference Manuel : Statistics, Graphics, Data Management*, vol 3. – Texas, Stata Corporation, College Station.

WUNSCH Guillaume, 1988. – *Causal Theory and Causal Modeling*. – Louvain-la-Neuve, Leuven University Press, 200 p.

INDEX

D

E

F

G

H

I

K

L

M

O

P

Q

R

S

T

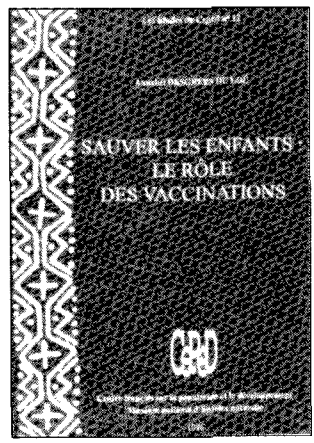
U

V

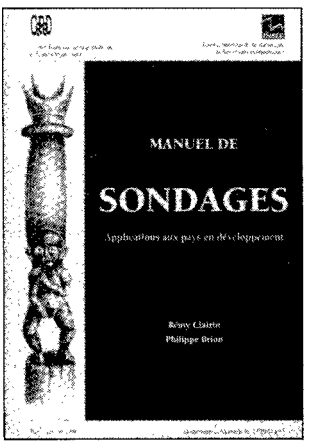
W

LES PUBLICATIONS DU CEPED

Collection Les Études du CEPED



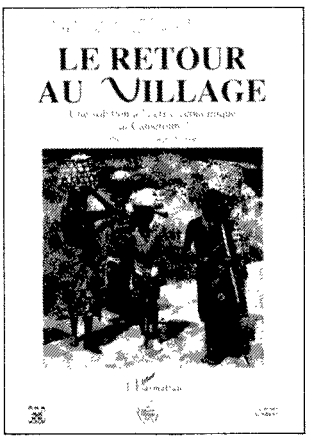
- n°13 : *Crise et population en Afrique*, par Jean COUSSY et Jacques VALLIN (dir.) (1996). (À paraître).
- n°12 : *Sauver les enfants : le rôle des vaccinations*, par Annabel DESGRÉES DU LOÛ (1996), 261 p. (100 F).
- n°11 : *L'économie algérienne à l'épreuve de la démographie*, par Lhaocine AOURAGH (1996), 337 p. (100 F).
- n°10 : *Conséquences démographiques du sida en Abidjan : 1986-1992*, par Michel GARENNE *et al.* (1995), 198 p. (100 F).
- n° 9 : *La maternité chez les Bijago de Guinée-Bissau*, par Alexandra DE SOUSA et Dominique WALTISPERGER (collab.) (1995), 114 p. (100 F).
- n° 8 : *La crise de l'asile politique en France*, par Luc LEGOUX (1995), 344 p. (100 F).
- n° 7 : *L'entrée en vie féconde. Expression démographique des mutations socio-économiques d'un milieu rural sénégalais*, par Valérie DELAUNAY (1994), 326 p. (90 F).
- n° 6 : *La traite des esclaves au Gabon du XVIIe au XIXe siècle, essai de quantification pour le XVIIIe siècle*, par Nathalie PICARD-TORTORICI et Michel FRANÇOIS (1993), 156 p. (90 F).
- n° 5 : *Croissance urbaine, migrations et population au Bénin*, par Julien GUINGNIDO GAYE (1992), 114 p. (100 F).
- n° 4 : *Un siècle de démographie tamoule*, par Christophe GUILMOTO (1992), 175 p. (120 F).
- n° 3 : *Mobilité spatiale et mobilité professionnelle dans la région nord-andine de l'Équateur*, par Jean PAPAIL (1991), 87 p. (80 F).
- n° 2 : *Mortal, logiciel d'analyse de la mortalité*, par Jean-Michel COSTES et Dominique WALTISPERGER (1988), 99 p. + disquette. (épuisé).
- n° 1 : *De l'homme au chiffre, réflexions sur l'observation démographique en Afrique*, par Louis LOHLÉ-TART et Rémy CLAIRIN (1988), 329 p. (150 F).



Collection Documents et Manuels du CEPED

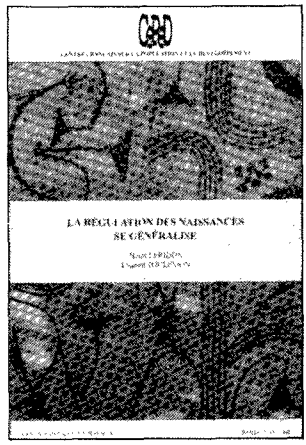
- n° 4 : *L'analyse des enquêtes biographiques à l'aide du logiciel STATA*, par Philippe BOCQUIER (1996), 208 p. (120 F).
- n° 3 : *Manuel de sondages. Applications aux pays en développement*, par Rémy CLAIRIN et Philippe BRION (1996), 104 p. (80 F).
- n° 2 : *Clins d'œil de démographes à l'Afrique et à Michel François*, Jacques VALLIN (éd.) (1995), 244 p. (80 F). (épuisé).
- n° 1 : *La démographie de 30 États d'Afrique et de l'Océan Indien*, CEPED (1994), 352 p. (épuisé).

COÉDITIONS



- *Le retour au village*, par Patrick Gubry *et al.* (1996), CEPED/IFORD/MINREST/L'Harmattan, 206 p. (120 F).
- *Les familles dakaroises face à la crise*, par Philippe ANTOINE *et al.* (1995), CEPED/IFAN/ORSTOM, 212 p. (80 F).
- *Populations africaines et sida*, sous la direction de Jacques VALLIN (1994), CEPED-La Découverte, 218 p. (149 F).
- *Intégrer Population et Développement*, (1994) Académia-CEPED-CIDEP-l'Harmattan-UCL, 824 p. (400 F).
- *La population de l'Afrique. Manuel de démographie*, par Francis GENDREAU (1993), CEPED-Karthala, 463 p. (180 F).
- *Politiques de développement et croissance démographique rapide en Afrique*, par Jean Claude CHASTELAND, Jacques VÉRON et Magali BARBIERI (éds.) (1993), 314 p. (INED-CEPED-PUF) (180 F).
- *Migration, urbanisation et développement au Cameroun*, par Joseph-Pierre TIMNOU, (1993), Les cahiers de l'IFORD, n° 4, CEPED-IFORD, 115 p. (gratuit).
- *Migration, urbanisation et développement au Congo*, par Gabriel TATI (1993), Les cahiers de l'IFORD, n° 5, CEPED-IFORD, 94 p. (gratuit).
- *Condition de la femme et population : le cas de l'Afrique francophone*, édité par Thérèse LOCOH (1992), CEPED-FNUAP-ONU-URD, 116 p. (épuisé).
- *Comores, les enfants du volcan.* Le recensement général de la population des Comores en (Septembre 1991), film vidéo 30 minutes, AFEP-CEPED. (gratuit).
- *Le spectre de Malthus, déséquilibres alimentaires, déséquilibres démographiques*, édité par Francis GENDREAU (1991), CEPED-EDI-ORSTOM, 444 p. (230 F).

Collection Les Dossiers du CEPED (30 F/numéro)



- n° 42 : *La polyandrie chez les Bashilele du Kasaï occidental (Zaïre) : fonctionnement et rôles*, par Séraphin NGONDO A PITSHANDENGE, 20 p.
- n° 41 : *La régulation des naissances se généralise*, par Henri LERIDON et Laurent TOULEMON, 19 p.
- n° 40 : *Ho Chi Minh Ville : de la migration à l'emploi*, par Truong SI ANH, Patrick GUBRY, Vu Ti HONG et Jerrold W. HUGUET, 52 p.
- n° 39 : *La population de Cuba : principales caractéris-tiques et tendances démographiques*, par Sonia I. CATASUS CERVERA, 35 p.
- n° 38 : *Effets de la guerre civile au Centre-Mozambique et évaluation d'une intervention de la croix rouge*, par Michel GARENNE, Rudi CONINX et Chantal DUPUY, 25 p.
- n° 37 : *Ressources économiques et comportements démographiques des ménages agricoles : le cas des Éwé du Sud-Togo*, par Kokou VIGNIKIN, 35 p.
- n° 36 : *Structure de production et comportement procréateur en Côte d'Ivoire*, par Aka KOUAMÉ et Mburano RWENGE, 31 p.
- n° 35 : *Les migrations comoriennes en France : histoire de migrations coutumières*, par Géraldine VIVIER, 38 p.
- n° 34 : *La transition démographique. Trente ans de bouleversements (1965-1995)*, par Jean-Claude CHESNAIS, 25 p.
- n° 33 : *Pluralisme thérapeutique et stratégies de santé chez les Évhé du sud-est Togo*, par Nadia LOVELL, 20 p.
- n° 32 : *Peut-on échapper à la polygamie à Dakar ?*, par Philippe ANTOINE et Jeanne NANITELAMIO, 31 p.
- n° 31 : *Familles Africaines, population et qualité de la vie*, par Thérèse LOCOH (1995), 48 p. (3^{e} tirage).
- n° 30 : *La mortalité dans le monde : tendances et perspectives*, par France MESLÉ et Jacques VALLIN (1995), 25 p. (3^{e} tirage).
- n° 29 : *Planification sanitaire et ajustement structurel au Cameroun*, par Antoine KAMDOUM (1994), 40 p. (épuisé).
- n° 28 : *Migration et sida en Afrique de l'Ouest, un état des connaissances*, par Richard LALOU et Victor PICHÉ (1994), 52 p. (2^{e} tirage).
- n° 27 : *Éducation de la mère et soins aux enfants à Ouagadougou*, par Christine OUEDRAOGO (1994), 37 p.
- n° 26 : *Réflexions sur l'avenir de la population mondiale*, par Jacques VALLIN (1994), 24 p. (3^{e} tirage).
- n° 25 : *Facteurs de fécondité en milieu rural forestier ivoirien*, par KOFFI N'GUESSAN (1993), 40 p.

- n° 24 : *Les disparités régionales de la mortalité au Bénin*, par Martin LAOUROU (1993), 36 p.

- n° 23 : *Contribution à l'étude de l'évolution de la population de l'Afrique Occidentale Française 1904-1960*, par Raymond R. GERVAIS (1993), 50 p.

- n° 22 : *Solidarité dans la crise ou crise des solidarités familiales au Cameroun ?* par Parfait Martial ELOUNDOU-ENYEGUE (1992), 40 p.

- n° 21 : *La mortalité des enfants à Luanda*, par Maria Julia VAZ-GRAVE (1992), 39 p.

- n° 20 : *Mortalité maternelle : deux études communautaires en Guinée*, par Pierre CANTRELLE, Patrick THONNEAU et Boubacar TOURE (1992), 43 p.

- n° 19 : *Vingt ans de planification familiale en Afrique sub-saharienne*, par Thérèse LOCOH (1992), 27 p. (épuisé).

- n° 18 : *Les déterminants de la mortalité des enfants dans le tiers-monde*, par Magali BARBIERI (1991), 33 p. (épuisé).

- n° 17 : *La fécondité en Mauritanie*, par Keumaye IGNEGONGBA (1991), 39 p. (épuisé).

- n° 16 : *Dix problèmes de population en perspective - Hommage à Jean Bourgeois-Pichat et à Alfred Sauvy*, par Léon TABAH (1991), 31 p. (épuisé).

- n° 15 : *La mesure de l'infécondité et de la sous-fécondité*, par Evina AKAM (1990), 39 p. (épuisé).

- n° 14 : *Statut de la femme, structure familiale, fécondité : transitions dans le Golfe du Bénin*, par Laurent Mensan ASSOGBA (1988), 28 p. (épuisé).

- n° 13 : *Estimer la mortalité maternelle à l'aide de la méthode des sœurs*, par Véronique FILIPPI et Wendy GRAHAM (1990), 29 p. (épuisé).

- n° 12 : *La montée du célibat féminin dans les villes africaines. Trois cas : Pikine, Abidjan et Brazzaville*, par Philippe ANTOINE et Jeanne NANITELAMIO (1990), 27 p. (épuisé).

- n° 11 : *Deux études sur l'emploi dans le monde arabe*, par Jacques CHARMES (1990), 37 p. (épuisé).

- n° 10 : *Facteurs culturels et sociaux de la santé en Afrique de l'Ouest*, par Pierre CANTRELLE et Thérèse LOCOH (1990), 36 p. (épuisé).

- n° 9 : *Éléments du débat population - développement*, par Jacques VÉRON (1989), 48 p. (2^{e} tirage).

- n° 8 : *Transformations agraires et mobilités de la main d'oeuvre dans la région Nord Andine de l'Équateur*, par LE CHAU et Jean PAPAIL (1989), 18 p.

- n° 7 : *Prospective des déséquilibres mondiaux - démographie et santé*, par Pierre CANTRELLE et Francis GENDREAU (1989), 33 p. (épuisé).

- n° 6 : *Les politiques de population en matière de fécondité dans les pays francophones : l'exemple du Togo*, par Thérèse LOCOH (1989), 20 p. (épuisé).

- n° 5 : *Rétention de la population et développement en milieu rural : à l'écoute des paysans Mafa des Monts Mandara (Cameroun)*, par Patrick GUBRY (1988), 24 p. (épuisé).

- n° 4 : *État et besoins de la recherche démographique dans la perspective des recommandations de la conférence de Mexico et de ses réunions préparatoires*, par Jean-Claude CHASTELAND (1988), 23 p. (épuisé).
- n° 3 : *La fécondité en Afrique noire : un progrès rapide des connaissances mais un avenir encore difficile à discerner*, par Thérèse LOCOH (1988), 26 p. (épuisé).
- n° 2 : *Politiques africaines en matière de fécondité : de nouvelles tendances*, par Patrick GUBRY et Mpembele SALA DIAKANDA (1988), 50 p. (épuisé).
- n° 1 : *La connaissance des effectifs de population en Afrique : bilan et évaluation - Hommage à Rémy Clairin*, par Rémy CLAIRIN et Francis GENDREAU (1988), 35 p. (épuisé).

Collections en langues étrangères (anglais et espagnol)

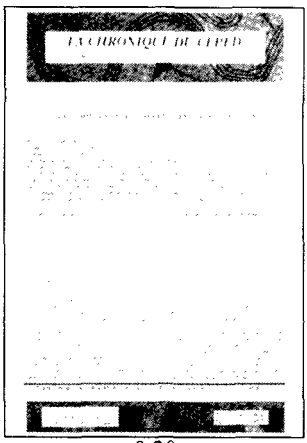
The CEPED Series (37 F/numéro)

- n° 1 : *Mortality in the world : trends and prospects*, by France MESLÉ and Jacques VALLIN, 24 p. (Translated from French by Isabelle Wallerstein).

Los Documentos del CEPED (37 F/numéro)

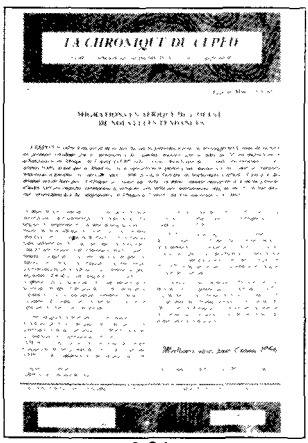
- n° 1 : *La mortalidad en el mundo : tendencias y perspectivas*, para France MESLÉ y Jacques VALLIN, 24 p. (Traducido del francès para Maria Celina AÑAÑOS).

La Chronique du CEPED, bulletin de liaison trimestriel (10 F/numéro ou 30 F/an)
n° 1 (Printemps 1991) à n° 22 (Juillet-septembre 1996)



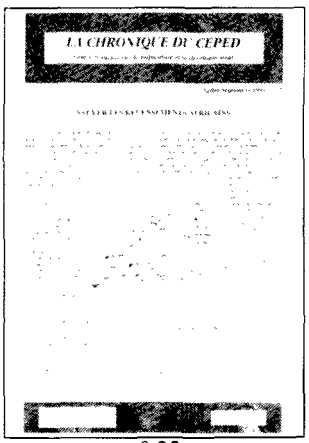
n° 20

Migrations en Afrique de l'Ouest : de nouvelles tendances



n° 21

Les familles africaines en plein remue-ménage



n° 22

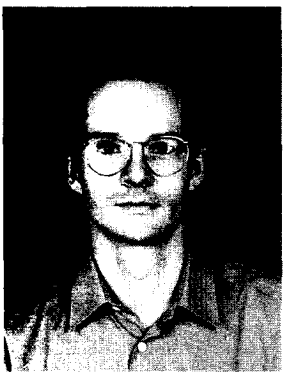
Sauver les recensements africains

Collection Données de base sur la population (gratuit)
(dossiers réalisés par N. LOPEZ-ESCARTIN)
(* encore disponible)

n° 1 : Cameroun
n° 2 : Madagascar*
n° 3 : Gabon
n° 4 : Togo*
n° 5 : Tchad
n° 6 : Bénin
n° 7 : Sénégal
n° 8 : Congo
n° 9 : Rwanda*
n° 10 : Guinée Équatoriale
n° 11 : Gambie
n° 12 : Djibouti
n° 13 : Mali
n° 14 : Mauritanie
n° 15 : Burundi*
n° 16 : Centrafrique
n° 17 : Angola
n° 18 : Côte d'Ivoire*
n° 19 : Zaïre*
n° 20 : Guinée*
n° 21 : Burkina Faso*
n° 22 : Maurice*
n° 23 : Comores
n° 24 : Niger*
n° 25 : Guinée Bissau*
n° 26 : Seychelles*
n° 27 : Cap Vert*
n° 28 : Sao Tome e Principe*
n° 29 : Mozambique*
Nigéria*
Viet Nam

Reproduit par l'Imprimerie MOREAU
Route de Mettray
37390 LA MEMBROLLE-SUR-CHOISILLE
Tél. : 02 47 41 28 35
Fax : 02 47 54 70 52

Dépôt légal 3ème trimestre 1996



Philippe BOCQUIER, sociologue et démographe, a obtenu un Master en statistiques à la London school of economics and political science. Il a travaillé de 1988 à 1992 au centre ORSTOM de Dakar sur l'insertion et la mobilité professionnelle, recherches qui l'ont mené au doctorat en démographie et l'Université Paris V René Descartes. Chercheur de l'ORSTOM, il travaille depuis 1992 au Centre d'études et de recherche sur la population pour le développement (CERPOD) à Bamako, où il co-anime le Réseau migrations et urbanisation en Afrique de l'Ouest (REMUAO).

L'analyse des enquêtes biographiques est réputée d'accès difficile. Pourtant, les techniques ont évolué et ont montré leur efficacité. L'objectif de ce manuel est de rendre accessibles aux chercheurs en sciences sociales de disciplines les plus diverses, des techniques développées en démographie et qui peuvent être utilisées dès lors que les données font intervenir la dimension temporelle.

Le manuel propose d'abord au lecteur un travail conceptuel en lui soumettant une réflexion sur la temporalité et la causalité. La logique de l'analyse multivariée est ensuite expliquée à l'aide des régressions statistiques simples et du modèle de régression logistique. Les tables de séjour introduisent les notions de temps et d'événement.

Les principes de la régression et des tables de séjour sont combinés pour analyser les biographies. Les événements que connaît l'individu au cours de sa vie sont analysés en relation les uns avec les autres. On peut par exemple déterminer l'effet de la migration sur la vie professionnelle, de l'emploi sur la vie matrimoniale, du mariage sur l'itinéraire résidentiel, etc. Ce type d'analyse permet de maîtriser au mieux la dimension temporelle, et, de fait, l'analyse causale.

En suivant les exemples détaillés, le lecteur sera à même de maîtriser les outils d'analyse des biographies. En utilisant les programmes fournis sur disquette avec ce manuel, il pourra traiter ses propres données et arriver à une quasi-autonomie sur un matériel informatique courant.

Contenu de la disquette

CEPED
15 rue de l'École de médecine
75270 Paris cedex 06
Téléphone : (33) 1 44 41 82 30
Télécopie : (33) 1 44 41 82 31

Couverture :
Baga, génie de l'eau
Collection particulière
Maquette : CEPED

Prix : 120 FF TTC